

Qualitative and Quantitative Analysis of Scientific and
Scholarly Communication

Henk F. Moed

Applied Evaluative Informetrics



Springer

Qualitative and Quantitative Analysis of Scientific and Scholarly Communication

Series editors

Wolfgang Glänzel, Katholieke Universiteit Leuven, Leuven, Belgium
Andras Schubert, Hungarian Academy of Sciences, Budapest, Hungary

More information about this series at <http://www.springer.com/series/13902>

Henk F. Moed

Applied Evaluative Informetrics



Springer

Henk F. Moed
Amsterdam
The Netherlands

ISSN 2365-8371 ISSN 2365-838X (electronic)
Qualitative and Quantitative Analysis of Scientific and Scholarly Communication
ISBN 978-3-319-60521-0 ISBN 978-3-319-60522-7 (eBook)
DOI 10.1007/978-3-319-60522-7

Library of Congress Control Number: 2017950268

© Springer International Publishing AG 2017

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

In 1976, Francis Narin, founder and for many years president of the information company CHI Research, published a seminal report to the US National Science Foundation entitled *Evaluative Bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. The current book represents a continuation of his work. It is also an update of an earlier book published by the current author in 2005, *Citation Analysis in Research Evaluation*. In the past 15 years, many new developments have taken place in the field of quantitative research assessment, and the current book aims to describe these, and to reflect upon the way forward.

Research assessment has become more and more important in research policy, management and funding, and also more complex. The role of quantitative information has grown substantially, and research performance is more and more conceived as a *multi-dimensional* concept. Currently not only the classical indicators based on publication and citation counts are used, but also new generations of indicators are being explored, denoted with terms such as altmetrics, webometrics, and usage-based metrics, and derived from multiple multi-disciplinary citation indexes, electronic full text databases, information systems' user log files, social media platforms and other sources. These sources are manifestations of the computerization of the research process and the digitization of scientific-scholarly communication. This is why the current book uses the term *informetrics* rather than *bibliometrics* to indicate its subject.

Informetrics and quantitative science, technology, and innovation (STI) studies have enforced their position as an academic discipline, so that STI indicator development is determined at least partially by an internal dynamics, although external factors play an important role as well, not in the least the business interests of large information companies. As its title indicates, the current book deals with the *application* of informetric tools. It dedicates a major part of its attention to how indicators are used in practice and to the benefits and problems related to this use. It also discusses the relationships between the informetric domain and the research policy and management sphere, and launches proposals for new approaches in research assessment and for the development of new informetric tools.

Following Francis Narin's publication from 1976, the term *evaluative* in the book's title reflects its focus on *research assessment*. But this term refers to the *application domain* and delineates the *context* in which informetric tools are being used. It does *not* mean that informetrics is *by itself* evaluative. On the contrary, this book defends the position that informetricians should maintain *in their informetric work* a *neutral* position towards evaluative criteria or political values.

Target Audience

This book presents an introduction to the field of applied evaluative informetrics. It sketches the field's history, recent achievements, and its potential and limits. It also discusses the way forward. It is written for interested scholars from all domains of science and scholarship, and especially for the following categories of readers.

- All those subjected to research assessment;
- Research students at advanced master and Ph.D. levels;
- Research managers and science policy officials;
- Research funders;
- Practitioners and students in informetrics and research assessment.

Structure

The book consists of six parts as follows:

- **Part I** presents an *introduction* to the use of informetric indicators in research assessment. It provides an historical background of the field and presents the book's basic assumptions, main topics, structure, and terminology. In addition, it includes a *synopsis*, summarizing the book's main conclusions; readers who are interested in the main topics and conclusions of this book but who do not have the time to read it all could focus on this part.
- **Part II** presents an overview of the various types of *informetric indicators* for the measurement of *research performance*. It highlights the multi-dimensional nature of research performance and presents a list of 28 often used indicators, summarizing their potential and limits. It also clarifies common misunderstandings in the interpretation of some often used statistics.
- **Part III** discusses the *application context* of quantitative research assessment. It describes research assessment as an evaluation science and distinguishes various assessment models. It is in this part of the book that the domain of informetrics and the policy sphere are disentangled analytically. It illustrates how external, non-informetric factors influence indicator development and how the policy context impacts the setup of an assessment process.

- **Part IV** presents *the way forward*. It expresses the current author's views on a series of problems in the use of informetric indicators in research assessment. Next, it presents a list of new features that could be implemented in an assessment process. It highlights the potential of informetric techniques and illustrates that *current* practices in the use of informetric indicators could be *changed*. It sketches a perspective on *altmetrics* and proposes new lines in longer term, strategic indicator research.
- **Part V** presents five *lectures with historical overviews* of the field of bibliometrics and informetrics, starting from three of the field's founding fathers: Derek de Solla Price, Eugene Garfield, and Francis Narin. It presents 135 slides and is based on a doctoral course presented by the author at the Sapienza University of Rome in 2015, and on lectures presented at the European Summer School of Scientometrics (ESSS) during 2010–2016, and in the CWTS Graduate Courses during 2006–2009.
- Finally, **Part VI** presents two full articles published recently by the current author in collaboration with his co-authors on hot topics of general interest in which the use of informetric indicators plays a key role. These topics are a critical comparison of five *world university rankings* and a comparison of *usage* indicators based on the number of *full text downloads* with *citation-based* measures.

Acknowledgements

The author wishes to thank the following colleagues for their valuable contributions.

- Dr. Gali Halevi at The Levy Library of the Icahn School of Medicine at Mount Sinai, New York, USA, for her contribution as a co-author of four articles presented in this book, on the multi-dimensional assessment of scientific research (Chap. 8); international scientific collaboration in Asia (Chap. 12); the comparison between Google Scholar and Scopus (Chap. 14); and on a comparative analysis of usage and citations (Chap. 19).
- Prof. Cinzia Daraio at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome for her contribution to the text on ontology-based data base management in Sect. 12.3, and for her valuable comments to a draft version of Chap. 6.
- Prof. Judit Bar-Ilan at the Department of Information Science in Bar-Ilan University, Tel Aviv, Israel, for her contribution as a co-author to the paper on Google Scholar and Scopus (Chap. 14).
- The members of the Nucleo di Valutazione of the Sapienza University of Rome for stimulating discussions about the interpretation and the policy significance of world university rankings discussed in Sect. 10.6.

Executive Summary

This book presents an introduction to the field of applied evaluative informetrics. Its main topic is application of informetric indicators in the assessment of research performance. It gives an overview of the field's history and recent achievements, and its potential and limits. It also discusses the way forward, proposes informetric options for future research assessment processes, and new lines for indicator development.

It is written for interested scholars from all domains of science and scholarship, especially those subjected to quantitative research assessment, research students at advanced master and Ph.D. levels, and researchers in informetrics and research assessment, and for research managers, science policy officials, research funders, and other users of informetric tools.

The use of the term informetrics reflects that the book deals not only with bibliometric indicators based on publication and citation counts, but also with altmetrics, webometrics, and usage-based metrics derived from a variety of data sources, and does not only consider research output and impact, but also research input and process.

Research performance is conceived as a multi-dimensional concept. Key distinctions are made between publications and other forms of output, and between scientific-scholarly and societal impact. The pros and cons of 28 often used indicators are discussed.

An analytical distinction is made between *four* domains of intellectual activity in an assessment process, comprising the following activities.

- *Policy and management:* The formulation of a policy issue and assessment objectives; making *decisions* on the assessment's organizational aspects and budget. Its main outcome is a policy decision based on the outcomes from the evaluation domain.
- *Evaluation:* The specification of an evaluative framework, i.e., a set of evaluation criteria, in agreement with the policy issue and assessment objectives. The main outcome is a *judgment* on the basis of the evaluative framework and the empirical evidence collected.

- *Analytics*: Collecting, analyzing and reporting *empirical* knowledge on the subjects of assessment; the specification of an assessment *model or strategy*, and the *operationalization* of the criteria in the evaluative framework. Its main outcome is an analytical report as input for the evaluative domain.
- *Data collection*: The collection of relevant data for analytical purposes, as specified in an analytical model. Data can be either quantitative or qualitative. Its main outcome is a dataset for the calculation of all indicators specified in the analytical model.

Three *basic assumptions* of this book are the following.

- Informetric analysis is positioned in the analytics domain. A basic notion holds that from what *is* cannot be inferred what *ought to be*. Evaluation criteria and policy objectives are not informetrically demonstrable values. Of course, empirical informetric research may study quality perceptions, user satisfaction, the acceptability of policy objectives, or effects of particular policies, but they cannot provide a foundation of the validity of the quality criteria or the appropriateness of policy objectives. Informetricians should maintain in their informetric work a neutral position towards these values.
- If the tendency to replace reality with symbols and to conceive these symbols as an even a higher form of reality, are typical characteristics of *magical thinking*, jointly with the belief to be able to change reality by acting upon the symbol, one could rightly argue that the un-reflected, unconditional belief in indicators shows rather strong similarities with *magical thinking*.
- The future of research assessment lies in the intelligent *combination* of *indicators* and *peer review*. Since their emergence, and in reaction to a perceived lack of transparency in peer review processes, bibliometric indicators were used to break open peer review processes, and to stimulate peers to make the foundation and justification of their judgments more explicit. The notion of informetric indicators as a support tool in peer review processes rather than as a replacement of such processes still has a great potential.

Five strong points of the use of informetric indicators in research assessment are highlighted: it provides tools to demonstrate performance; shapes one's communication strategies; offers standardized approaches and independent yardsticks; delivers comprehensive insights that reach beyond the perceptions of individual participants; and provides tools for enlightening policy assumptions.

But severe criticisms were raised as well against these indicators. Indicators may be imperfect and *biased*; they may suggest a *façade of exactness*; most studies adopt a *limited time horizon*; indicators can be *manipulated*, and may have *constitutive effects*; measuring *societal impact* is problematic; and when they are applied, an *evaluative framework* and assessment model are often *lacking*.

The following views are expressed, partly supportive, and partly as a counter-critique towards these criticisms.

- Calculating indicators *at the level of an individual* and claiming they measure *by themselves* the individual's performance, suggests *a façade of exactness* that cannot be justified. A valid and fair assessment of individual research performance can be conducted properly only on the basis of sufficient background knowledge on the particular role they played in the research presented in their publications, and by taking into account also other types on information on their performance.
- The notion of making a *contribution to scientific-scholarly progress* does have a basis in reality, that can best be illustrated by referring to an *historical* viewpoint. *History will show* which contributions to scholarly knowledge are valuable and sustainable. In this sense, informetric indicators do *not* measure contribution to scientific-scholarly progress, but rather indicate attention, visibility, or short-term impact.
- *Societal value* cannot be assessed in a politically neutral manner. The foundation of the criteria for assessing societal value is not a matter in which scientific experts have *qualitate qua* a preferred status, but should eventually take place in the policy domain. One possible option is moving away from the objective to evaluate an activity's societal *value*, towards measuring in a neutral manner researchers' *orientation* towards any articulated, lawful need in society.
- Studies on *changes in editorial and author practices* under the influence of assessment exercises are most relevant and illuminative. But the issue at stake is *not* whether scholars' practices *change* under the influence of the use of informetric indicators, but rather whether or not the application of such measures enhances *research performance*. Although this is in some cases difficult to assess without extra study, other cases clearly show traces of mere indicator manipulation with no positive effect on performance at all.
- A typical example of a constitutive effect is that research quality is more and more conceived as what citations measure. More empirical research on the size of constitutive effects is needed. If there is a genuine constitutive effect of informetric indicators in quality assessment, one should not point the critique on current assessment practices merely towards informetric indicators as such, but rather towards any claim for an absolute status of a particular way to assess research quality. Research quality is not what citations measure, but at the same time peers may assess it wrongly.
- If the role of informetric indicators has become too dominant, it does not follow that the notion to intelligently combine peer judgments and indicators is fundamentally flawed and that indicators should be banned from the assessment arena. But it does show the combination of the two methodologies has to be organized in a more balanced manner.
- In the proper use of informetric tools, an *evaluative framework and an assessment model* are indispensable. To the extent that in a practical application an evaluative framework is absent or implicit, there is *a vacuum* that may be easily filled either with ad-hoc arguments of evaluators and policy makers, or with un-reflected assumptions underlying informetric tools. Perhaps the role of such ad hoc arguments and assumptions has nowadays become too dominant.

It can be reduced only if evaluative frameworks become stronger and more actively determine which tools are to be used and how.

The following alternative approaches to the assessment of academic research are proposed.

- A key assumption in the assessment of academic research has been that it is not the *potential* influence or importance of research, but the *actual* influence or *impact* that is of primary *interest to policy makers* and evaluators. But an academic assessment policy is conceivable that rejects this assumption. It embodies a shift in focus from the measurement of performance itself to the assessment of *preconditions* for performance.
- Rather than using citations as indicator of research importance or quality, they could provide a tool in the assessment of *communication effectiveness* and express the extent to which researchers bring their work to the attention of a broad, potentially interested audience. This extent can in principle be measured with informetric tools. It discourages the use of citation data as a *principal* indicator of importance.
- The *functions* of publications and other forms of scientific-scholarly output, as well as their *target audiences* should be taken into account more explicitly than they have been in the past. Scientific-scholarly journals could be systematically categorized according to their function and target audience, and separate indicators could be calculated for each category. More sophisticated indicators of internationality of communication sources can be calculated than the journal impact factor and its variants.
- One possible approach to the use of informetric indicators in research assessment is a systematic exploration of indicators as tools to set *minimum performance standards*. Using baseline indicators, researchers will most probably change their research practices as they are stimulated to meet the standards, but if the standards are appropriate and fair, this behavior will actually increase their performance and that of their institutions.
- At *the upper part* of the quality distribution, it is perhaps feasible to distinguish entities which are '*hors catégorie*', or '*at Nobel Prize level*'. Assessment processes focusing on the very top of the quality distributions could further operationalize the criteria for this qualification.
- Realistically speaking, *rankings of world universities* are here to stay. Academic institutions could, individually or collectively, seek to influence the various systems by formally sending to their creators a request to consider the implementation of a series of new features: more advanced analytical tools; more insight into how the methodological decisions influence rankings; and more information in the system about additional, relevant factors, such as teaching course language.
- In response to major criticisms towards current national research assessment exercises and performance-based funding formula, an alternative model would require less efforts, be more transparent, stimulate new research lines and reduce

to some extent the Matthew Effect. The basic unit of assessment in such a model is the emerging research group rather than the individual researcher. Institutions submit emerging groups and their research programs, which are assessed in a combined peer review-based and informetric approach, applying minimum performance criteria. A funding formula is partly based on an institution's number of acknowledged emerging groups.

The practical realization of these proposals requires a large amount of informetric research and development. They constitute important elements of a wider R&D program of *applied evaluative informetrics*. The further exploration of measures of communication effectiveness, minimum performance standards, new functionalities in research information systems, and tools to facilitate alternative funding formula, should be conducted in a close collaboration between informetricians and external stakeholders, each with their own domain of expertise and responsibilities.

These activities tend to have an applied character and often a short-term perspective. Strategic, longer term research projects with a great potential for research assessment are proposed as well. They put a greater emphasis on the use of techniques from *computer science* and the newly available *information and communication technologies*, and on *theoretical models* for the interpretation of indicators.

- It is proposed to develop new indicators of the *manuscript peer review process*. Although this process is considered important by publishers, editors, and researchers, it is still strikingly opaque. Applying classical humanistic and computational linguistic tools to peer review reports, an understanding may be obtained for each discipline what is considered a reasonable quality threshold for publication, how it differs among journals and disciplines, and what distinguishes an acceptable paper from one that is rejected. Eventually, it could lead to better indicators of journal quality.
- To solve a series of challenges related to the management of informetric data and standardization of informetric methods and concepts, it is proposed to develop an Ontology-Based Data Management (OBDM) system for research assessment. The key idea of OBDM is to create a three-level architecture, constituted by the ontology, a conceptual, formal description of the domain of interest; the data sources; and the mapping between these two domains. Users can access the data by using the elements of the ontology. A strict separation exists between the conceptual and the logical–physical level.
- The creation is proposed of an informetric *self-assessment tool* at the level of individual authors or small research groups. A challenge is to create an online application based on key notions expressed decades ago by Eugene Garfield about author benchmarking, and by Robert K. Merton about the formation of a *reference group*. It enables authors to check the indicator data calculated about themselves, decompose the indicators' values, learn more about informetric indicators, and defend themselves against inaccurate calculation or invalid interpretation of indicators.

- As an illustration of the importance of theoretical models for the interpretation of informetric indicators, a model of a country's scientific development is presented. Categorizing national research systems in terms of *the phase* of their scientific development is a meaningful alternative to the presentation of rankings of entities based on a single indicator. In addition, it contributes to the solution of ambiguity problems in the interpretation of indicators.

It is proposed to dedicate in doctoral programs more attention to the ins and outs, potential and limits of the various assessment methodologies. Research assessment is an activity one can learn.

Contents

Part I General Introduction and Synopsis

1	Introduction	3
1.1	The Value and Limits of Informetric Indicators in Research Assessment	3
1.2	A Short History of Bibliometrics and Informetrics	7
1.2.1	The Start	7
1.2.2	Recent Developments	9
1.3	Basic Assumptions	12
1.4	This Book's Main Topics	13
1.5	Structure of Book	15
1.6	A Note on Terminology	16
1.7	Re-Use of Paragraphs of Previously Published Articles	17
2	Synopsis	19
2.1	Part I. Introduction	19
2.1.1	Chapter 1	19
2.2	Part II. Informetric Indicators of Research Performance	21
2.2.1	Chapter 3	21
2.2.2	Chapter 4	21
2.2.3	Chapter 5	24
2.3	Part III. The Application Context	24
2.3.1	Chapter 6	24
2.3.2	Chapter 7	26
2.3.3	Chapter 8	27
2.4	Part IV. The Way Forward	28
2.4.1	Chapter 9	28
2.4.2	Chapter 10	30
2.4.3	Chapter 11	33
2.4.4	Chapter 12	34

2.5	Part V. Lectures	38
2.5.1	Chapter 13.	38
2.5.2	Chapter 14.	38
2.5.3	Chapter 15.	38
2.5.4	Chapter 16.	39
2.5.5	Chapter 17.	39
2.6	Part VI. Papers	39
2.6.1	Chapter 18.	39
2.6.2	Chapter 19.	40
Part II Informetric Indicators of Research Performance		
3	Multi-dimensional Research Performance	45
3.1	Introduction	46
3.2	Research Outputs	47
3.3	Research Impacts	48
3.4	Research Infrastructure	49
3.5	Summary Table of 28 Important Informetric Indicators	50
4	Informetric Tools.	61
4.1	Introduction	61
4.2	Indicators	62
4.2.1	Publication-based Indicators	62
4.2.2	Citation-based Indicators	63
4.2.3	Journal Metrics	64
4.2.4	Patent-based Indicators	66
4.2.5	Usage-based Indicators	67
4.2.6	Altmetrics	68
4.2.7	Webometric Indicators.	69
4.2.8	Economic Indicators	70
4.2.9	Reputation and Esteem-based Indicators	71
4.2.10	Indicators of Research Collaboration and Cross-Disciplinarity.	72
4.2.11	Indicators of Research Infrastructure	73
4.3	Big Informetric Data.	74
4.4	Science Maps	76
5	Statistical Aspects	79
5.1	Introduction	79
5.2	Journal Impact Factors Are no Good Predictors of Citation Rates of Individual Articles	80
5.2.1	Aftermath	82
5.3	Errors or Biases in Data Samples	83
5.4	How to Interpret Correlation Coefficients?	84

Part III The Application Context

6 Research Assessment as an Evaluation Science	89
6.1 Introduction	89
6.2 Evaluation Science	90
6.2.1 Research Versus Management Tools	90
6.2.2 Comprehensive Theory of Change	91
6.2.3 Performance Management Versus Evaluation	92
6.2.4 Summative Versus Formative Evaluation	93
6.2.5 Normative Versus Criterion Based Reference Framework	93
6.2.6 Evaluation Versus Assessment	93
6.3 Types of Intellectual Activity in Assessment	94
6.4 Assessment Models and Strategies	98
6.4.1 Base Distinctions	98
6.4.2 Assessment of Basic Science Combining Peer Review and Bibliometric Indicators	99
6.4.3 Program Assessment: Empowerment Evaluation (EE)	100
6.4.4 Field-Specific Evaluation: The Becker Model	100
6.5 Costs of Research Assessment	101
7 Non-informetric Factors Influencing Indicator Development	103
7.1 Introduction	103
7.2 How Evaluative Assumptions Shape Indicators	104
7.2.1 Size Dependent Versus Size Independent Indicators	104
7.2.2 Focus on the Top or the Bottom of a Performance Distribution?	108
7.2.3 Which Indicator Normalizations Should Be Implemented?	109
7.2.4 How to Define a Proper Reference Framework?	109
7.2.5 Short Term Versus Long Term Perspective	110
7.3 The Influence of the Wider Context on Indicator Development	111
7.4 Indicator Development and Business Interests	115
8 The Policy Context	119
8.1 Introduction	119
8.2 The Multi-dimensional Research Assessment Matrix	120
8.2.1 Units of Assessment	121
8.2.2 Objectives and Performance Dimensions	121
8.3 Systemic Characteristics of the Units of Assessment	121
8.3.1 The Use of Journal Impact Factors for Measuring International Orientation	123
8.3.2 The Use of Publication Counts in the Assessment of Being Research Active	124

8.4	Meta-Analyses	125
8.5	Policy Considerations	126
Part IV The Way Forward		
9	Major Problems in the Use of Informetric Indicators	131
9.1	The Problem of Assessing Individual Scholars	131
9.2	The Effect of a Limited Time Horizon	132
9.3	The Problem of Assessing Societal Impact	134
9.4	The Effects of the Use of Indicators upon Authors and Editors	135
9.5	Constitutive Effects of Indicators and Magical Thinking About Research Quality	137
9.6	The Need for an Evaluative Framework and an Assessment Model	139
10	The Way Forward in Quantitative Research Assessment	141
10.1	Introduction	141
10.2	Communication Effectiveness as a Precondition for Performance	142
10.3	Some New Indicators of Multi-dimensional Research Output	143
10.3.1	Journal Functions and Target Audiences	144
10.3.2	A Note on Journal Coverage of the Citation Indexes	145
10.3.3	Research Training and Scientifically Developing Countries	146
10.4	Definition of Minimum Performance Standards	146
10.5	Policy Towards World University Rankings	148
10.6	An Alternative Approach to Performance Based Funding	150
10.7	Concluding Remark	152
11	A Perspective on Altmetrics	153
11.1	Introduction	153
11.2	The Computerization of the Research Process	155
11.3	Michael Nielsen's "Reinventing Discovery"	156
11.4	Useful Distinctions	157
11.5	Concluding Remarks	159
12	The Way Forward in Indicator Development	161
12.1	Towards New Indicators of the Manuscript Peer Review Process	161
12.1.1	Introduction	161
12.1.2	Analyses	162
12.1.3	Concluding Remarks	164

12.2	Towards an Ontology-Based Informetric Data Management System	164
12.2.1	Introduction	164
12.2.2	An OBDM Approach	165
12.2.3	Design of Indicators	166
12.2.4	Concluding Remarks	167
12.3	Towards Informetric Self-Assessment Tools	168
12.3.1	Introduction	168
12.3.2	Why an Informetric Self-Assessment Tool Is Useful	168
12.3.3	What an Informetric Self-Assessment Tool Could Look like	169
12.4	Towards Informetric Models of Scientific Development	170
12.4.1	Introduction	170
12.4.2	A Model of Scientific Development	170
12.4.3	Application to South-East Asian Countries	172
12.4.4	Application to the Persian Gulf Region	173
12.4.5	Concluding Remarks	174

Part V Lectures

13	From Derek Price's Network of Scientific Papers to Advanced Science Mapping	177
13.1	Networks of Scientific Papers	177
13.2	Modelling the Relational Structure of Subject Space	183
13.3	Mapping Software	191
14	From Eugene Garfield's Citation Index to Scopus and Google Scholar	193
14.1	Science Citation Index and Web of Science	193
14.2	Scopus Versus Web of Science	200
14.3	Google Scholar Versus Scopus	203
14.4	Concluding Remarks	207
15	From Francis Narin's Science-Technology Linkages to Double Boom Cycles in Technology	209
15.1	Citation Analysis of the Science-Technology Interface	210
15.2	Theoretical Models of the Relationship Between Science and Technology	213
15.3	Double Boom Cycles in Product Development	218
15.4	Selected Case Studies	221

16 From Journal Impact Factor to SJR, Eigenfactor, SNIP, CiteScore and Usage Factor	229
16.1 Journal Impact Factors	229
16.2 Effect of Editorial Self-citations	234
16.3 Alternative Journal Metrics.	236
17 From Relative Citation Rates to Altmetrics.	245
17.1 Citation-Based Indicators	245
17.2 Usage-Based Indicator and Altmetrics	250
17.3 Efficiency Indicators	253
Part VI Papers	
18 A Comparative Study of Five World University Rankings	261
18.1 Introduction	261
18.2 Analysis of Institutional Overlap	265
18.3 Geographical Distributions	267
18.4 Indicator Scores and Their Distributions.	269
18.4.1 Missing Values	269
18.4.2 From Data to Indicators.	270
18.4.3 Skewness of Indicator Distributions.	271
18.5 Statistical Correlations	272
18.6 Secondary Analyses	277
18.6.1 Characteristics of National Academic Systems	277
18.6.2 QS Versus Leiden Citation-Based Indicators	279
18.6.3 THE Research Performance Versus QS Academic Reputation.	280
18.6.4 ARWU Highly Cited Researchers Versus Leiden Top Publications Indicator	281
18.7 Discussion and Conclusions	282
18.8 Concluding Remarks.	285
19 Comparing Full Text Downloads and Citations	287
19.1 Introduction	287
19.2 Data Collection.	289
19.3 Results	289
19.3.1 Downloads Versus Citations of an Individual Article	289
19.3.2 Downloads by User Institution	291
19.3.3 Downloads Time Series Per Journal and Document Type	292
19.3.4 Download-Versus-Citation Ratios	294
19.3.5 Statistical Correlations Between Downloads and Citations at the Journal and Article Level	296

Contents	xxi
19.4 Discussion and Conclusions	297
19.4.1 Analyses by User Country and Institution	297
19.4.2 Downloads Time Series Per Journal and Document Type	298
19.4.3 Download-Versus-Citation Ratios	298
19.4.4 Statistical Correlation Between Downloads and Citations	299
19.4.5 Factors Differentiating Between Download and Citations	299
References	301

Part I

General Introduction and Synopsis

Chapter 1

Introduction

Abstract This chapter presents an introduction to the book. It starts with an overview of the value and limits of informetric indicators in research assessment, and presents a short history of the field. It continues with the book's main assumptions, scope and structure. Finally, it clarifies the terminology used in the book.

Keywords Academic research · Altmetrics · Assessment model · Citation analysis · Citation index · Constitutive effect · Desktop bibliometrics · Developing country · Document views · Evaluative framework · Funding formula · Google scholar · Manipulation · Mendeley · Metrics obsession · Ocular surgery · Peer review · Performance based funding · Price medal · ResearchGate · Science citation index · Scopus · Social media · Societal impact · Usage · Value for money · Web of science · Webometrics

1.1 The Value and Limits of Informetric Indicators in Research Assessment

What is the value of informetric indicators in the assessment of scientific-scholarly research?

- Informetric tools may help researchers and their organizations to demonstrate their performance. As an example, consider an institute developing and applying new tools in the field of ocular surgery. There is evidence that the work of the institute's staff is widely known across the globe, especially in the leading clinical centers in the field. Patients from many different countries come to the institute and undergo interventions. Specialists from all over the world seek to obtain a fellowship in the institute to learn the newest techniques. Clinical surveys confirm the success of these techniques. A way to further substantiate the contribution the institute made to research and clinical practice in its subject field is to examine traces of influence of its work in the scientific literature. Table 1.1 presents a citation analysis of the group's performance.

Table 1.1 Citation count and rank of the institute's five most frequently cited articles

Rank	Year	Total citations	Rank in annual volume	Nr. docs in annual volume
1	1998	416	1	96
2	2006	312	1	267
3	2004	273	2	151
4	1999	237	4	267
5	1999	224	3	220

Source Scopus. Data collected by this book's author on 28 Dec 2016. The first data row shows that the most highly cited article published from the institute in 1998 has been cited 416 times. The total number of documents published in this journal in the same year by researchers from all over the world amounts to 96. In a ranking of these 96 articles based on citation counts, the institute's paper ranks first. The table shows that all 5 articles included in the table are highly ranked in this way. A further analysis of a larger set of their highly cited work reveals that the group published both in the very specialist journals and in more general journals, and in journals with a European, British or American orientation

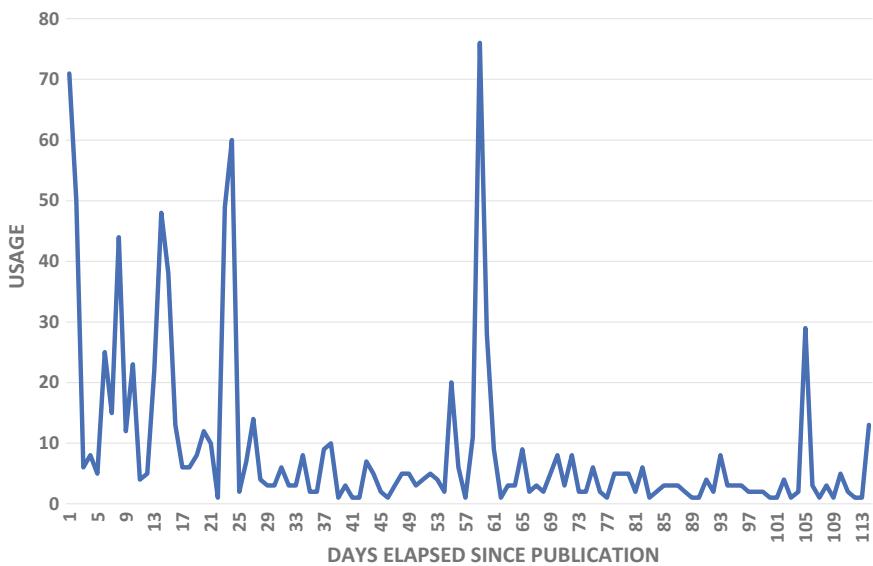


Fig. 1.1 Usage counts (online views and full txt downloads) for the paper *Toward new indicators of a journal's manuscript peer review process*, by H.F. Moed, published in *Frontiers in Research Metrics and Analytics*. Usage starts soon after its online publication. The daily fluctuations during the first weeks since publication are probably due to weekend holidays. About 60 days after publication the paper was presented at the OECD Blue Sky Conference in Ghent, Belgium, 19–21 Sept. 2016. Around this date, usage has increased substantially

- Informetric indicators can be useful tool for authors who are interested in tracking the degree of attention to their work, and in assessing the effectiveness of their communication strategies. Figure 1.1 shows a useful indicator that has become available on a large scale during the past decade, the number of online

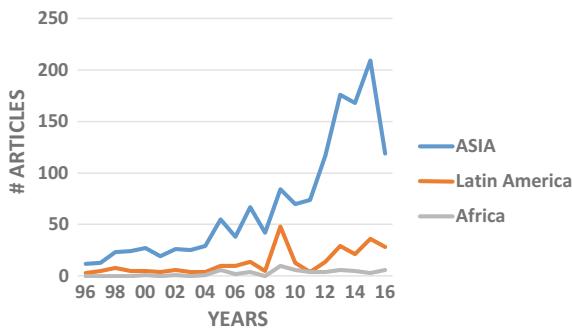


Fig. 1.2 Number of articles published in the journal *Scientometrics* by authors from three geographical regions: Asia, Latin America and Africa. The absolute number of articles published in this journal increased over the years, from 80 to 100 per year during 1996–2004 to over 300 in 2016, with peaks even above 500 in 2013 and 2015. The figure shows a steady increase in the share of articles from Asia, with over 100 papers in 2016. Major publishing countries in this region are China, Taiwan, South Korea, Japan and Iran. As from 2010, there is an increase also in papers from Latin America, with 28 papers in 2016, major countries being Brazil, Mexico and Chile. The peak for this region in 2009 is probably due to the fact that in this year one of the major international conferences in the field, the Conference of the International Society for Scientometrics and Informetrics was held in Brazil. Africa shows publication activity in the journal as from 2000. The most important African country is South Africa

views or full text downloads of articles published in an electronic journal. Figure 1.2 presents the number of document views and downloads—denoted as usage—of one particular paper, presented at an international conference of policy makers. It shows a substantial increase in usage around the day it was presented at that conference. This case illustrates also how strongly usage counts may depend upon the behavior of the authors themselves, and therefore are, to some extent, manageable.

- The use of a well-documented and validated informetric method in an assessment process enables an evaluator to achieve a certain degree of standardization in the process, and to compare units of assessment against an independent yardstick. These characteristics are sometimes indicated with the term ‘objective’. Use of such a method reduces the risk that the outcomes of an assessment are biased in favor of particular external interests. Rather than considering the collection and interpretation of indicators as an individual matter of participating evaluators, the decision to formally use in a peer review process a balanced informetric method process may contribute to an appropriate, informed outcome of the assessment.
- Informetric tools may produce insights and relationships beyond the horizon of an individual expert’s knowledge. This is true not only for indicators, but especially for science maps, revealing an informetric image of innumerable relationships among objects, such as single scientific articles, authors, institutions, journals, research topics and disciplines. They may provide ‘aerial views’ of aggregate data reflecting the behavior of large numbers of actors, views that are complementary to the outcomes of more qualitative and disaggregated approaches.

- They can also be used as tools to critically and empirically investigate how quality perceptions are formed, and examine the validity of policy assumptions about the functioning of the scientific-scholarly system, and the effectiveness of policy measures. As argued in Sect. 8.4, informetric studies of a research system can also shed light on which type of research assessment process and which types of indicators are the most promising in a performance analysis of that system.

Informetric tools in research assessment have their *limitations* too. Severe criticism is raised against the current use of citation-based indicators and other informetric measures in research assessment. The following often debated issues are further discussed in Chap. 9, in which the current author gives his view and suggests possible solutions.

- *Indicators may be imperfect or biased.* For instance, generating attention and making a contribution to scientific-scholarly progress are not identical concepts. If one agrees that performance tends to attract attention, the degree of received attention may be interpreted as an indicator of performance. But other factors influence the level of attention as well, for instance, the way at which a piece of work is being exposed to a wide audience. The awareness that all performance indicators are ‘partial’ or ‘imperfect’—influenced as they may be not only by the various aspects of performance but also by other factors that have little to do with performance—is as old as the use of performance indicators itself. Indicators may be *imperfect* or *biased*, but in the application of such indicators this is not seldom forgotten.
- *Most studies adopt a limited time horizon.* The time horizon in an assessment is the considered length of the time period for an objective to meet, outcome to be created or impact to be generated. Therefore, the question which *time horizon* is employed in an assessment is crucial. In most informetric assessments, the time horizon is 10 years or less, and the focus is on recent past performance, as it is believed to increase the policy relevance, and reduce data collection costs. In the calculation of altmetrics and usage counts, time windows in impact measurement can be even less than one year.
- *Indicators may be manipulated.* An often articulated critique on the use of bibliometric indicators states that if publishing articles and being cited is the norm, researchers change their behavior in order to obtain the highest possible score. This may lead to ‘strategic’ behavior and even to indicator manipulation. For instance, there is evidence that some journal editors seek to influence the value of journal impact factor of their journals.
- *Indicators may have constitutive effects.* Another criticism states that in the minds of those concerned the meaning of a concept that an indicator claims to measure is more and more narrowed to the definition of that indicator. Thus, scientific production is more and more equated with publishing articles, and research quality with being well cited. This phenomenon is denoted as the constitutive effect of an indicator (Dahler-Larsen, 2014).

- *Measuring societal impact is problematic.* The notion of multi-dimensional impact has created a growing interest in the societal—technological, social, economic, educational, cultural—value of research. But the time delays involved in generating impact in these extra-scientific domains may be typically 10 years or even longer. Also, societal merit cannot be measured in a politically neutral manner. What is socially valuable according to one political view, may be considered inappropriate in an alternative view.
- *An evaluative framework and assessment model are often lacking.* An evaluative framework aims to set evaluation criteria in an assessment, derived from the policy issue at stake and from the assessment objectives. Informetric indicators are often available without offering such a framework. This may make their role too dominant and give space to constitutive effects. Also, they may be applied without a well-defined assessment model, specifying how the assessment will take place, and ensuring it is not only efficient and fit-to-purpose, but also fair.

1.2 A Short History of Bibliometrics and Informetrics

1.2.1 The Start

Pioneering work has been conducted the 1960s and 1970s by Derek de Solla Price, a visionary who applied bibliometric techniques in a ‘science of science’, Eugene Garfield, the founder of the Science Citation Index, and Francis Narin, who introduced the term ‘evaluative bibliometrics’. Chapters 13, 14, 15, 16 and 17 in *Part V* of this book give an overview of their contributions.

A theoretical basis of the use of citation-based indicators as measures of intellectual influence is provided by the notion developed in Robert K. Merton’s sociological studies that references give credit where credit is due, acknowledge the community’s intellectual debts to the discoverer, and can be conceived as registrations of intellectual property and peer recognition (Merton, 1957; 1996). Moed (2005a, p. 193 a.f.) gives an overview of the views of a series of authors on what cited references and citations measure. It should be emphasized that according to Merton’s theory, a reference acknowledges the *source* of a knowledge claim, but not necessarily the claim’s *validity*. In other words, it does not provide any theoretical justification of the claim that the more cited a piece of work is, the more valid are its results.

One of the very first studies applying citation-based indicators as measures of intellectual influence was conducted by Stephen Cole and Jonathan Cole (1967, 1971). They used these indicators as a sociological *research tool*. This type of use should be distinguished from the application of indicators in an *evaluative context* for the assessment of research performance of individuals or groups. The former type of use eventually aims at testing some kind of research *hypothesis* or revealing a structure. The latter may lead to statements on the performance of particular, designated individual scientists in the research system. As Stephen Cole (1989) puts it:

Citations are a very good measure of the quality of scientific work for use in sociological studies of science; but because the measure is far from perfect it would be an error to reify it and use it to make individual decisions. [...] In sociological studies our goal is not to examine individuals but to examine the relationships among variables (Cole, 1989, p. 11).

The pioneering work in the USA by Narin (1976) showed that international scientific influence as measured by citations is a crucial parameter in the measurement of research performance. In Europe, Ben Martin and John Irvine further developed his approach. They proposed workable definitions of key concepts such as ‘indicator’, ‘influence’ and ‘impact’.

The citation rate is a partial indicator of the impact of a scientific publication: that is, a variable determined partly by (a) the impact on the advance of scientific knowledge, but also influenced by (b) other factors, including various social and political pressures such as the communication practices [...], the emphasis on the numbers of citations for obtaining promotion, tenure or grants, and the existing visibility of authors, their previous work, and their employing institution (Martin and Irvine, 1983, p. 70).

Anthony van Raan emphasized the *complementarity* between bibliometric indicators and peer review.

I do not plead for a replacement of peer review by bibliometric analysis. Subjective aspects are not merely negative. In any judgment there must be room for the intuitive insights of experts. I claim, however, that for a substantial improvement of decision making an advanced bibliometric method [...] has to be used in parallel with a peer-based evaluation procedure (van Raan, 2004a, p. 27).

At the end of the 1980s van Raan founded at the University of Leiden the Centre for Science and Technology Studies (CWTS), of which he was the director for almost 30 years. He established the Science and Technology Indicators Conference Series, edited the first handbook on science and technology indicators (van Raan, ed., 1987), and supervised numerous Ph.D.s students, including the author of this book.

Tibor Braun, Wolfgang Glanzel and Andras Schubert at the Academy of Sciences in Budapest were in the early 1980s the first to systematically calculate a series of bibliometric macro-indicators derived from the Science Citation Index for all countries in the world. As from the late 1980s, many other authors started using citation and publication-based indicators in a large number of studies related to various entities: individual researchers, research groups, departments or institutes, universities, countries, journals and subject fields.

Many valuable and informative historical overviews of the field of informetrics have been published during the past decades (e.g. Mingers and Leydesdorff, 2015). A comprehensive review of citation-based indicator development is presented by Waltman (2016). Table 1.2 gives a list of the institutions of the recipients of the Derek de Solla Price Memorial Medal, periodically awarded to scientists with outstanding contributions to the fields of quantitative studies of science, by a panel comprised of the editors and members of the advisory board of the journal *Scientometrics* together with former Price awardees.

Table 1.2 shows that the overwhelming part of the awardees were active in institutions located in European and North American countries. The medalists and

Table 1.2 Institutions of recipients of the Derek de Solla Price Award

Region	Country	Institution	Year of Award
Europe	Belgium	Univ Hasselt	2001
	Belgium	Univ Antwerp	2001
	Czechoslovakia	Czechoslovak Acad Sci	1989
	Denmark	Royal Library School	2005
	France	OST/INRA	2009
	Hungary	Hungarian Acad Sciences	1986; 1993; 1999; 2009
	Netherlands	Leiden Univ	1995; 1999
	Netherlands	Univ Amsterdam	2003
	Sweden	Umea Univ	2011
	UK	City Univ London	1989
	UK	Univ. Sussex	1997
	UK	Loughborough Univ	2015
North America	USA	Inst. Scientific Inform.	1984; 1987
	USA	Univ Oregon	1985
	USA	CHI Research	1988
	USA	Columbia Univ	1995
	USA	Drexel Univ	1997; 2005; 2007
	USA	Indiana Univ	2013
Other	Israel	Bar-Ilan Univ	2017
	USSR	Moscow State Univ	1987

their colleagues made excellent contributions to the field, and their awards are well deserved. While in the first half of the time period awardees tend to be located in the USA, in the second half they are mainly affiliated with European institutions. But Fig. 1.2 reveals that institutions in other geographical regions are emerging. Both the absolute number of the percentage share of articles published in the journal *Scientometrics* by authors from Asia and Latin America has increased substantially over the years. In 2016, almost 40% of papers was authored by an author from Asia, and about 10% by a Latin American author. This outcome illustrates that an indicator based on the number of awards is to some extent conservative, and does not keep pace with the most recent developments. The bibliometric trend forecasts award winners from Asia and Latin America in the near future.

1.2.2 Recent Developments¹

In the current economical atmosphere where budgets are strained and funding is difficult to secure, ongoing, diverse and thorough assessment is of immense

¹This sub-section re-uses with permission selected paragraphs from Moed and Halevi (2015).

importance for the progression of scientific and research programs and institutions. Research assessment is an integral part of any scientific activity. It is an ongoing process aimed at improving the quality of scientific-scholarly research. It includes evaluation of research quality and measurements of research inputs, outputs and impacts, and embraces both qualitative and quantitative methodologies, including the application of bibliometric indicators and peer review.

The manner by which research assessment is performed and executed is a key matter for a wide range of stakeholders including program directors, research administrators, policy makers and heads of scientific institutions as well as individual researchers looking for tenure, promotion or to secure funding to name a few. These stakeholders have also an increasing concern regarding the quality of research performed especially in light of competition for talent and budgets and mandated transparency and accountability demanded by overseeing bodies (Hicks, 2009).

The following trends can be identified in the *science policy domain* during the past ten years.

- *Emphasis on societal value and value for money.* In most OECD countries, there is an increasing emphasis on the effectiveness and efficiency of government-supported research. Governments need systematic evaluations for optimizing their research allocations, re-orienting their research support, rationalizing research organizations, restructuring research in particular fields, or augmenting research productivity. In view of this, they have stimulated or imposed evaluation activities of their academic institutions. Universities have become more diverse in structure and are more oriented towards economic and industrial needs.
- *Performance-based funding.* Funding of scientific research—especially in universities—tends to be based more frequently upon performance criteria, especially in countries in which research funds were in the past mainly allocated to universities by the Ministry responsible for research as a block grant, the amount of which was mainly determined by the number of enrolled students. It must be noted that in the U.S. there has never been a system of block grants for research; in this country research funding was, and is still, primarily based on peer review of the content of proposals submitted to funding organizations. Government agencies as well and funding bodies rely on evaluation scores to allocate research budgets to institutions and individuals. Such policy requires the organization of large scale research assessment exercises (OECD, 2010; Hicks, 2010) especially in terms of monetary costs, data purchasing, experts' recruitment and processing systems (Jonkers & Zacharewicz, 2016).
- *Research competition in a global market.* Research institutions and universities operate in a global market. International comparisons or rankings of institutions are being published on a regular basis with the aim to inform students, researchers and knowledge seeking external groups. Research managers use this information to benchmark their own institutions against their competitors (Hazelkorn, 2011). Indicators from numerous world university ranking systems such as the Shanghai, Times Higher Education, QS and Leiden Ranking, and U-Multirank play an important role.

- *Internal research assessment systems.* More and more institutions implement internal research assessment processes and build research information systems (see for instance EUROCRIS, 2013) containing a variety of relevant input and output data on the research activities within an institution, enabling managers to distribute funds based on research performance.

Due to the computerization of the research process and the digitization of scholarly communication, more and more policy-relevant data sources are becoming available. Quantitative research assessment becomes a ‘big data’ activity. The following specific trends can be observed.

- *Multiple, comprehensive citation indexes.* While the Science Citation Index founded by Eugene Garfield (1964) and published by the Institute for Scientific Information (currently Clarivate’s Web of Science) has for many years been the only database with a comprehensive coverage of peer reviewed journals, in 2004 two new indexes entered the market, namely Elsevier’s *Scopus* and *Google Scholar*.
- *Full texts in digital format.* Due to electronic publishing, more and more full texts of research publications are available in a digital format. While in the past bibliometric studies were mostly based on publication *meta-data*—including cited references—, current informetric techniques increasingly analyze *full texts*.
- *Usage data from publishers’ sites.* Major publishers make their content electronically available on-line, and researchers as well as administrators are able to measure the use of their scientific output as a part of an assessment process (Luther, 2002; Bjork & Paetau, 2012).
- *Construction of large publication repositories:* Disciplinary or institutionally oriented publication repositories are being built, including meta-data and/or full text data of publications made by an international research community in a particular subject field, or by researchers active in a particular institution, respectively (Fralinger & Bull, 2013; Burns, Lana & Budd, 2013).
- *Altmetric and other new data sources:* More and more researchers use social media such as Twitter, reference managers like Mendeley, and scholarly blogs, to communicate with each other and to promote their work. Traces of this use are stored in electronic files that can be analyzed with informetric techniques. Also, patent data of major patent offices are available in electronic form, and the Word Wide Web itself can be used as a data source in webometric analysis.

The above trends in the science policy domain and in the computerization of the research and communication processes generated an increasing interest in the development, availability and practical application of new indicators for research assessment.

- *Development of new indicators.* In the past decade many new indicators have been developed. They cover new dimensions of research communication and performance, and have become more sophisticated. Typical examples are altmetrics, webometrics and usage based measures, or citation-context analyses. Their development attracted more and more specialists from other disciplines, including statistical physics, molecular biology, econometrics and computer science.

- *More indicators are becoming available.* Currently, indicators such as author h-indices and total citation and publication counts are available in the three large literature databases Web of Science, Scopus and Google Scholar, or in special assessment tools such as Clarivate's Incites and Elsevier's SciVal. Many measures are produced by small, specialized firms, such as Altmetric.com or Plum Analytics, or by spin-offs such as Scimago.com and CWTS.com. Reference managers such as Mendeley and ResearchGate provide indicators as well, based on data from their own systems.
- *Desktop bibliometrics.* Calculation and interpretation of science metrics are not always made by bibliometric experts. "Desktop bibliometrics", a term coined by Katz and Hicks (1997) and referring to an evaluation practice using bibliometric data in a quick, often unreliable manner, is becoming a reality.
- *More and more critique on the use of indicators.* In a comment published in Nature, Benedictus and Miedema (2016) criticized what they denote as an "obsession with metrics" in assessment practices at Dutch academic institutions, especially academic hospitals. They argued that a pressure to publish as many papers and generate as many citations as possible created a tendency to give far too little weight to its influence of research work on patient care. Their experiences are good illustrations of the constitutive effects of indicators mentioned in the previous section.

1.3 Basic Assumptions

A basic notion holds that from what *is* cannot be inferred what *ought to be*. This notion has implications for the role of informetrics in the foundation of the evaluative criteria to be applied in an assessment, or of the political objectives of the assessment. Evaluation criteria and policy objectives are not informetrically demonstrable values. Of course, empirical informetric research may study quality perceptions, user satisfaction, the acceptability of policy objectives, or effects of particular policies, but they cannot provide a theoretical foundation of the validity of the quality criteria or the appropriateness of policy objectives. Informetricians should maintain in their informetric work a neutral position towards these values.

The current author conceives informetrics as a value free, empirical science. Being value free is conceived as a *methodological* requirement. This book shows how statistical definitions of indicators of research performance are based on theoretical assumptions on what constitutes performance. The informetric component and the domain of evaluative or political values in an assessment are disentangled by distinguishing between quantitative-empirical, informetric evidences on the one hand, and an evaluative framework based on normative views on what constitutes research performance and which policy objectives should be achieved, on the other. This distinction is further discussed in Sect. 6.3.

Above it was argued that informetricians should maintain *in their informetric work* a neutral position towards evaluative or political values. This statement should be further qualified in the following manner. Maintaining a neutral position towards evaluative criteria or political objectives is a methodological requirement. But informetricians are also researchers and members of the wider research community. There are *also* potential subjects of research assessment themselves. As researchers and as assessed subjects they do have views on what constitutes performance and what are appropriate and less appropriate political objectives. The methodological requirement of a value-free informetrics does not mean that they are not ‘allowed’ to have these views, nor that they are not allowed to communicate their views to the outside world. But they should make such views *explicit*, and make at the same time clear that these are their own views that are *not* informetrically *demonstrable*.

As regards the application of informetric tools in research assessment, a key assumption underlying this book is that the future of research assessment lies in the intelligent *combination* of *indicators* and *peer review*. From the very beginning, and in reaction to a perceived lack of transparency in peer review processes, and to the critical view of peer review as an instrument to consolidate an ‘old boys’ network, bibliometric indicators were used to break open peer review processes, and stimulate peers to make the foundation and justification of their judgments more explicit. The notion of informetric indicators as a support tool in peer review processes rather than as a replacement of such processes has still a great potential, and this book aims to further explore it. A necessary condition for achieving this is a thorough awareness of the potentialities and limitations of *both* methodologies.

1.4 This Book’s Main Topics

This book aims to inform the ongoing debate about the validity and usefulness of informetric indicators in research assessment. It gives an overview of recent insights into their potential and limits, and does not only include *classical* bibliometric indicators based on publication and citation counts, but also newly developed measures such as *altmetrics* derived from social media appearances, and *usage* metrics based on full text downloads or online views of electronic publications.

Contrary to the situation in the 1960s, informetric tools are currently not merely used by a selected group of librarians and scientific information experts, but by many researchers in their daily practices, managers of research institutions in various types of assessment processes, politicians in the development of research funding mechanisms, the daily and weekly press in presenting the wider public detailed rankings of world universities, and information companies not only in their product portfolios but also in their marketing. The five main topics of this book are as follows.

- *An overview of new informetric tools.* A first topic is the description of a series of important *new* databases, methodologies, indicators and products introduced

during the past ten years in the field of informetrics and its application in research assessment. The field has attracted interested experts from many other research disciplines, including statistical physics, social network analysis, molecular biology, and especially, computer science and artificial intelligence. And the number of new informetric tools has increased enormously. This book focuses on a series of hot topics: the use of the database Google Scholar; development phases of scientifically developing countries; new indicators derived from social media or based on full text downloads; and the publication of numerous rankings of world universities.

- *Often used informetric indicators and their pros and cons.* As outlined in the previous sections, the number of indicators available in research assessment has increased substantially over the years. A second topic is the categorization and a comprehensive overview of the most important indicators or indicator families, the aspects or dimensions they are assumed to be measuring, and their principal pros and cons. This discussion does *not* focus on the *technical* and *statistical* aspects of the indicators, but on their basic *theoretical* notions and assumptions.
- *The relationship between the informetric and the policy domain.* A third key topic in this book is a critical discussion of the current role of indicators in research assessment, and to reflect upon the relationship between the domain of informetrics and the policy domain. It is argued that, on the one hand, quantitative research assessment should not become a plaything in the hands of the policy makers and managers using informetric tools, nor of the big and smaller information companies producing them. It is a scientific-scholarly field with its own, independent criteria of validity and good practice. On the other hand, informetric experts should not sit on the chair of a policy maker or manager, and decide—implicitly or explicitly—on normative, political issues. Instead, they should inform the policy domain, by offering instrumental support to its policies and insights and enable a reflection upon its policy objectives.
- *Options for consideration when designing an assessment process.* A fourth topic relates to the *application* of informetric indicators. The book addresses a series of basic issues and proposes a series of *options* that could be considered when designing and implementing a research assessment process. The list of options aims to illustrate the potential of informetric techniques. It aims to create openness and space for further reflection by illustrating that *current* practices in the use of informetric indicators could be *changed*. In this way it shows what *could* be done, *not* what *should* be done. As argued above, the latter issue is to be solved in the definition of an evaluative framework, integrating policy needs and informetric evidences with a view of what is valuable and must be achieved. Informetricians, including the author of this book, should maintain a neutral position towards the normative aspects of such a framework.
- *Future research and indicator development.* A last topic relates to the development of new indicators for research assessment. The book proposes new lines of research that can facilitate the development of new indicators. Two main components are distinguished. The first is an emphasis on the need to develop theoretical models as guides for interpretation and adequate use of indicators.

During the past decades, *data* collection and handling have been a key priority. But nowadays it becomes clear more data does not always mean a better understanding. Such models can in principle come from all interested disciplines. A second component is an emphasis on the use of methods from computer science, and of the potential of the new information and communication technologies. Perhaps many indicators that are currently still being applied, as well as the way in which they are used, are determined by techniques that were developed some 50 years ago. New technical and analytical approaches can bring more progress than they have achieved thus far.

1.5 Structure of Book

The book consists of six parts.

- **Part I** presents an introduction to the use of informetric indicators in research assessment. It provides an historical background to the topic, and presents the book's basic assumptions, main topics, structure and terminology. In addition, Chap. 2 presents a synopsis, summarizing the book's main conclusions of each of its 25 chapters in about 1000 words.
- **Part II** presents an overview of the use of informetric *indicators for the measurement of research performance*. While Chap. 3 focuses on the multi-dimensional nature of research performance, Chap. 4 presents a list of 28 informetric indicators (or indicator families) that are often used as measures of research performance, and summarizes their pros and cons. Chap. 5 discusses two common misunderstandings, namely that journal impact factors are good predictors of the citation rate of individual articles, and that in large data sets errors or biases are always canceled out. In addition, it dedicates attention to the interpretation of particular often used statistics, namely (linear) correlation coefficients between two variables.
- **Part III** discusses the *application context* of quantitative research assessment. Chapter 6 describes research assessment as an evaluation science. It is in this chapter that the domain of informetrics and the policy sphere are disentangled analytically. The distinctions function as a framework for an outline presented in Chap. 7 and 8 of how external, non-informetric factors influence indicator development. Whilst Chap. 7 discusses the influence of *non- or extra-informetric factors*, relating to evaluative assumptions, the wider social context, and business interests, Chap. 8 focuses on the *policy context*. It introduces the notion of the multi-dimensional research assessment matrix and that of meta-analyses generating background insights for policy makers and evaluators responsible for an assessment.
- **Part IV** presents *the way forward*. It starts in Chap. 9 with a discussion of a series of major problems in the use of informetric indicators in research assessment: the assessment of individual scholars; the use of a limited time horizon; the

assessment of societal impact; the effects of the use of indicators upon authors and editors, and their constitutive effects; and the need for an evaluative framework and an assessment model. The author of this book expresses his views on these problems, and on how they could be dealt with. This chapter forms an introduction to Chaps. 10, 11 and 12 discussing the way forward.

- The way forward in *quantitative research assessment* is the subject of Chap. 10. It presents a list of new features that could be implemented in an assessment process. They highlight the potential of informetric techniques, and illustrate that *current* practices in the use of informetric indicators could be *changed*. Chapter 11 sketches a perspective on *altmetrics*, also termed as ‘alternative’ metrics, but many propositions and suggestions relate to the use in research assessment of any type of informetric indicator. This chapter, as well as Chap. 12 propose new lines in indicator research that put a greater emphasis on *theoretical models*, and on the use of techniques from *computer science* and the newly available *information and communication technologies*. They advocate the development of models of scientific development, and new indicators of the manuscript peer review process, ontology-based data management systems, and informetric author self-assessment tools.
- Part V presents five *lectures* with *historical overviews* of the field of bibliometrics and informetrics, starting from three of the field’s founding fathers: Derek de Solla Price, Eugene Garfield and Francis Narin. It is based on a doctoral course presented by the author at the Sapienza University of Rome in 2015, and on lectures presented at the European Summerschool of Scientometrics (ESSS) during 2010–2016, and in the CWTS Graduate Courses during 2006–2009. Main topics addressed are: citation networks; science mapping; informetric databases; the science-technology interface; journal metrics; and research performance indicators.
- Finally, Part VI presents two full articles published recently by the author of this book on hot topics of general interest in which the use of informetric indicators play a key role. These topics are: A critical comparison of five *world university ranking* systems; and how *usage indicators* based on the number of full text downloads or online reads of research articles compare with citation-based measures. These articles provide background information on the chapters presented in Part IV.

1.6 A Note on Terminology

This book uses the term ‘informetrics’ as a generic term indicating the study of quantitative aspects of information. It comprises all studies denoted as ‘bibliometric’, including the classical publication- and citation-based studies, but it is broader, as it does not merely analyze books and other media of written communication, but also altmetric and usage data, webometric, economic and research input data, and survey

data on scholarly reputation. It does not cover all aspects of informetrics, but those related to research assessment.

The term *assessment* is used as an *overarching* concept, denoting the total of activities in assessment or evaluation processes, or the act of evaluating or assessing *in general*. This book uses the term evaluation exclusively in relation to the setting of criteria for, and the formation of judgments on, the ‘worth’ of a subject. This use is further explained in Chap. 6. It can be said that an assessment process often contains an evaluative component, but there may be assessment without such a component. This book avoids the use of the term ‘evaluation’ as a noun, but uses its adjective form in combination with nouns such as ‘framework’ or ‘criteria’.

This book focuses on *academic* research, primarily intended to increase scholarly knowledge, but often motivated by and funded for specific technological objectives such as the development of new technologies such as medical breakthroughs. It comprises both ‘curiosity-driven’ or ‘pure’ as well as ‘strategic’ or ‘application oriented’ research. The latter type of research may be fundamental in nature, but is undertaken in a quest for a particular application, even though its precise details are not yet known. Indicators (also denoted synonymously as metrics throughout this book) are conceived as instruments used to measure the various components of research activity.

1.7 Re-Use of Paragraphs of Previously Published Articles

Several chapters in this book re-use text fragments from articles published by the author during the past four years (2014–2017). Such re-use, resulting in paragraphs in the book that are very similar in content to text segments in earlier published papers, is indicated in a footnote, acknowledging their source. The current author does not have the intention to plagiarize his own earlier papers, nor to re-publish them, but rather to provide the re-used paragraphs with a new, synthesizing context.

All re-used text fragments were taken from the author copies, published in arXiv, of the following articles. Moed, 2005a (Chap. 5.4); Halevi & Moed, 2014a (Chap. 12.5); Halevi & Moed, 2014b (Chap. 19); Moed & Halevi, 2015 (Chaps. 1.3, 4, 6.5, 6.6 and 8); Daraio, Lenzerini, Leporelli, Moed, Naggar, Bonaccorsi & Bartolucci, 2016 (Chap. 12.3); Moed, 2016a (Chap. 11); Moed, 2016c (Chap. 12.5); Moed, 2016d (Chap. 12.2); Moed, 2016e (Chaps. 7.5, 12.4); Moed & Halevi, 2017 (Chap. 19); Moed, 2017a (Chap. 12.2); Moed, 2017b (Chap. 18).

Chapter 2

Synopsis

Abstract This chapter provides summaries of the main topics and conclusions of each chapter.

Keywords Altmetrics · Architecture of attention · Business interests · Citation-based indicators · Cognitive intelligence · Communication effectiveness · Constitutive effects · Correlation coefficient · Cross-disciplinarity · Economic indicators · Esteem · Evaluation strategy · Evaluative framework · Individual scholars · Informetric arguments · Journal internationality · Journal metrics · Minimum performance standards · Multi-dimensionality · Nobel Prize level · Ontology · Open Science · Patent-based indicators · Peer review · Performance based funding · Policy context · Precondition for performance · Predictor · Publication-based indicators · Random error · Reputation · Research infrastructure · Scientific collaboration · Scientific development · Scientific migration · Self assessment · Self-selection · Size-dependent · Societal impact · Socio-political context · Systemic Time horizon · Unit of assessment · Usage-based indicators · Webometrics · World university ranking

2.1 Part I. Introduction

2.1.1 *Chapter 1*

Chapter 1 highlights five strong points of the use of informetric indicators in research assessment. It helps to demonstrate one's performance; it gives information on shaping one's communication strategies; it offers standardized approaches and independent yardsticks; it delivers comprehensive insights; and provides tools for enlightening policy assumptions.

Five main points of critique are: Indicators may be biased and not measure what they are supposed to measure; most studies adopt a limited time horizon; indicators can be manipulated, and may have constitutive effects; measuring societal impact is

problematic; and when they are used, an evaluative framework and assessment model are often lacking.

The chapter describes a series of trends during the past decade in the domain of science policy: an increasing emphasis on societal value and value for money, performance-based funding and on globalization of academic research, and a growing need for internal research assessment and research information systems.

Due to the computerization of the research process and the digitization of scholarly communication, research assessment is more and more becoming a ‘big data’ activity, involving multiple comprehensive citation indexes, electronic full text databases, large publications repositories, usage data from publishers’ sites, and altmetric, webometric and other new data sources.

The above trends created an increasing interest in the development, availability and application of new indicators for research assessment. Many new indicators were developed, and more and more measures have become available on a large scale. Desktop bibliometrics is becoming a common assessment practice.

Two principal assumptions of the author are as follows.

- From what is cannot be inferred what ought to be. Evaluation criteria and policy objectives are not informetrically demonstrable values. Informetric research may study such values empirically, but cannot provide a theoretical foundation of the validity of the quality criteria or the appropriateness of policy objectives. Informetricians should in their informetric work maintain a neutral position towards these values.
- The future of research assessment lies in the intelligent combination of indicators and peer review. From the very beginning, bibliometric indicators stimulated peers to make the foundation and of their judgments more explicit. The notion of informetric indicators as a support tool in peer review processes rather than as a replacement of such processes still has a great potential, and this book aims to further explore it. A necessary condition for achieving this is a thorough awareness of the potentialities and limitations of both methodologies.

The main subjects of the book are:

- An overview of important new databases, methodologies, indicators and products introduced during the past 10 years in the field of informetrics and its application in research assessment.
- A comprehensive overview of the most important indicators or indicator families, the aspects or dimensions they are assumed to measure, and their potential and limits.
- A clarification of the relationship between the informetric, the evaluative and the policy domain.
- Possible new features that could be implemented in a research assessment process.
- New lines of research that are expected to lead to the development of new, useful indicators.

This book uses the term '*informetrics*' as a generic term indicating the study of quantitative aspects of information. It does not only analyze bibliometric data based on publication and citation counts, but also altmetric and usage data, webometric data, economic data, research input data, and survey data on scholarly reputation.

The term '*assessment*' is used as an *overarching* concept, denoting the total of activities in assessment or evaluation processes, or the act of evaluating or assessing *in general*. This book uses the term evaluation exclusively in relation to the setting of criteria for, and the formation of judgments on, the 'worth' of a subject.

This book focuses on *academic* research, primarily intended to increase scholarly knowledge, but often motivated by and funded for specific technological objectives such as the development of new technologies such as medical breakthroughs. It comprises both 'curiosity-driven' or 'pure' as well as 'strategic' or 'application oriented' research.

2.2 Part II. Informetric Indicators of Research Performance

2.2.1 Chapter 3

The multi-dimensional nature of research performance is highlighted. Four main components of research activity are distinguished: input, process, output and impact. *Input* includes funding, manpower and research infrastructure. Indicators of research infrastructure are not primarily performance indicators but relate rather to a *precondition* for performance. *Process* indicators focus on research collaboration and efficiency.

Output can be publication based, such as a journal article or monograph, or be delivered in a non-publication format, such as research dataset, and be directed to the scientific-scholarly community or to society and the wider public. A key distinction is made between scientific-scholarly and societal *impact*. Societal impact embraces a wide spectrum of aspects outside the domain of science and scholarship itself, including technological, social, economic, educational and cultural aspects.

Research performance is reflected in *all four* components. For instance, research funding, especially competitive funding, can be assumed to depend on the past performance and the reputation of the grant applicant, and therefore relates both to input and to impact. Efficiency is both a process and an impact indicator, as it aims to measure output or impact per unit of input. A table is presented that lists the 28 important indicators or indicator families and their pros and cons.

2.2.2 Chapter 4

Chapter 4 presents the main characteristics of the most important indicator families.

- *Publication-based indicators.* In academic institutions, publications in all scientific-scholarly subject fields constitute an important format of academic output. However, in the private sector, and also in academic departments with a strong applied orientation, which aim primarily to produce new products or processes, publishing for the general public, peer reviewed literature often does not have the highest priority; in this case, other output forms must be considered as well. Publication counts may be used to define minimum performance standards.
- *Citation-based indicators.* Citation analysis offers a certain degree of standardization, and compares units of assessment against an independent yardstick, which makes an evaluator more independent from the views of the subjects of the analysis and of the commissioning entity. Citations can be interpreted as proxies of more direct measurements of intellectual influence, but they are by no means indicators of the validity of a knowledge claim. Citation impact and quality are not identical concepts. As all indicators, citations be affected by disturbing factors and suffer from serious biases.
- *Journal metrics.* The quality or impact of the journals in which a unit under assessment has published is a performance aspect in its own right. But the relationship between a journal's impact factor and the rigorousness of its manuscript peer review process is unclear. Journal metrics cannot be used as a surrogate of actual citation impact; they are no good predictors of the citation rate of individual papers. Moreover, their values can to some extent be manipulated, and may be affected by editorial policies.
- *Patent-based indicators.* Analyses of inventors of the findings described in patents reveal the extent to which scientists with academic positions contributed to technological developments. Patent-to-patent citations may reveal a patent's technological value, and patent citations to scientific papers a technology's science base. But the propensity to apply for patents differs across countries because of legislation or culture, and also across subject fields. In addition, patents are a very poor indicator of the commercialization of research results.
- *Usage-based indicators.* Data on downloads of an electronic publication in html or full text format enable researchers to assess the effectiveness of their communication strategies, and may reveal attention of scholarly audiences from other research domains or of non-scholarly audiences. Downloaded articles may be selected according to their face value rather than their value perceived after reflection. Also, there is an incomplete data availability across providers, and counts can be manipulated. It is difficult to ascertain whether downloaded publications were actually read or used.
- *Altmetrics* relates to different types of data sources with different functions. Mentions in social media may reveal impact upon non-scholarly audiences, and provide tools to link scientific expertise to societal needs, but cannot be used to measure scientific-scholarly impact. Their numbers can be manipulated, and interdependence of the various social media may boost figures. Readership counts in scholarly reference managers are potentially faster predictors of emerging scholarly trends than citations are, but results depend upon readers' cognitive and professional background.

- *Webometrics.* Indicators of Web presence and impact are extracted from a huge universe of documents available on the Web. They do not merely relate to institutions' research mission, but also to their teaching and social service activity. But there is no systematic information on the universe of Web sources covered and their quality; and an institution's Web presence depends upon its internal policies towards the use of the Web, and upon the propensity of its staff to communicate via the Web.
- *Economic indicators.* Indicators of economic value and efficiency are relevant measures in research assessment. But not all contributions to economic development can be easily measured. The relationship between input and output is not necessarily linear, and may involve a time delay. Accurate, standardized input data is often unavailable; and comparisons across countries are difficult to make, due to differences in the classification of economic data. Indicators of funding from industry are useful measures of economic value; but funding levels differ across disciplines, and data may be difficult to collect.
- *Reputation and esteem based measures.* Receiving a prestigious prize or award is a clear manifestation of esteem. But absolute numbers tend to be low; the evaluation processes on which the nominations are based are not always fully transparent; and there are no agreed equivalences that apply internationally and facilitate comparison across disciplines. Reputation can be measured in surveys using validated methods from social sciences. But response rates are often very low; mentions may be based on 'hear-say' rather than on founded judgement and may refer to performance made in a distant past.
- *Measures of scientific collaboration, migration and cross-disciplinarity.* Data on co-authorship and on how authors migrate over time from one institutional affiliations to another, provide useful indicators of intra-institutional, national, international scientific collaboration and migration. But instances that have not resulted in publications remain bibliometrically invisible. Indicators of cross-disciplinary measure the relevance of a piece of research for surrounding disciplines, or the cognitive breadth of research impact. Their calculation presupposes a valid, operational classification of research into disciplines.
- *Indicators of research infrastructure* are not primarily performance indicators, but focus on preconditions for performance. They measure basic facilities that support research, the scale of the research activities, and their sustainability. But research practices differ across disciplines; large research teams or laboratories are mostly found in the natural and life sciences. It is difficult to obtain reliable, comparable institutional data, as there is no agreement on the basis of a full cost calculation of research investment. There is no clear, generally accepted definition of being research active across universities, countries and disciplines.

The calculation of informetric indicators of research performance more and more becomes a '*big data*' activity. Not only the increasing volume of informetric datasets is of interest, but especially their *combination* creates a large number of new possibilities. For instance, the combination on an article-by-article basis of citation indexes and usage log files of full text publication archives, enables one to

investigate the relationships between downloads and citations, and develop ways to generate a more comprehensive, multi-dimensional view of the impact of publications than each of the sources can achieve individually.

There is an increasing interest in mapping techniques, and science mapping is to be qualified as one of the most important domains of informetrics as a big data science. It can be defined as the development and application of computational techniques for the visualization, analysis, and modeling of a broad range of scientific and technological activities as a whole.

2.2.3 *Chapter 5*

Although this book does not present details on the technical and statistical aspects of most informetric indicators, Chap. 5 does discuss three common misunderstandings as regards interpretation of particular, often used, statistical measures or techniques, related to journal impact factors as means of skewed distributions, errors in large datasets, and the interpretation of linear correlation coefficients. The conclusions are as follows.

- Journal impact factors are *no* good predictors of the citation rate of individual articles in a journal.
- Only random errors tend to cancel out in large datasets; systematic biases may remain.
- When interpreting a correlation coefficient, never rely merely on its numerical value. Consider always a scatter plot of the underlying data.

2.3 Part III. The Application Context

2.3.1 *Chapter 6*

In Chapter 6 an analytical distinction is made between *four* domains of intellectual activity in an assessment process, including the following activities.

- *Policy or management:* The formulation of a policy issue and assessment objectives; making *decisions* on the assessment's organizational aspects and budget. Its main outcome is a policy decision based on the outcomes from the evaluation domain.
- *Evaluation:* A specification of the evaluative framework, i.e., a set of evaluation criteria in agreement with the constituent policy issue and assessment objectives. The main outcome is a *judgment* on the basis of the evaluative framework and the empirical evidence collected.

- *Analytics.* Collecting, analyzing, and reporting *empirical* knowledge on the subjects of assessment; The specification of an assessment *model or strategy*, and the *operationalization* of the criteria in the evaluative framework. Its main outcome is an analytical report as input for the evaluative domain.
- *Data collection.* Collection of relevant data for analytical purposes, as specified in the analytical model. Data can be either quantitative or qualitative. Its main outcome is a dataset for the calculation of all indicators specified in the analytical model.

A main objective of this analytical categorization is to distinguish between scientific-methodological principles and considerations on the one hand, and policy-related, political or managerial considerations on the other. Focusing on the role of informetricians, the chapter argues as follows.

- What is of worth, good, desirable, or important in relation to the functioning of a subject under assessment, *cannot* be established in informetric, or, more general, in quantitative-empirical research. The principal reason is that one cannot infer what *ought to be* from what actually *is*.
- What informetric investigators *can* do is empirically examine value *perceptions* of researchers, the conditions under which they were formed and the functions they fulfil, but the foundation of the validity of a value is *not* a task of quantitative-empirical, informetric research. The same is true for the formation of *evaluative judgements*. The latter two activities belong to the domain of evaluation.
- Informetricians should maintain a *methodologically* neutral position towards the constituent policy issue, the criteria specified in the evaluative framework, and the goals and objectives of the assessed subject. As professional experts, their competence lies *primarily* in the development and application of analytical models *given* the established evaluative framework.
- Obviously, informetricians are allowed to form and express ‘normative’ views while assessing a unit’s worth, but when doing so they should make these explicit and give them methodologically speaking a hypothetical status.

Several types of assessment models are distinguished: peer review based versus indicator based assessments, and self-assessment versus external assessment. It distinguishes four classes of *evaluation strategies*:

- Scientific-experimental focusing on impartiality and objectivity;
- Management-oriented systems based on systems theory;
- Qualitative anthropological approaches emphasizing the importance of observation and space for subjective judgement;
- Participant oriented strategies, giving a central role to ‘consumers’.

A genuine challenge is to combine the various models and create hybrid assessment models.

2.3.2 Chapter 7

Chapter 7 illuminates the influence of non-informetric or extra-informetric factors on the development of indicators, and in this way aims to *disentangle* informetric arguments and evaluative principles, one of the key objectives of this book. Typical examples are given as regards the following issues: size dependent versus size independent indicators; focus on the top or the bottom of a performance distribution; indicator normalizations; definition of benchmark sets; and the application of a short term or a long term perspective.

For instance, a series of citation impact indicators seek an ‘optimal’ combination of publication and citation counts, and address the issue whether this optimum should be size-normalized or not. Under the surface of this seemingly technical debate, a confrontation takes place of distinct views of what constitutes genuine research performance or ‘quality’.

- According to one view, a citation-per-paper ratio is the best indicator, because it helps to detect ‘saturation’, which occurs if a research group increases its annual number of published papers while the citation impact per paper declines.
- A second view holds that such a ratio penalizes groups with a large publication output, while a large publication output should be rewarded rather than penalized.
- A third view aims to reduce the role of absolute publication numbers by proposing an indicator counting only the ‘best’ papers in terms of citation impact.
- A fourth view claims that the only good performance indicator is an efficiency indicator dividing output or impact by ‘input’ measures.

It illustrates how in seemingly technical discussions on the construction and statistical properties of science indicators, *‘evaluative’, theoretical assumptions on what constitutes research performance* play an important, though often implicit, role. Such values are denoted as extra-informetric, as their validity cannot be grounded in informetric research.

Chapter 7 also presents a brief history of some of the main lines in bibliometric indicator development from the early 1960s up to date. It focuses on the *wider socio-political context* in which indicators were developed. It describes the context of their launch, not so much in terms of the *intentions* of the developers, but, at a higher analytical level, in terms of how they fit into – or are the expression of – a more general tendency in the policy, political or cultural environment in which they were developed. A base assumption is that knowledge of the wider context in which specific indicators were developed contributes to a better understanding of how and under which conditions they can be properly applied.

It is argued that in the early decades newly proposed indicators primarily aimed to solve specific policy problems and fitted in with specific national or institutional policy contexts, but during the past 10–15 years the following two tendencies emerged: on the one hand, previously developed indicators were used in more and

more policy contexts, including contexts in which they are only partially or hardly fit-for-purpose; on the other hand, indicator development was more and more driven by an internal dynamics powered by mathematical-statistical considerations.

Finally, Chap. 7 discusses the influence of *business interests* of the information industry upon the development of indicators. It concludes that since Eugene Garfield introduced the Journal Impact Factor as an ‘objective’ tool to expand the journal coverage of his citation index independently of journal publishers, the landscape of scientific information providers and users has changed significantly. While, on the one hand, politicians and research managers at various institutional levels need valid and reliable fit-for-purpose metrics in the assessment of publicly funded research, there is, on the other hand, a tendency that indicators increasingly become a tool in the business strategy of companies with product portfolios that may include underlying databases, social networking sites, or even indicator products. This may be true both for ‘classical’ bibliometric indicators and for alternative metrics.

2.3.3 *Chapter 8*

Chapter 8 argues that in the design of a research assessment process, one has to decide which methodology should be used, which indicators to calculate, and which data to collect. To make proper decisions about these matters, one should address the following key questions, each of which relates to a particular aspect of the research assessment process.

- What is the unit of the assessment? A country, an institution, a research group, an individual, or a research field or an international network? In which discipline(s) is it active?
- Which dimension of the research process must be assessed? Scientific-scholarly impact? Social benefit? Multi-disciplinarity? Participation in international networks?
- What are the purpose and the objectives of the assessment? Allocate funding? Improve performance? Increase regional engagement?
- What are relevant, general or ‘systemic’ characteristics of the units of assessment? For instance, to which extent are they oriented towards the international research front?

The answers to these question determine which indicators are the most appropriate in a particular assessment process. Indicators that are useful in one context, may be less so in another. A warning is issued against a practice in which particular indicators are used in a given context simply because they are available, and because they have been successfully applied in *other* contexts.

Knowledge on general characteristics of the system of units of assessment plays an important role in the formulation of the objectives of an assessment. Such

assumptions do not focus on *individual* units, but on more general or systemic characteristics of these units *as a group*. Therefore, they can be denoted as ‘meta’ assumptions, and illuminate the assessment’s *policy context*.

For instance, if an analysis of the state of a country’s science system provides evidence that researchers tend to publish mainly in national journals without a serious manuscript peer review process, it is from an informetric viewpoint *defensible* to use the number of publications in the top quartile of journals in terms of citation impact as an indicator of research performance, not so much as an *evaluation tool*, but rather as an instrument to *change* certain communication *practices* among researchers.

However, if in internationally oriented, leading universities one has to assess candidates submitting their job application, it is questionable whether it makes sense comparing applicants according to the average citation impact of the journals in which they published their papers. Due to *self-selection*, the applicants will probably publish at least a large part of the papers in good, international journals, and in this group journal impact factors hardly discriminate between high and lower performance.

2.4 Part IV. The Way Forward

2.4.1 Chapter 9

This chapter discusses a series of problems in the use of informetric indicators for evaluative purposes. Its main conclusions are as follows. Their implications for the application of indicators and for future indicator development are further discussed in Chaps. 10–12.

- *The problem of assessing individual scholars.* Calculating indicators at the level of an individual and claiming they measure *by themselves* the individual’s performance suggests a façade of exactness that cannot be justified. A valid and fair assessment of individual research performance can be conducted properly only on the basis of sufficient background knowledge on the particular role they played in the research presented in their publications, and by taking into account also other types on information on their performance.
- *The effect of a limited time horizon.* The notion of making a contribution to scientific-scholarly progress, does have a basis in reality, that can best be illustrated by referring to an *historical* viewpoint. *History will show* which contributions to scholarly knowledge are valuable and sustainable. In this sense, informetric indicators do *not* measure contribution to scientific-scholarly progress, but rather indicate attention, visibility or short term impact.
- *The problem of assessing societal impact.* Societal value cannot be assessed in a politically neutral manner. The foundation of the criteria for assessing societal value is not a matter in which scientific experts have *qualitate qua* a preferred

status, but should eventually take place in the policy domain. One possible option is moving away from the objective to evaluate an activity's societal *value*, towards measuring in a neutral manner researchers' *orientation* towards any articulated, lawful need in society, as reflected for instance in *professional contacts*. Due to time delays, emphasis on societal impact on the one hand, and assessment focus on *recent* past performance on the other, are at least partially conflicting policy incentives.

- *The effects of the use of indicators upon authors and editors.* Studies on changes in editorial and author practices under the influence of assessment exercises are most relevant and illuminative. The issue at stake is *not* whether scholars' practices *change* under the influence of the use of informetric indicators, but rather whether or not the application of such measures enhances their *research performance*. Although this is in some cases difficult to assess without extra study, other cases clearly show traces of mere indicator manipulation with no positive effect on performance at all.
- *How to deal with constitutive effects of indicators.* A typical example of a constitute effect is when research quality is more and more perceived as what citation measures. If the tendency to replace reality with symbols and to conceive these symbols as an even a higher form of reality, are typical characteristics of *magical thinking*, jointly with the belief to be able to change reality by acting upon the symbol, one could rightly argue that the un-reflected, unconditional belief in indicators shows rather strong similarities with *magical thinking*.
- More empirical research on the size of constitutive effects is urgently needed. If there is a genuine constitutive effect of informetric indicators in quality assessment at all, one should not point the critique on current assessment practices merely towards informetric indicators as such, but rather towards any claim for an *absolute status* of a particular *way* to assess quality. If the role of informetric indicators has become too dominant, it does *not* follow that the idea to intelligently combine peer judgements and indicators is fundamentally flawed and that indicators should be banned from the assessment arena. But it does show the combination of the two methodologies has to be organized in a more sophisticated and balanced manner.
- *The need for an evaluative framework and an assessment model.* Chapter 6 underlines the need to define an evaluative framework and an assessment model. To the extent that in a practical application an evaluative framework is absent or implicit, there is a vacuum, that may be easily filled either with ad-hoc arguments of evaluators and policy makers, or with un-reflected assumptions underlying informetric tools. Perhaps the role of such ad hoc arguments and assumptions has nowadays become too dominant. It can be reduced only if evaluative frameworks become stronger, and more actively determine which tools are to be used, and how. To facilitate this, informetricians should make the normative assumptions of their tools explicit, and inform policy makers and evaluators about the potential and the limits of these tools.

2.4.2 Chapter 10

Chapter 10 critically reflects on the assumptions underlying current practices in the use of informetric indicators in research assessment, and proposes a series of alternative approaches, indicating their pros and cons.

Communication Effectiveness as a Precondition for Performance

In academic institutions, especially in research universities, it is considered appropriate to stimulate academic researchers to make a solid contribution to scientific-scholarly progress. But is it defensible to require that they generate impact? What should be of primary interest to academic policy makers: importance (potential influence) or impact (actual influence)? An academic assessment policy is conceivable that rejects the claim that impact rather than importance is the key aspect to be assessed, and discourages the use of citation data as a *principal* indicator of importance.

Such an assessment process would *not* aim at measuring importance or *contribution to scientific-scholarly progress* as such, but rather *communication effectiveness*, a concept that relates to a *precondition for performance* rather than to performance itself. It expresses the extent to which researchers bring their work to the attention of a broad, potentially interested audience, and can in principle be measured with informetric tools.

Some New Indicators of Multi-dimensional Research Output

The *functions* of publications and other forms of scientific-scholarly output, as well as their *target audiences* should be taken into account more explicitly than they have been in the past. Journals with an educative or enlightening function are important in scientific scholarly communication, and tend to have a substantial societal value. Since their visibility at the international research front as reflected in citations and journal impact factors may be low, in a standard bibliometric analysis based on publication and citation counts this value may not be visible.

Scientific-scholarly journals could be systematically categorized according to their function and target audience, and separate indicators could be calculated for each category. In an analysis of research output in journals directed towards *national* audiences, citation-based indicators are *less relevant*. At the same time, in citation analyses based on the large international citation indexes focusing on the international research front, it would be appropriate to disregard such journals. It is proposed to develop indicators of *journal internationality* based on the geographical distribution of publishing, citing or cited authors. A case study shows that journal impact and internationality are by no means identical concepts. Whether or not a journal is indexed in one or more of the large citation indexes is not in all cases a good indicator of its international orientation.

Definition of Minimum Performance Standards

One possible approach to the use of informetric indicators in research assessment is a systematic exploration of indicators as tools to set minimum performance standards and define in this way a performance baseline. Important considerations in favor of this approach are as follows.

- There is evidence that citation rates are a good predictor of how peers discriminate between a ‘valuable’ and a ‘less valuable’ past performance, but that they do not properly predict within the class of ‘valuable’ performances, peers’ perception of ‘genuine excellence’. This outcome underlines the potential of informetric indicators in the assessment of the *lower part* of the quality distribution.
- Using indicators to define a baseline, researchers will most probably change their research practices as they are stimulated to meet the standards, but if the standards are appropriate and fair, this behavior will actually increase their performance and that of their institutions.
- Focusing on minimum criteria involves a shift in perspective from measuring performance *as such* towards assessing the *preconditions* for performance. Expert opinion and background knowledge will play a crucial role, not only in the definition of the standards themselves, but also in the assessment processes in which these are applied.
- The definition of minimum standards could also be applied to journal impact measures. Rather than focusing on the most highly cited journals and rewarding publications in this top set, it would be possible to discourage publication in the bottom set of journals (e.g., the bottom quartile) with the lowest citation impact.

At the *upper part* of the quality distribution, it is perhaps feasible to distinguish entities which are ‘*hors catégorie*’, or ‘*at Nobel Prize level*’. Assessment processes focusing on the *top* of the quality distributions could further operationalize the criteria for this qualification.

Policy Towards World University Rankings

Chapter 18 in *Part VI* of this book presents a comparative analysis of 5 World University Ranking Systems. Realistically speaking, rankings of world universities are here to stay. When university managers use their institution’s position in world university rankings primarily for marketing purposes, they should not disregard the negative effects such use may have upon researchers’ practices within their institution. They should also critically address the validity of the methodological claims made by producers of these ranking systems.

The following strategy towards these ranking systems is proposed. Academic institutions could, individually or collectively, seek to influence the various systems by formally sending them a request to consider the implementation of the following new features.

- Offer more advanced analytical tools, enabling a user to analyze the data in a more sophisticated manner than ranking systems currently offer.
- Provide more insight into how the methodological decisions of the producers influence the ranking positions of given universities.
- Enhance the information in the system about additional factors, such as teaching course language.

In addition, academic institutions could proceed as follows.

- Create a special university webpage providing information on a university's internal assessment and funding policies, and on its various types of performance, and giving comments on the methodologies and outcomes of ranking systems.
- Request ranking producers to make these pages directly accessible via their systems.

An Alternative Approach to Performance Based Funding

Major criticisms towards national research assessment exercises underline their bureaucratic burden, costs, lack of transparency, and their Matthew effect. Adopting an informetric viewpoint, an alternative assessment model is described that would require less efforts, be more transparent, stimulate new research lines, and reduce to some extent the Matthew Effect. The main features are as follows.

- The base unit of assessment is an *emerging group*, a small research group with a great scientific potential. Acknowledged as a 'rising star', the director has developed a promising research program, and has already been able to establish a small research unit.
- The profile of an emerging group should be further operationalized into a set of minimum quantitative criteria, taking into account the communication practices and funding opportunities in the group's subject field.
- In the assessment procedure, institutions submit groups rather than individual staff. Submissions provide information on the group's past performance and a future research programme, that are evaluated in a peer review process, informed by appropriate informetric indicators.
- The primary aim of the peer review is to define the minimum standards in operational terms and assess whether the submitted groups comply with these standards. These standards constitute a precondition for the group's future performance.
- There would be no need to rank groups, assign ratings, discriminate between 'top' and 'almost top' groups, or make funding decisions. Funding decisions take place within their institution.
- To stimulate the implementation of quality control processes within an institution, the availability of a certain amount of funding from internal,

performance-based allocation processes could be posed as a necessary condition.

- A part of public funding (block grant) could be allocated to institutions as a lump sum on the basis of the number of acknowledged emerging groups.

The practical realization of these proposals requires a large amount of informetric research and development. They constitute important elements of a wider R&D program in *applied evaluative informetrics*. These activities tend to have an applied character and often a short term perspective, and focus on the development side of R&D. They should be conducted in a close collaboration between informetricians and external stakeholders. Chapters 11 and 12 propose strategic, longer term research projects with a great potential for research assessment.

2.4.3 *Chapter 11*

Chapter 11 discusses the potential of altmetrics. A multi-dimensional conception of altmetrics is proposed, namely as traces of the computerization of the research process, and conceived as a tool for the practical realization of the ethos of science and scholarship in a computerized or digital age. Three drivers of development of the field of altmetrics are distinguished.

- In the policy domain: An increasing awareness of the multi-dimensionality of research performance, and an emphasis on societal merit.
- In the domain of technology: The development of information and communication technologies (ICTs), especially social media.
- From the scientific-scholarly community itself: The Open Science movement.

Four aspects of the computerization of the research process are highlighted: the computerization of research data collection and analysis; the digitization of scientific information; the use of computerized communication technologies; and informetritization of research assessment.

Michael Nielsen's set of creative ideas constitute a framework in which altmetrics can be positioned. He argued that "to amplify cognitive intelligence, we should scale up collaborations, increasing cognitive diversity and the range of available expertise as much as possible". The role of altmetrics and other informetric indicators would not merely be, passively, to provide descriptors, but also actively, or proactively, to establish and optimize, Nielsen's "architecture of attention", a configuration that combines the efforts of researchers and technicians on the one hand, and the wider public and the policy domain on the other, and that "directs each participant's attention where it is best suited—i.e., where they have maximal competitive advantage".

It should not be overlooked that a series of distinctions made in 'classical' research assessment are most relevant in connection with altmetrics as well: scientific-scholarly versus societal impact; attention versus influence; opinion

versus scientific fact; peer-reviewed versus non-peer reviewed; intended or unintended versus constitutive effects of indicators.

2.4.4 *Chapter 12*

Chapter 12 proposes a series of alternative approaches to the development of new informetric indicators for research assessment that put a greater emphasis on *theoretical models* for the interpretation of informetric indicators, and on the use of techniques from *computer science* and the newly available *information and communication technologies*.

Towards New Indicators of the Manuscript Peer Review Process

A proposal to develop new indicators of the manuscript peer review process is based on the following considerations.

- Manuscript peer review is considered important by journal publishers, editors and researchers. But the process itself is still strikingly opaque.
- Reviewers tend to receive little training in what is one of the key academic activities, and there is little evidence of any standardization in how review reports are composed.
- There is little systematic, objective information on the quality of the process across journals and subjects, and on its effect upon the quality of submitted papers.
- With peer review largely a black box, proxies for its quality have grown up, most notably the journal impact factor (JIF) based on citation counts.
- But the digitization of scientific information offers great potential for the development of tools to allow peer review to be analyzed directly.

The objectives and set-up of the project are as follows.

- The aim of the project is to build up an understanding for each discipline of what is considered a reasonable quality threshold for publication, how it differs among journals and disciplines, and what distinguishes an acceptable paper from one that is rejected.
- The analysis consists of two phases, an explorative phase in which a classical-humanities approach is dominant, aimed to develop a conceptual model; and a data mining phase, applying techniques from digital humanities.
- Taking into account a journal's scope and instructions to reviewers, the various elements of a review report should be analyzed. Statements are categorized in terms of aspect and modality. Standards and assumptions applied by a reviewer are identified.

- Relevant concepts and their indicators are developed, including the formative content of a review report, and the distance a reviewer maintains towards his own methodological and theoretical views.

The project could have the following outcomes.

- It provides insight into the effects and ‘added value’ of peer review upon manuscript quality.
- It defines a set of minimum quality standards per journal and per discipline. This information enhances the transparency of the review process, and is useful for editors, reviewers and authors.
- It proposes indicators characterizing the various aspects of the process, for instance, its formative effect, and other tools to monitor the process.
- The results may be used to improve the quality of the peer review process.
- Ultimately, perhaps journal-level metrics can be validated that can supersede proxies such as journal impact factors.

Towards an Ontology-Based Informetric Data Management System

During the past decades, development and application of informetric indicators in research assessment have posed a series of challenges to the management—collection, handling, integration, analysis and maintenance—of informetric data, and in the design of S&T indicators. There are data-related, concept-related and maintenance-related issues.

To solve these issues, it is proposed to develop an Ontology-Based Data Management (OBDM) system for research assessment, along the lines set out in Daraio et al (2016). The key idea of OBDM is to create a three-level architecture, constituted by a) the ontology; b) the data sources; and c) the mapping between the two.

An *ontology* can be defined as a conceptual, formal description of the domain of interest to a given entity (e.g., organization or community of users), expressed in terms of relevant concepts, *attributes* of concepts, *relationships* between concepts, and *logical assertions* characterizing the domain knowledge. The *sources* are the data repositories accessible by the organization in which data concerning the domain are stored. The mapping is a precise specification of the correspondence between the data contained in the *data sources* on the one hand, and the elements of the *ontology* on the other.

The main advantages of an OBDM approach are as follows

- Users can access the data by using the elements of the ontology. A strict separation exists between the conceptual and the logical-physical level.
- By making the representation of the domain explicit, the acquired knowledge can be easily re-used.

- The mapping layer explicitly specifies the relationships between the domain concepts in the ontology and the data sources. It is useful also for documentation and standardization purposes.
- The system is more flexible. It is for instance not necessary to merge and integrate all the data sources at once, which could be extremely time consuming and costly.
- The system can be more easily extended. New elements in the ontology or data sources can be added incrementally when they become available. In this sense, the system is dynamical and develops over time.

Towards Informetric Self-assessment Tools

The creation of an informetric self-assessment tool at the level of individual authors or small research groups is proposed. Such an application would be highly useful, but is currently unavailable. A challenge is to make an optimal use of the potentialities of the current information and communication technologies and create an online application based on key notions expressed decades ago by Eugene Garfield about author benchmarking, and by Robert K. Merton about the formation of a *reference group*.

The general concept is as follows.

- In a first step, the application enables an author to define a set of publications he/she wishes to take into account in the assessment. It is important that there is a proper data verification tool at hand.
- In a next step, a benchmark set is created of authors with whom the assessed author can best be compared, along the lines adopted by Eugene Garfield in his proposal for an algorithm to create for a given author under assessment, a set of ‘candidate’ benchmark authors who have bibliometric characteristics that are similar to those of the given author.
- Garfield’s idea could be further developed by creating a flexible benchmarking feature as the practical realization of Robert K. Merton’s notion of a reference group, i.e., the group with which individuals compare themselves, but to which they do not necessarily belong but aspire to.
- The calculated indicators should be the result of simple statistical operations on absolute numbers. Not only the outcome, but also the underlying numbers themselves should be visible.
- In addition, researchers must have the opportunity to decompose and reconstruct indicators. It should also be possible to insert particular data manually.

An adequate assessment of individual research performance can take place only by taking into account multiple sources of information about their performance. This does not mean that bibliometric measures in the assessment of individuals are irrelevant, especially when used as self-assessment tools. The proposed tool would be useful for the following reasons.

- In their self-assessments, researchers may wish to calculate specific fit-for-purpose indicators that are not ‘standard’ and therefore unavailable at the websites of indicator producers. The proposed self-assessment tool could be a genuine alternative to using journal impact factors or h-indices.
- Even if one is against the use of informetric indicators in individual assessments, one cannot ignore their availability to a wider public. Therefore, it would be useful if researchers had an online application to check the indicator data calculated about them, and to decompose the indicators’ values.
- In this way they would learn more about the ins and outs of evaluative informetrics, and, for instance, become aware of how the outcomes of an assessment depend upon the way benchmark sets are being defined. This would enable researchers to defend themselves against inaccurate calculation or invalid interpretation of indicators.

Towards Informetric Models of Scientific Development

Scientifically developing countries need tools to monitor the effectiveness of their research policies in a framework that categorizes national research systems in terms of *the phase* of their scientific development. Leaving out the dynamical aspects of a system gives an incomplete picture. The current section presents a model of a country’s scientific development using bibliometric indicators based on publications in international, peer-reviewed journals.

The model aims to provide a framework in which the use of informetric indicators of developing countries makes sense, as an alternative to common bibliometric rankings based on publication and citation counts from which the only signal is that such countries tend to feature in the bottom.

A simplified and experimental bibliometric model for different phases of development of a national research system distinguishes four phases:

- *Pre-development phase.* Low research activity without clear policy of structural funding of research
- *Building up.* Collaborations with developed countries are established; national researchers enter international scientific networks
- *Consolidation and expansion.* The country develops its own infrastructure; the amount of funds available for research increases
- *Internationalization.* Research institutions in the country start as fully-fledged partners, increasingly taking the lead in international collaboration

The distinction into phases is purely *analytical* and does not imply a *chronological order*. The model assumes that during the various phases of a country’s scientific development, the number of published articles in peer-reviewed journals shows a more or less continuous increase, although the rate of increase may vary substantially over the years and between countries.

It is the share of a country's internationally co-authored articles that discriminates between the various phases in the development. The model also illustrates the *ambiguity* of this indicator, as a high percentage of internationally co-authored papers at a certain point in time may indicate that a country is either in the building up or the internationalization phase.

The model is applied to empirical data on South-East Asian countries, and on Arab Gulf states and neighboring countries in the Middle East, and provided evidence, for instance, that while Saudi Arabia is in the building-up phase, Iran has already entered the internationalization phase.

2.5 Part V. Lectures

2.5.1 *Chapter 13*

Chapter 13 presents two visionary papers published by Derek de Solla Price, the founding father of the science of science. It presents his view on the scientific literature as a network of scientific papers, and introduces important informetric concepts, including research front and immediacy effect. Next, the chapter shows how his pioneering work on modelling the relational structure of subject space evolved into the creation of a series of currently available, advanced science mapping tools.

2.5.2 *Chapter 14*

A comparative analysis of three big, multi-disciplinary citation indexes is presented in Chap. 14. It starts with a presentation of the basic principles of the Science Citation Index (SCI, later Thomson Reuters' Web of Science, currently Clarivate Analytics), a multi-disciplinary citation index created by Eugene Garfield in the early 1960s. Next, it presents a study conducted in 2009 comparing the Web of Science with Scopus, a comprehensive citation index launched by Elsevier in 2004, and a recent study comparing Scopus with an even more comprehensive citation index, Google Scholar, also launched in 2004.

2.5.3 *Chapter 15*

Chapter 15 discusses studies on the relationship between science and technology. It starts with presenting the pioneering work by Francis Narin and co-workings on the citation analysis of the linkage between science and technology. Next, it discusses

several theoretical models of the relationship between science and technology. As an illustration of an analysis of the development of a technological field, it presents key findings from a study on industrial robots. The chapter ends with illustrating the limitations of citation analysis of the scientific literature for the measurement of technological performance.

2.5.4 *Chapter 16*

Chapter 16 deals with journal metrics. It starts with a discussion of the journal impact factor, probably the most well-known bibliometric measure. It shows some of its technical limitations and dedicates in an analysis of editorial self-citations special attention to its sensitivity to manipulation. Next, a series of alternative journal citation measures is presented, SJR, Eigenfactor, SNIP, CiteScore, and indicators based on usage.

2.5.5 *Chapter 17*

Definitions and properties of a series of informetric indicators discussed in earlier chapters of this book, are presented in Chap. 17: relative citation rates, h-index, Integrated Impact Indicator, usage-based indicators, social media mentions, and research efficiency or productivity measures. It highlights their potential and limits, and gives typical examples of their application in research assessment.

2.6 Part VI. Papers

2.6.1 *Chapter 18*

To provide users insight into the value and limits of world university rankings, Chap. 18 presents a comparative analysis of 5 World University Ranking Systems: ARWU, the Academic Ranking of World Universities, also indicated as ‘Shanghai Ranking’; The Leiden Ranking created by the Centre for Science and Technology Studies (CWTS); The Times Higher Education (THE) World University Ranking; QS World University Rankings; and an information system denoted as U-Multirank created by a consortium of European research groups.

As all ranking systems claim to measure essentially academic excellence, one would expect to find a substantial degree of consistency among them. The overarching issue addressed in this Chap. 6 is the assessment of this consistency-between-systems. To the extent that a lack of consistency is found,

what are the main causes of the observed discrepancies? A series of analyses is presented, from which the following conclusions were drawn.

- Each ranking system has its proper orientation or ‘profile’; there is no ‘perfect’ system. There is only a limited overlap between the top 100 segments of the 5 rankings.
- What appears in the top of a ranking depends to a large extent upon a system’s geographical coverage, rating methodologies applied, indicators selected and indicator normalizations carried out.
- Current ranking systems are still one-dimensional in the sense that they provide finalized, seemingly unrelated indicator values rather than offer a dataset and tools to observe patterns in multi-faceted data.
- To enhance the level of understanding and adequacy of interpretation of a system’s outcomes, more insight is to be provided to users into the methodological differences between the various systems, especially on how their institutional coverage, rating methods, the selection of indicators and their normalizations influence the ranking positions of given institutions.

2.6.2 *Chapter 19*

A statistical analysis of full text downloads of articles in Elsevier’s ScienceDirect covering all scholarly disciplines reveals large differences between disciplines, journals, and document types as regards their download frequencies, their skewness, and their correlation with Scopus-based citation counts. Download counts tend to be two orders of magnitude higher and are less skewedly distributed than citations. Differences between journals are discipline-specific.

Despite the fact that in all analysed journals download and citation counts per article positively correlate, the following factors differentiate between downloads and citations.

- *Usage leak.* Not all full text downloads of a publisher archive’s documents may be recorded in the archive’s log files.
- *Citation leak.* Not all relevant sources of citations may be covered by the database in which citations are counted.
- *Downloading* the full text of a document does not necessarily mean that it is fully *read*.
- *Reading and citing populations may be different.* For instance, industrial researchers may read scientific papers but not cite them as they do not publish papers themselves.
- *Number of downloads depends upon type of document.* For instance, editorials and news items may be heavily downloaded but poorly cited compared to full length articles.

- *Downloads and citations show different obsolescence functions.* Download and citation counts both vary over time, but in a different manner, showing different maturing and decline rates.
- *Downloads and citations measure different aspects.* Short term downloads tend to measure readers' awareness or attention, whereas citations result from authors' reflection upon relevance.
- *Downloads and citations may influence one another in multiple ways.* More downloads may lead to more citations, but the reverse may be true as well. Articles may gain attention and be downloaded because they are cited.
- *Download counts are more sensitive to manipulation.* While citations tend to be regulated by the peer review process, download counts are more sensitive to manipulation.
- *Citations are public, usage is private.* While citations in research articles in the open, peer reviewed literature are public acts, downloading documents from publication archives is essentially a private act.

Part II

**Informetric Indicators of Research
Performance**

Chapter 3

Multi-dimensional Research Performance

Abstract This chapter distinguishes the main components of the research process. It continues with a presentation of a typology of the various dimensions of research input, process, output and impact. A key distinction is between scientific-scholarly and societal impact. The chapter ends with an overview of 28 important informetric indicators or indicator families, providing information on the dimensions they cover, and their main pros and cons.

Keywords Advisory committees · Advisory work · Artworks · Book chapter · Citation per article · Citations · CiteScore · Co-authorships · Commercialization · Cross-disciplinarity · Design · Device · Doctoral completions · Doctorates · Early career researchers · Editing · Efficiency indicators · Eigenfactor · Encyclopedia · End-user esteem · Exhibit · Fractional scientific strength · Full text downloads · Handbook · Highly cited publications · H-index · Image · Integrated impact indicator · Intellectual property · Intellectual rights · Invited lectures · Invited talks · Journal impact factor · Journal metrics · Journal paper · Licenses · Medical guidelines · Mendeley readers · Minimum performance standards · Monograph · Multi-dimensional performance · Negative-binomial distribution · Newspaper article · Online contribution · Online course · Patent · Patent citations · Percentile-based indicator · Policy document · Press story · Professional guidelines · Professional practice · Publication counts · R&D investment · Readership indicators · Relative citation rate · Reputation survey · Research active · Research dataset · Research efficiency · Research impact · Research infrastructure · Research input · Research mobility · Research output · Research process · Research report · Reviewing · Scale · Scholarly awards · Scholarly prizes · SJR · SNIP · Social media mentions · Software downloads · Spin-off · Start-up company · Super computing power · Sustainability · Syllabus · Textbook · Textbook sales · Top publications · Video · Web followers

3.1 Introduction

Four major components of the research process are distinguished: input, output, process and impact. They are listed in Table 3.1. *Input* relates to the human, physical and financial commitments devoted to research. Important sub-components are the amount of funding and human resources available, and the research infrastructure. *Process* indicators measure how the research that is conducted, including its management and evaluation. They include indicators of the type and degree of collaboration, and the efficiency of the process, relating output to input. A report of the Expert Group on the Assessment of University-Based Research gives as a typical example the total of human resources to support and fulfil technology transfer activities (AUBR, 2010, p. 36).

Output focuses on the quantity of research products, and *impact* to their benefits. The term ‘impact’ is used in a broad sense, and comprises both short term, measurable effects—in some classifications denotes as outcomes—and broader, longer term benefits (see for instance Harding, n.d.). A distinction is made between *scientific-scholarly* and *societal* outputs or impacts. The term ‘societal’ embraces a wide spectrum of aspects outside the domain of science and scholarship itself, including technological, social, economic, educational and cultural aspects. Some authors prefer to use the term ‘social’ rather than ‘societal’. The latter term is used in this book to distinguish between social and cultural aspects.

Research performance is reflected in *all four* components. Research funding, especially competitive funding, can be assumed to depend amongst other factors on the past performance and the reputation of the grant applicant. Efficiency, in Table 3.1 categorized as a process indicator, is also a performance indicator, as it aims to measure output or impact per unit of input. Moreover, research outputs may have both scientific-scholarly and societal benefits. Also, some indicators may be used both as output and as impact measures. For instance, the number of patents is an output indicator, but is, given the potential technological or economic value of patents, also used as an indicator of societal impact.

Table 3.1 Components of the research process

Component	Sub-component	Typical examples of indicators
Input	Funding and manpower	Amount of funds available; FTE Academic staff available
	Research infrastructure	Total R&D investment; Value of infrastructure and facilities
Output	Scientific-scholarly	Nr. journal articles or book chapters; research data files
	Societal	Nr. new designs; art performances
Process	Efficiency	Nr. publications per FTE academic staff
	Collaboration	Nr. collaborations or partnerships
Impact	Scientific-scholarly	Citation impact; prizes and awards
	Societal	Revenues from commercialization of intellectual property

Research output and process indicators are further discussed in Sect. 3.2, impact indicators in Sect. 3.3, and input indicators, especially those concerning the research infrastructure in Sect. 3.4. The European project ACUMEN—acronym of Academic Careers Understood through Measurements and Norms—in the 7th European Framework Programme has developed a portfolio aimed to provide an integral picture of a researcher's achievements and capabilities. It allows researchers to present themselves through a brief narrative in which they can highlight their past performance and future goals (Bar Ilan, 2014). Sections 3.2 and 3.3 combine categorizations from the ACUMEN portfolio with those made in Moed and Halevi (2015) and in the AUBR report. Section 3.4 on research infrastructure is fully based on the AUBR Report (AUBR, 2010).

3.2 Research Outputs

Table 3.2 presents many of the most important output forms, and gives typical examples of these, but it is not fully comprehensive. Documents related to Research Excellence Framework (REF) in the UK give a more complete overview of the output forms taken into account in the assessment of research in the various major disciplines (REF, 2012).

Table 3.2 Research outputs

Primary orientation	Publication	Non-publication
Scientific-scholarly	<ul style="list-style-type: none"> • Scientific journal paper • book chapter • scholarly monograph • conference paper; • editorial • review 	<ul style="list-style-type: none"> • Research dataset • software; tool; instrument • video of experiment • Registered intellectual rights (discoveries)
Educational/Teaching	<ul style="list-style-type: none"> • Teaching course book; • Syllabus • Text- or hand book; 	<ul style="list-style-type: none"> • Online course • Students completed • Degrees attained (e.g., doctorates);
Technological	<ul style="list-style-type: none"> • Patent; • commissioned research report; 	<ul style="list-style-type: none"> • Product; process • device; design; image
Economic		<ul style="list-style-type: none"> • Spin-off; start-up company • Registered industrial rights • Revenues from commercialization of intellectual property

(continued)

Table 3.2 (continued)

Primary orientation	Publication	Non-publication
Social	<ul style="list-style-type: none"> • Professional guidelines • Policy document 	<ul style="list-style-type: none"> • Scientific advisory work
Cultural (Communication to the general public)	<ul style="list-style-type: none"> • Newspaper article • Press story • Encyclopaedia article • Popular book or article 	<ul style="list-style-type: none"> • Interview • Event; performance; exhibit • Online presence; online contribution

3.3 Research Impacts

Table 3.3 distinguishes the various types of research impact, and gives typical examples of indicators that may be used to assess these. It combines categories and indicators from the ACUMEN portfolio, the AUBR Report, and from Moed and Halevi (2015). The two main categories are scientific-scholarly and societal impact. The term ‘societal’ embraces a wide spectrum of aspects outside the domain of

Table 3.3 Types of research impact and their indicators

Type of impact	Short Description; Typical examples	Indicators (examples)
<i>Scientific-scholarly</i>		
Impact on knowledge growth	Contribution to scientific-scholarly progress: creation of new knowledge	<ul style="list-style-type: none"> • Indicators based on publications and citations in peer reviewed journals and books, e.g., total and average citations; Article citations (top 3); Age corrected H-index; Book citations (top 3); • Indicators of multi-, inter- or cross-disciplinary research • Scholarly prizes and awards • Editing and reviewing • Membership of scientific committees • Invited talks • Mendeley readers • Participation in online discussions • Social scholarly web followers
Impact on communication	Effectiveness of publication strategies; visibility of used publication outlets	<ul style="list-style-type: none"> • Journal impact factors and other journal metrics; • Diversity of used outlets.
<i>Societal</i>		
Educational/teaching		<ul style="list-style-type: none"> • Awards • Online views • Syllabus mentions • Textbook sales • Invited lectures • Research datasets or software downloads

(continued)

Table 3.3 (continued)

Type of impact	Short Description; Typical examples	Indicators (examples)
Social	Stimulating new approaches to social issues; informing public debate and improve policy-making; informing practitioners and improving professional practices; providing external users with useful knowledge; Improving people's health and quality of life; Improvements in environment and lifestyle;	<ul style="list-style-type: none"> • Citations in medical guidelines or policy documents to research articles (e.g., Haunschild & Bornmann, 2017) • Funding received from end-users • End-user esteem (e.g., appointments in (inter)national organizations, advisory committees) • Juried selection of artworks for exhibitions • Mentions of research work in social media • Advice; • Professional practice; • Laws, regulations, guidelines
Technological	Creation of new technologies (products and services) or enhancement of existing ones based on scientific research	<ul style="list-style-type: none"> • Number of patents • Number of citations from patents or articles to patents
Economic	Improved productivity; adding to economic growth and wealth creation; enhancing the skills base; increased innovation capability and global competitiveness; uptake of recycling techniques;	<ul style="list-style-type: none"> • Revenues created from the commercialization of research generated intellectual property (IP) • Number patents, licenses, spin-offs • Number of PhD and equivalent research doctorates • Employability of PhD graduates • Income • Consultancies • Citations from patents to articles • Citations to patents • Spin-offs, start ups
Cultural	Supporting greater understanding of where we have come from, and who and what we are; bringing new ideas and new modes of experience to the nation.	<ul style="list-style-type: none"> • Media (e.g. TV) performances • Essays on scientific achievements in newspapers and weeklies • Mentions of research work in social media (tweets or blog posts)

science and scholarship itself, including technological, educational, social, economic and cultural aspects. The latter type is denoted in the ACUMEN list as ‘communications towards the general public’.

3.4 Research Infrastructure

This dimension relates to the basic facilities that support research, the scale of the research activities, and their sustainability. These aspects are discussed in a report published by an Expert Group on University Based Research, installed by the

Table 3.4 Indicators of research infrastructure, sustainability and scale

Aspect	Indicator
Research infrastructure	Value of infrastructure and facilities (expressed in money)
	Research active academics
	Percentage research-active academics per total academic staff
	Total R&D investment
Sustainability and scale	Postgraduate research student load
	Involvement of early career researchers in teams
	Number of collaborations and partnerships
	Doctoral completions

Source AUBR (2010, p. 47–48)

European Commission (AUBR, 2010). Indicators of research infrastructure, sustainability and scale are *not* primarily performance indicators but focus on the *precondition* for performance. This is why these indicators are categorized as input measures in Table 3.1.

The relevance of indicators of research infrastructure is grounded in the notion that any research assessment directed towards the future does not only take into account the scientific-scholarly or societal value of the research is at stake, but also the question as to whether the research environment available to the research teams involved is adequate and continues to be so in the foreseeable future. In the assessment of research teams or their grant proposals the term ‘viability’ is sometimes used to indicate this dimension.

Infrastructure includes library and digital access, availability of super-computing power and other research equipment indispensable for the tasks to be performed, and a stable basic financial support. As the AUBR Report puts it, indicators of research infrastructure “measure the research environment as a predictor of research capability and success” (AUBR, 2010, p. 79). Most indicators are not bibliometric. Research intensity, total investment in research and the degree of integration with international networks are important aspects.

3.5 Summary Table of 28 Important Informetric Indicators

Table 3.5 presents a list of 28 often used indicators, giving a short definition and a summary of their main potential and limits. Indicators are grouped on the basis of how they are technically created, especially from which data sources they are derived. Between parentheses is indicated what they measure, using the categorization presented in Tables 3.1, 3.2, 3.3 and 3.4. Table 3.5 is further discussed in Chap. 4.

Table 3.5 Potentialities and limits of 28 often used indicators of research performance

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
<i>Publication and citation based measures (scientific-scholarly outputs and impacts)</i>			
Publication counts	Number of publications made by an entity during a specific publication time window	<ul style="list-style-type: none"> Publications constitute in all scientific-scholarly subject fields an important output form. Useful tool to identify at the bottom of the performance distribution not sufficiently research active entities if its value is below a certain (subject field dependent) minimum Although mostly applied to journal articles, the measure may also include books and other textual output forms 	<ul style="list-style-type: none"> If numbers exceed a certain minimum level, differences between them cannot be interpreted in terms of performance Despite large differences in the level of publication output between subject fields, publication counts are not field-normalized Have a limited value in the private sector, and in technical science Collaboration and multi-authorship is a rule rather than an exception. How to assess the contribution of an entity to the output of team work?
Citations (general)	Citations received during a specific (fixed or variable) citation time window by an entity's publication	<ul style="list-style-type: none"> Citation counts can be interpreted in terms of intellectual influence Can be collected according to an objective methodology that gives the analyst a certain degree of independence Help to bring relevant background knowledge on performance into the open May focus on short or longer citation impact Are mostly applied to journal articles, but more and more other types are processed 	<ul style="list-style-type: none"> Citations are partial indicators of actual influence, but may be affected by other factors For a proper interpretation background knowledge on assessed entities is indispensable Large differences in citation practices across subject fields Biased towards journal publications as sources of citations Only partially valid in social sciences, humanities, and applied or technical sciences

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
		and analyzed as well, e.g., books	
Citations per article	Ratio of total citations and number of published articles	<ul style="list-style-type: none"> Is size independent: reveals citation impact relative to size of publication volume Can be used to indicate ‘impact saturation’ 	<ul style="list-style-type: none"> Should be used with caution when comparing entities with very different publication volumes or active in highly specialized subjects Is not field-normalized Is the mean of a skewed citation distribution
Relative citation rate	Citations per article, normalized by the world citation average in an entity’s subject field(s)	<ul style="list-style-type: none"> Corrects for differences in citation practices among subject field May also take into account document type (e.g., review, full length article), and age of cited article Takes into account the full citation distribution 	<ul style="list-style-type: none"> Should be used with caution when comparing entities with very different publication volumes or active in highly specialized subjects Field delimitation must be sound
Number of top publications	Number of an entity’s articles in the top 1, 5 or 10 per cent worldwide in the subject fields in which the entity is active	<ul style="list-style-type: none"> Corrects for differences in citation practices among subject fields Focuses on the top of the citation distribution (‘highly cited publications’) 	<ul style="list-style-type: none"> Maps all actual values onto a 0–100 scale; One may lose the sense of underlying absolute differences, and undervalue extraordinary papers Is size-dependent (but <i>percentage</i> top publications relative to total output is size-independent)
H-Index	A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np–h) papers have no more than h citations each	<ul style="list-style-type: none"> Combines an assessment of both quantity (published papers) and impact (citations) Tends to be insensitive to highly cited outliers and to poorly cited papers 	<ul style="list-style-type: none"> Its value is biased in favor of senior researchers compared to juniors; Impact of the most cited papers hardly affects its value

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
			<ul style="list-style-type: none"> • Does not correct for differences between subject fields
Integrated Impact Indicator	Sum of an entity's field-normalized citation scores expressed as percentile ranks	<ul style="list-style-type: none"> • Combines an assessment of both quantity (published papers) and impact (citations) • Corrects for differences in citation practices among subject field • Takes into account all publications, and thus rewards entities with a large article production 	<ul style="list-style-type: none"> • Its value is biased in favor of senior researchers compared to juniors; • Maps all actual values onto a 0–100 scale; • One may lose the sense of underlying absolute differences, and undervalue extraordinary papers
<i>Journal-based indicators (scientific-scholarly impact, communication)</i>			
Journal metrics (general)		<ul style="list-style-type: none"> • The quality or impact of the journals in which an entity has published is a performance aspect in its own right 	<ul style="list-style-type: none"> • The relationship between a journal's impact factor and the rigorousness of its manuscript peer review process is unclear • Journal metrics cannot be used as a surrogate of actual citation impact; • Are no good predictors of the citation rate of individual papers • Their values can to some extent be manipulated and may be affected by editorial policies
Journal impact factor	Average number of citations in a particular citing year to a journal's 1–2 year old document, divided by the number of the journal's 1–2 year old 'citable' documents	<ul style="list-style-type: none"> • Is a relatively simple measure • Is already used for decades by researchers, librarians and publishers • Is widely available for large numbers of journals 	<ul style="list-style-type: none"> • Does not correct for differences in citation practices between subject fields • Discrepancies between document types included in numerator and denominator

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
			<ul style="list-style-type: none"> • Is sensitive to citation outliers • Biased in favor of review journals
SNIP	Field normalized measure relating a journal's citation rate to the average length of the reference lists in papers in its subject field	<ul style="list-style-type: none"> • Corrects for differences in citation frequencies between subject fields • Is independent of an a-priori classification of journals into subject categories • A journal's subject field is defined as the set of articles citing that journal 	<ul style="list-style-type: none"> • Different versions of SNIP exist • Is difficult to explain to laypersons • Is sensitive to citation outliers • Biased in favor of review journals
SJR, Eigenfactor	Measure derive from the journal citation network, weighting each received citation with the importance of the citing journal	<ul style="list-style-type: none"> • Corrects for differences in citation frequencies between subject fields • Takes into account the full journal-to-journal citation network 	<ul style="list-style-type: none"> • Scale and range of obtained values is different from that of the journal impact factor • Evidence of a Matthew effect: inequalities among journals increase; strong journals become stronger
Glanzel's negative-binomial model	Characterizes a journal in terms of the two parameters of a fitted negative-binomial distribution (Glanzel, 2008; 2009)	<ul style="list-style-type: none"> • Abandons the idea to characterize a complex citation distribution with one single parameter • Is based on purely statistical considerations and in this sense neutral in terms of evaluative assumptions 	<ul style="list-style-type: none"> • The two parameters are difficult to interpret in terms of communication or performance aspects • Users tend to prefer one single measure
<i>Patent-based indicators (societal-technological impact)</i>			
Number of patents	Number of patent applications or granted patents by applicant or by inventor	<ul style="list-style-type: none"> • Inventions may be disclosed in patents; • Patent data is available at a global level • Analyses of inventors reveal the extent to which scientists with 	<ul style="list-style-type: none"> • Not all inventions are patentable or actually patented • The propensity to patents differs across countries because of legislation or culture,

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
		academic positions contributed to technological developments	and also across subject fields • Patent application date and granted date may differ several years
<i>Altmetrics, usage-based and web-based measures (scientific-scholarly or societal output and impact)</i>			
Full text downloads	Number of downloads of an electronic publication in full text format	<ul style="list-style-type: none"> Patent-to-patent citations may reveal a patent's technological value Patent-to-paper citations may reveal a paper's technological impact or a technology's science base 	<ul style="list-style-type: none"> Data on cited references in patents to scientific papers is not well standardized Patents constitute a poor indicator of the commercialization of research
Social media mentions	Mentions of publications in Twitter, Facebook, blogs, and other social media	<ul style="list-style-type: none"> Are in principle available immediately after publication Enable researchers to assess the effectiveness of their communication strategies May reveal attention of scholarly audiences from other research domains or of non-scholarly audiences 	<ul style="list-style-type: none"> Downloaded articles may be selected according to their face value rather than their value perceived after reflection Incomplete data availability across providers Affected by differences in reading behavior between disciplines and institutions Counts can be manipulated Difficult to ascertain whether downloaded publications were actually read or used

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
		<ul style="list-style-type: none"> • Measure attention rather than influence • May provide tools to link scientific expertise to societal needs 	<ul style="list-style-type: none"> • Interdependence of the various social media may boost up numbers • Altmetrics are easily be manipulated, due to lack of quality control and of identification of users
Readership indicators	Readers of publications in reference managers or scholarly network sites	<ul style="list-style-type: none"> • Citations in articles in preparation are in principle available before publication • Are potentially predictors of emerging trends 	<ul style="list-style-type: none"> • Results depend on readers' cognitive and professional background
Webometric indicators	Web-based indicators of Web presence and Web impact in terms of number of inlinks	<ul style="list-style-type: none"> • Webometric measures give an indication of the presence or visibility at the Web • They do not merely relate to their <i>research</i> mission, but also on their <i>teaching</i> and <i>services</i> activities, • It extracts linkage data from a universe of Web documents that is much larger than that covered in the multi-disciplinary citation-indexes 	<ul style="list-style-type: none"> • There is no systematic information on the universe Web sources covered and their quality • The degree of an institution's Web presence depends upon her internal policies towards the use of the Web, and upon the propensity of her staff to communicate via the Web
<i>Reputation and esteem based measure (scientific-scholarly impact)</i>			
Reputation Survey	Number of mentions in surveys sent to peers, users of or customers	<ul style="list-style-type: none"> • Survey distributed among large sample of experts • Votes take into account and integrate a range of aspects • Surveys can be conducted based in scientifically-founded methods developed in social sciences 	<ul style="list-style-type: none"> • Possible biases due to selective responses and 'amplifications' via social media • Votes are based on impressions, not on an 'objective' methodology • Impressions may be based upon performances in a distant past

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
			<ul style="list-style-type: none"> • Votes may be based on ‘hear-say’ rather than on founded judgement
Scholarly prizes and awards	Number of prizes and/or awards granted by professional organizations on the bases of peer review	<ul style="list-style-type: none"> • Receiving a prestigious price/award is a clear manifestation of esteem 	<ul style="list-style-type: none"> • Status of prizes and awards may be difficult to assess • Absolute numbers tend to be low; overwhelming majority of scholars has a zero score • Evaluation processes are not always fully transparent
<i>Economic/econometric indicators (economic impact; process indicators)</i>			
Efficiency indicators	Publication output or citation impact per FTE Research or per dollar spent	<ul style="list-style-type: none"> • Relate output to input; allow calculation of output-per-dollar ratios • Efficiency is an important aspect of a research system and deserves a systematic analysis • Are size-independent measures but do not use publication counts as a proxy of size 	<ul style="list-style-type: none"> • Relationship between input and output is not necessarily linear, and may involve a time delay • Any input indicators also reflect output • Accurate, standardized input data is often unavailable • Comparisons across countries are difficult to make, due to differences in the classification of administrative and financial data • Focuses on a system’s efficiency but disregards its effectiveness in reaching external goals
Measures of economic value	Number of licenses, spin-offs	<ul style="list-style-type: none"> • May indicate economic value of research 	<ul style="list-style-type: none"> • Not all contributions of research to economic development can be easily measured

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
Funding related measures	<ul style="list-style-type: none"> External research income; Number and percentage competitive grants won 	<ul style="list-style-type: none"> Willingness of industry to pay for research is a useful indicator of its anticipated contribution to innovation and the economy Comparable data, verifiable through audit, is useful for comparing research performance across the system and within universities 	<ul style="list-style-type: none"> Levels of external funding vary greatly across disciplines Data collection may be difficult in case of funding by end users because this information is not known to the University administration
<i>Measures of collaboration, migration and cross-disciplinarity (scientific-scholarly impact, process indicators)</i>			
Co-authorships	Various measures based on co-authorship, e.g., international co-authorship	<ul style="list-style-type: none"> Measure formal scientific-scholarly collaboration May indicate degree of integration in international networks 	<ul style="list-style-type: none"> Purely informal forms of collaboration are invisible Standard measures do not reflect the nature of the role in these networks
Research mobility	Measures of research mobility	<ul style="list-style-type: none"> Institutional, international and disciplinary mobility and migration can be studied in a longitudinal publication analysis 	<ul style="list-style-type: none"> Affiliations' and authors' names as mentioned in the publish papers are not always standardized Events that have not resulted in a publication remain invisible
Cross-disciplinarity	Various measures of multi-, inter- or cross-disciplinary research	<ul style="list-style-type: none"> Measure relevance of a piece of research for surrounding disciplines Measures impact breadth 	<ul style="list-style-type: none"> Assumes a valid, operational classification of research into disciplines
<i>Measures of research infrastructure (input indicators)</i>			
Sustainability and scale	<ul style="list-style-type: none"> the ratio of research students per academic staff number of early career researchers involved in research; 	<ul style="list-style-type: none"> the ratio of research students per academic staff, number of early stage researchers involved in research; and the number of PhD completions, and 	<ul style="list-style-type: none"> Research practices differ across disciplines Large research teams or laboratories are found mostly in the

(continued)

Table 3.5 (continued)

Indicator	Short definition	Potentialities; strong points	Limits; points to be taken into account
	<ul style="list-style-type: none"> • number of PhD completions, and total investment in R&D are useful indicators 	total investment in R&D are useful indicators	natural and life sciences
Research infrastructure	<ul style="list-style-type: none"> • Total research funding • Total value of research facilities • Amount of literature sources accessible • Super computing power available 	<ul style="list-style-type: none"> • An adequate infrastructure is an important precondition of research performance 	<ul style="list-style-type: none"> • It is difficult to obtain reliable, comparable institutional data, as there is no agreement on the basis on which full cost of research investment should be calculated
Being research active	<ul style="list-style-type: none"> • Combination of indicators, e.g., publication counts, funding obtained, PhD students supervised 	<ul style="list-style-type: none"> • Sets minimum performance standards for a specific period 	<ul style="list-style-type: none"> • There is no clear, generally accepted definition of being research active across universities, countries and disciplines

Chapter 4

Informetric Tools

Abstract This chapter presents a detailed discussion of the 28 indicators or indicator families presented in Table 3.5 in Chap. 3. It explains how they are calculated, and what their main pros and cons are. The chapter continues with an overview of ‘big’ informetric datasets, and how they can be combined. Finally, it briefly introduces mapping tools of science and technology.

Keywords Article recommender systems • Assessment context • Assessment objectives • Authoritativeness of citations • Big data • Citation context • Clickstreams • Cloud computing • Co-citation analysis • Cognitive linkages • Composite indicator • Cost-benefit measures • Cross-disciplinary research • Download immediacy index • Esteem • Facebook • Functional imperative • Intellectual influence • Journal influence weights • Journal usage factor • Mapping software packages • Mass media • Multi-authorship • Network analysis • Non-parametric • Page-rank algorithm • Percentile ranks • Persuasiveness of citations • Preconditions to performace • Publication types • Reading behavior • Recognition by peers • Reputation • Research collaboration • Research intensity • Research mission • Research productivity • Research specialties • Response rates • Scholarly blogs • Science base • Science maps • Science of science • Shanghai ranking • Size-dependent • Size-independent • Socially defined quality • Stock portfolios • Technological impact • Time delays • Twitter • Web impact factor • Webometrics

4.1 Introduction

In the past decade, several important monographs and handbooks were published on informetric or bibliometric indicators. In 2010, E. Geisler published his monograph *The Metrics of Science and Technology* (Geisler, 2010), and Peter Vinkler his book *The evaluation of research by scientometric indicators* (Vinkler, 2010). In 2016, Roberto Todeschini and Alberto Baccini published their Handbook

This chapter re-uses with permission selected paragraphs from Moed and Halevi (2015).

on Bibliometric Indicators (Todeschini & Baccini, 2016) that presents definitions, formulas, algorithms and comments on the bibliometric indicators known in the literature. An extensive review of available citation based indicators is given by Waltman (2016), while Wildgaard (2015) compiled a list of 87 citation-based indicators applied at the level of individual authors. A full discussion of the accuracy of citation counts and related aspects is presented in Chaps. 12–14 in Moed (2005).

Section 4.2 gives main characteristics of the 28 indicators presented in Table 3.5 in Chap. 3. As indicated in Chap. 1, this book does *not* discuss technical details of indicators, their formal definition and their statistical properties. The current chapter aims to present the main types of indicators, discuss their base ideas, explain which theoretical assumptions are underlying their definition, and highlight their potential and limits.

Typical examples of big datasets in informetric research are given in Sect. 4.3. The section shows how the *combination* of large datasets creates *compound databases* that can be used to investigate relevant informetric research questions. Finally, Sect. 4.4 deals with science mapping. It provides an introduction to Sects. 13.2 and 13.3 that presents a series of mapping methodologies, software packages and maps.

4.2 Indicators

4.2.1 Publication-based Indicators

According to Robert K. Merton, free access to scientific results is a *functional imperative* in science. Publishing research outcomes in the open, publicly available literature serves the advancement of scientific-scholarly knowledge. Publications have another function as well: the principal way for a scholar to be *rewarded* for his contribution to the advancement of knowledge is through recognition by peers. In order to receive such an award, scholars publish their findings openly, so that these can be used and acknowledged by their colleagues. These notions provide a theoretical basis and justification for the use of publication counts in the assessment of research performance. However, several comments should be made.

Although publications are important research outputs in all scientific-scholarly subject fields, *different publication formats* exist in the domain of science and scholarship, each with their own characteristics and special functions. For instance, in the natural and life sciences journal articles are the primary source of written communication, but in many parts of social sciences and humanities monographs and chapters in edited books are the most important types, while in many domains of the applied and technical sciences, articles in conference proceedings constitute the most important type. The implication is that there is *no* uniform counting scheme which is valid across all domains of science and scholarship. Despite large

differences in the level of publication output between subject fields, there are no field-normalized publication indicators.

In academic institutions, publications constitute in all scientific-scholarly subject fields an important form of academic output. But in the private sector, and also in academic departments with a strong applied orientation, primarily aiming to produce new products or processes, publishing in the public, peer reviewed literature often does not have the highest priority (see for instance Moya et al., 2014).

During the past decade more and more emphasis is laid on the importance of *other forms* of scholarly output than the traditional written forms mentioned above. Two main developments that are responsible for this shift are the *computerization* of the research process and scientific communication, and the increasing interest in the *societal influence* of research. Table 3.1 in Chap. 3 gives a list of some of these other forms. Typical examples are research data sets, and communications on research findings via social media, directed towards the general public.

Collaboration and multi-authorship in publications is a rule rather than an exception. Both at the level of individual researchers, research groups, institutions and national research systems, the number of co-publications with other entities increased dramatically. If research is more and more becoming team work, the issue as to how one can assess the contribution of an individual participant to the team's output becomes more and more important. This issue is further discussed in Sect. 9.1 in Part IV.

A final comment relates to the *validity* of publication counts as measures of research performance. In 1982, a Survey Committee installed by the Dutch government and assessing the performance of academic research groups in the Netherlands in the field of biochemistry explicitly stated that, in her view, publication counts were useful only to identify research units of which the output was below what was considered as an adequate *minimum level* of scientific production, but were *unapt* to discriminate in terms of performance between groups with counts *above* this threshold (Survey Committee Biochemistry, 1982). This is why the Committee decided to use citation-based measures in their assessment. The use of minimum publication thresholds in performance assessment is further discussed in Sect. 10.4 in Part IV.

4.2.2 *Citation-based Indicators*

Citation analysis is one of the most established methods in research assessment (van Raan, 1996; 2004a). As indicated above, according to Merton's theory the principal way for a scholar to be rewarded for his contribution to the advancement of knowledge is through recognition by peers. In order to receive such an award, scholars publish their findings openly, so that these can be used and acknowledged by their colleagues. At the same time, they have the obligation to acknowledge the sources containing the knowledge claims they have built upon in their own works. In short, they have to 'give credit where credit is due'. Scholars have no choice:

one's private property is established by giving it away, and in order to receive peer recognition, they must provide it to others.

These notions can provide a theoretical justification of the use of citations as *proxies* of more direct measurements of *intellectual influence*, such as peer judgments and honorific awards (Zuckerman, 1987), or as reflections of 'socially defined quality' (Cole & Cole, 1967). But alternative interpretations of citation data exist as well (Cronin, 1984), including the view of citation counts as reflections of authoritativeness (Gilbert, 1977) or persuasiveness and awareness (Cozzens, 1989). Moed (2005a, Chap. 15) gives a summary of a series of theoretical positions on what citations measure.

Even if citations are interpreted as a measure of scholarly influence, they may be affected by other factors as well. This is expressed clearly in the seminal paper by Martin and Irvine (1983) quoted in Sect. 1.2 in Part I. Moreover, one should distinguish between the *validity* of a claim and the acknowledgement of its *source*. In the Mertonian interpretation citations reflect the latter, but *not* the former aspect. Citations are *by no means* indicators of the *validity of a knowledge claim*. Attempts to demonstrate the validity of a claim by counting citations to the work in which the claim is made, or, for instance, to related articles by the authors publishing this work, have *no* theoretical basis in the Mertonian framework.

A strong point of citation analysis is that it can be conducted according to an objective methodology, which provides the analyst with a certain degree of independence from the preferences and views of subjects of the analysis and of those commissioning it. On the other hand, in order to interpret the outcomes of a citation analysis, background knowledge on the entities under assessment and their work is indispensable. In fact, citation analysis can help to bring such background knowledge into the open. It may focus on short or longer term impact, depending upon the time horizon taken into account. Although citation analysis traditionally focuses on journal publications both as sources and targets, more and more other publication types are taken into account, including monographs and edited books. These limitations are further discussed in Chap. 9 in Part IV.

There are many different types of citation-based indicators. Table 3.5 in Chap. 3 includes an absolute, 'size dependent' citation count, but also a 'size-independent' citation-per-article ratio; a field-normalized, relative citation rate, but also the number of highly cited publications in a subject field, an Integrated Impact Indicator based on percentile ranks, an 'impact productivity' measure relating citation impact to 'input' measures such as the amount of dollars spent, and, last but not least, the h-index.

4.2.3 Journal Metrics

Citation indicators of scientific-scholarly journals have always played an important role in bibliometrics. The journal impact factor, derived from the Science Citation Index (currently the Web of Science), and published by the Institute for Scientific

Information (Thomson Reuters, currently Clarivate Analytics) is probably the most frequently used informetric construct. Introduced by Eugene Garfield primarily to evaluate the journal coverage of the Science Citation Index independently of scientific publishers, it has functioned for many years as a ‘paradigm’ in indicator development. Great practical advantages of this measure are that it is a relatively simple measure, and widely available for large numbers of journals.

Several *alternative* journal measures of citation impact have been proposed during the past decades. One group of measures aims to correct for differences in citation practices among subject fields and calculates in this way *field-normalized* indicators, comparing a journal’s citation score with those of other journals in the same subject field. Many different ways have been proposed to express this notion in mathematical-statistical terms. A typical example is the SNIP or source-normalized impact per paper (Moed, 2010, 2016b; Waltman et al., 2013). Many indicators in this group have the property that a journal with a citation impact equal to the subject field ‘average’ has a value of 1.0.

A second group of indicators, including Scimago Journal Rank indicator and Eigenfactor, rejects the assumption that ‘all citations are equal’, and assigns to each received citation a weight proportional to the citation rate of the citing journal. In other words, a citation from a high impact journal counts for more than a citation from a less impactful outlet. Calculation of these indicators is based on an algorithm similar to the page-rank algorithm applied by Google to rank websites, which in its turn applies a methodology for calculating journal influence weights developed by Pinski and Narin (1976). Chapter 14 in Part V gives more information on the various journal indicators.

The citation distribution of articles in journals tends to be skewed: the major part of articles is not cited at all, while a few papers may be highly cited. All indicators mentioned above are based on the assumption that this type of distribution can be properly characterized in one single parameter. But several bibliometric approaches question the validity of this assumption. A good example is Wolfgang Glanzel’s analysis of journal citation distributions as negative-binomial distributions, estimating for a journal the values of its two parameters (Glanzel, 2008, 2010). Glanzel’s proposal is purely based on statistical considerations, and can be considered neutral in terms of evaluative assumptions. A further discussion of the role of such assumptions in the definition of indicators is presented in Sect. 7.2 in Part III.

The quality or impact of the journals in which an entity has published is a performance aspect in its own right. But the relationship between a journal’s impact factor and the quality or rigorousness of its manuscript peer review process is unclear. Sugimoto et al. (2013) did not find solid evidence that these two measures are significantly positively correlated. This issue is further discussed in Sect. 12.1 in Part IV.

Journal citation measures are often used as performance indicators of individual researchers. A justification of this type of use is that the quality or impact of the journals in which an individual or research group has published is a performance aspect in its own right. But this type of use has severe limitations as well. Chapter 8 in Part III defends the position that the adequacy of using journal citation measures

in the assessment of individuals strongly depends upon the assessment objectives and the assessment context, and that in many cases such use is *inappropriate*. In Sect. 5.2 it is argued that journal metrics cannot be used as a surrogate of actual citation impact, and that journal impact factors are no good predictors of the citation rate of individual papers. Section 9.4 shows that they can to some extent be manipulated and may be affected by editorial policies. A more detailed overview of the various types of journal metrics is presented in Chap. 14 in Part V.

4.2.4 Patent-based Indicators

Modern technology is more and more science-based. Academic researchers increasingly appear as inventors of patents, and analyses of inventors of the findings described in patents reveal the extent to which scientists with academic positions contributed to technological developments (Noyons et al., 2003; Schmoch, 2004). Combined analysis of scientific publications and patents reveals knowledge networks of academic scientists and industrial researchers. In addition, patents increasingly cite the scientific literature. Studies of citations in patents to the scientific literature show the science base of modern technology (e.g., Carpenter & Narin, 1983).

Patent analysis is a unique method that not only measures the number of patents associated with an applicant institution or an inventor, but also looks at its citations from two perspectives: citations in patents to other patents; and citations in patents to the scientific literature. Similar to citation analysis of journal articles, patent citation analysis identifies high citations to basic and applied research papers in patents as well as patents that are highly cited by recently issued patents. In both domains, groundbreaking studies were conducted by Francis Narin and co-workers (Carpenter & Narin, 1983; Albert et al., 1991; Narin, 1994; Narin et al., 1997).

A useful indicator of the value of a patent is the number of times that it is cited by a later patent ('forward patent citations'). Studies conducted by Narin and others provided evidence that patents describing a technology for which there is relatively high demand tend to be cited more often than less valuable patents. Patent citations can be used even to manage stock portfolios (Narin, Breitzman, & Thomas, 2004).

Although patent databases are available at the global level, patent records, and especially the cited references contained in patents are not fully standardized. Therefore, a major effort must be made in order to achieve an acceptable level of accuracy when counting citations to the scientific literature. Moreover, not all inventions are patentable or actually patented. Also, the propensity to apply for patents differs across countries because of legislation or culture, and also across subject fields. Patent application date and granted date may differ several years.

The AUBR Expert Group underlines that patents are a very poor indicator of the *commercialization* of research results (AUBR, 2010, p. 45). Yet, patents are almost the only form of public communication that can be used as indicator of *technological* innovation and thus it is used as a part of the assessment of institutions and

individuals. Chapter 15 in Part V provides more information on the construction and use of patent-based indicators, on the time delays between a basic research finding and its application into new products, and on the interplay between science and technology.

4.2.5 *Usage-based Indicators*

In their 2010 review article, ‘Usage Bibliometrics’, Michael Kurtz and Johan Bollen explored the addition of modern usage data to the traditional data used in bibliometrics, data from scientific publications (Kurtz & Bollen, 2010). The usage data includes clickstreams, full text downloads and views of scholarly publications recorded on an article level. A well-known phenomenon in research is the difference between the frequency at which articles are read, browsed or scanned on the one hand, and the number of times they are cited, on the other. This difference varies among disciplines and institutions depending on their reading and citing behavior. In addition, there are certain types of documents that might be read more than cited such as reviews, editorials, tutorials or other technical output (Paiva et al., 2012; Schloegl & Gorraiz, 2010). As a typical example, Fig. 1.1 in Chap. 1 showed a longitudinal analysis of the number of full text downloads of one particular article on a daily basis, revealing an upward trend around the day it was presented at an international conference.

Several usage factors similar to the journal impact factor have been proposed. For instance, the Usage Factor Project currently managed by COUNTER, a non-profit organization supported by a global community of library, publisher and vendor members, launched the Journal Usage Factor (JUF), defined as a ratio of the total usage over period x of items published in a journal during period y, and the total number of items published in that journal during period y (COUNTER, n.d.). In line with this scheme, Wan et al. (2010) define a journal’s Download Immediacy Index (DII) by counting items published in a particular year, and usage of these items during the same year. This is the exact analogue of the journal immediacy index, calculated on the basis of citations by Thomson Reuters and published in its Journal Citation Reports.

Usage analysis may be applied not only to *primary* sources and analyze for instance their full text downloads, but also to searching behavior in *secondary* literature databases such as Web of Science or Scopus. Another relevant distinction in the analysis of usage data is between an approach focusing on generating usage *counts*, and one aimed to establish from user clickstreams *relationships* between documents, and cluster them on the basis of their usage similarity. A typical example of the latter approach is to determine for a given, downloaded seed document a set of other documents that were downloaded in the same user sessions as the seed document, and therefore can be assumed to bear a cognitive relationship to it. *Article recommender* systems are often based on this principle.

An important limitation of usage data is that downloaded articles may be selected according to their face value rather than their value perceived after reflection. Other limitations and challenges highlighted by Bollen and van de Sompel (2008) are: incomplete data availability across providers; differences in reading behavior between disciplines and institutions which are difficult to account for; content crawling and automated downloads software tools that allow individuals to automatically crawl and download large amount of content which does not necessarily mean that it was read or viewed; and the difficulty to ascertain whether downloaded publications were actually read or used.

Chapter 19 in Part VI gives a technical introduction to usage analysis, comparing it systematically with citation data, and further discusses its significance in research assessment.

4.2.6 Altmetrics

Altmetrics is a relatively new area in indicator development (Piem et al., 2010). It emerged from the increasing numbers of social media platforms and their prolific use by scientists and researchers (Taylor, 2013). The rationale behind the use of these alternative measures is that mentions of a publication in social media sites can be counted as citations and should be taken into consideration when reviewing the impact of research, individual or institution (Adie & Roe, 2013; Barjak et al., 2007). Four types of altmetric data sources can be distinguished, covering distinct types of activity (Altmetric.com, 2014).

- Social media such as Twitter and Facebook, covering *social activity*.
- Reference managers or reader libraries such as Mendeley or ResearchGate covering *scholarly activity*.
- Various forms of scholarly blogs reflecting *scholarly commentary*.
- Mass media coverage, for instance, daily newspapers or news broadcasting services, informing *the general public*.

In addition, two types of indicators can be distinguished, namely *absolute counts*—for instance, the number of tweets referring to a particular scholarly article—and *composite metrics*, weighting counts from different sources and calculating one single number. Today there are a few companies that offer composite altmetric scores, including Altmetric.com and ImpactStory.com. Implementation of altmetrics indicators in primary and secondary literature databases is increasing and examples can be seen in Scopus.com and PLOS.

Thelwall (2014b) concludes that altmetrics may have value as indicators of research impact, as they may provide insights into aspects of research that were previously difficult to study, such as the extent to which articles from a field attract *readerships* from other fields or the value of social media publicity for articles. Altmetrics also have the potential to be used as impact indicators for individual

researchers. But he warns that “this information should not be used as a primary source of impact information since the extent to which academics possess or exploit social web profiles is variable (Thelwall, 2014)”. Altmetrics are easily manipulated, “since social websites tend to have no quality control and no formal process to link users to offline identities” (Thelwall, 2014b). Haustein (2016) presents a good overview of the heterogeneity, data quality and dependencies of altmetric data.

Indicators derived from reference managers and reader libraries such as Mendeley and ResearchGate deserve special attention. They are reflections of the publications that are being prepared by the users of these services, and of the documents that they find relevant when they are working on their own papers or are expanding their researcher networks. Manuscripts in preparation are potentially predictors of emerging trends and future citation impact. But the outcomes of their analysis depend on readers’ cognitive and professional background and are not necessarily representative of the global research community.

Chapter 11 in Part IV presents a critical discussion of the historical context, the limits and the potential of altmetrics in research assessment. It underlines the potential of altmetrics as tools to link scientific-scholarly expertise with societal needs.

4.2.7 Webometric Indicators

Webometrics can be defined as “the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibliometric and informetric approaches” (Björneborn & Ingwersen, 2004). The term webometrics was first coined by Almind and Ingwersen (1997). One of the topics in webometrics is the development of a Web Impact Factor as an analogue of a citation impact factor. The Web Impact Factor (WIF) of an entity is defined as the ratio of the number of external web links—also denoted as inlinks, site citations or sitiations—to the entity’s website, divided by the number of web-pages, a measure of the volume of the entity’s Web contents. It is noteworthy how important the paradigm of the ‘classical’ journal impact factor is in the construction of indicators for new types of data sources, such as usage and web links.

Web impact measured by inlinks, and web activity reflected in the volume of web contents, play a key role in the Webometrics Ranking of World Universities, published by the Cybermetrics Lab, a research group of the Spanish National Research Council (CSIC) located in Madrid (Aguillo et al., 2006). But rather than calculating a ratio of these two measures, as is performed in the WIF, the ranking is based on a composite indicator in which normalized forms of the two components impact and activity are summed up with equal weights (Webometrics Ranking, n.d.).

Webometric measures give an indication of the presence or visibility at the Web. A positive feature of the measurement of impact via web links, especially when calculated at the level of scientific institutions, is that it does not merely relate to

their *research* mission, but also on their *teaching* and *services* activities, and in this sense has a broader scope than the calculation of bibliographic citation-based measures. Also, it extracts linkage data from a universe of Web documents that is much larger than that covered in the multi-disciplinary citation-indexes Web of Science, Scopus or Google Scholar.

However, the vast size of the source universe has its limitations as well, as there is no systematic information on Web sources in the way in which the citation indexes inform users about their source coverage, and there is no systematic quality control of the source covered. Another limitation is that the degree of an institution's Web presence depends upon her internal policies towards the use of the Web, and the propensity of its academic staff to communicate via the Web.

4.2.8 Economic Indicators

Economic and econometric tools aim to measure the effect of science on industry, innovation and the economy as a whole. Within this framework one finds on the metrics side technology transfer measures and patentability potentialities of a piece of research carried out, as well as cost-benefit measures. The global economic crisis of the past decade has brought this type of evaluation to the forefront. As a result, programs and institutions are not only evaluated on the basis of their contribution to science but also according to the level of their contribution to the industry and commerce.

Commercialization of research via patents, licenses and formation of start-up companies has been a topic of research and analysis for a long time (Chen, Roco, & Son, 2013; Huang et al., 2013). However, the use of these measures are now being brought forward into the evaluation arena because of two reasons: (1) the ability to collect and analyze large scale datasets including patents, financial and technical reports globally; (2) the increasing demand by the public and government to calculate cost-benefit measures of programs within scientific institutions especially those publicly funded.

The AUBR Expert Group mentions four indicators of innovation and social benefit: external research income, i.e., the level of funding attracted from external sources; the level of funding won competitively; research funding obtained per academic staff; and the employability of Ph.D. students. The first three are useful for comparing research performance across a national or global system and within universities, as the willingness of industry to pay for research is a useful indicator of its anticipated contribution to innovation and the economy. Industry employment of Ph.D. graduates can be an indicator of the contribution of research to the highly educated and skilled workforce.

An important group of indicators based on economic models measure research productivity or research efficiency, by relating research outputs to 'input' measures of the resources that were used to produce the outputs. Typical examples are the publication output or citation impact per FTE Research or per dollar spent. Abramo

and D'Angelo (2014; 2016) strongly promote this type of indicators, and propose the use at the level of individuals or groups of a ratio of the (normalized) citation impact they generated and the total amount of money spent on their salaries. Efficiency is beyond any doubt an important aspect of a research system and deserves a systematic analysis. But effectiveness in reaching external goals is equally important (Glanzel, Thijs, & Debackere, 2016).

The use of research productivity measures defined as output-over-input ratios involves a series of problems. In research assessment, the research group or laboratory is the best level but often no data is available for such entities. Also, there is a lack of standardization. Not only are input and output data collected independently and use different classification systems, but also comparisons between institutions from different countries are difficult to make, due to national differences in the classification of administrative and financial data. In addition, outputs follow inputs with an often unknown, variable time lag structure, and also influence input levels.

Rather than using a classical production function approach, Bonacorsi and Daraio (2004) explore the use of an advanced, robust, non-parametric efficiency analysis that overcomes many of the problems mentioned above. They conceive S&T production as a non-deterministic, multi-input, multi-output relation, in which both inputs and outputs are not only qualitatively heterogeneous but also sometimes truly incommensurable. Section 17.3 presents a more detailed overview of the potential of efficiency analysis, based on the work by Daraio and Bonacorsi.

4.2.9 Reputation and Esteem-based Indicators

A count of the number of prestigious prizes or awards won either in total or per academic staff is used as an indicator of research performance. The Shanghai Ranking of World Universities includes this indicator in its rankings (ARWU, 2016). Receiving a prestigious prize or award is a clear manifestation of esteem. But the absolute numbers tend to be low; an overwhelming majority of scholars has a zero score. And the evaluation processes on which the nominations are based are not always fully transparent.

The AUBR Expert Group underlines that there are no agreed equivalences that apply internationally and facilitate comparison across disciplines. The Expert Group stated:

Some research communities—astronomy for example—are coherent and international; others have not developed international organisations, awards and cultures. But the status of prizes and awards may be difficult to assess. Prizes and other honours are a strong feature of science, technology and medicine research communities. They reinforce the influence of the key organisations that assess potential recipients and serve to promote research to governments and the broader community. Other research fields, notably in the Arts, Humanities and Social Sciences, are less well organized and resourced, and less influential with government and industry, and have not developed awards and honours to this extent (AUBR, 2010, p. 74).

Reputation can be measured in surveys. In the QS and THE World University Rankings, the outcomes of reputation surveys play an important role. Reputation indicators are based on the number of mentions in surveys sent via email to a large sample of peers, users or customers. Mentions integrate a range of aspects of institutional performance. Moreover, surveys can be conducted by applying scientifically-founded methods developed in social sciences.

But they have severe limitations as well. Although the sample sizes tend to be large, the response rates are very low (e.g., Rauvangers, n.d.). As a result, biases may occur due to selective responses and ‘amplifications’ via social media. Also, mentions may be based on ‘hear-say’ rather than on founded judgement. They tend to be based on impressions, not on an ‘objective’ methodology. In addition, they may refer to performance in a distant past and have little relevance for the current situation. Chapter 18 in Part VI presents a critical analysis of five World University Rankings.

4.2.10 Indicators of Research Collaboration and Cross-Disciplinarity

An important aspect studied in a network analysis of co-authorship is *scientific collaboration*. If a research article is co-published by authors from institutions located in different countries, it can be interpreted as a sign of international collaboration. Co-authorship is a well-documented form of collaboration (Glanzel & Schubert, 2004). A base assumption holds that all authors of a publication have made a substantial contribution to the collaboration. Co-authorship measures may indicate the degree of integration in institutional, national or international research networks. Section 12.5 in Part IV further analyses international co-authorship, within the framework of an analytical model of development of scientifically developing countries. But not all types of collaboration result in co-authorship. More informal types, such as participation in discussions about the line of research at internal seminars, or providing comments to draft versions of papers, may not result in co-authorship.

Network analysis is one of the more recent methods used for scientific assessment. The technological capability to trace and calculate collaborations between institutions and individuals on a large scale and through the years enables evaluators to have a novel view on how institutions and individuals work as a part of the domestic and global research network (Bozeman, Dietz, & Gaughan, 2001; Martinez et al., 2003). It is assumed that institutions and individuals who develop and maintain a prolific research network are not only more productive but also more active, visible and established.

The network analysis also allows benchmarking to be performed by evaluators by comparing collaborating individuals or institutions to each other. This type of comparison puts their research output and impact in context of the domestic and

international disciplinary activity and allows for a better understanding of their rank among their peers. Such network analytics is done on publication level but also, recently, on social media and public domain level where the scientific community shares outcomes and accomplishments openly.

Research mobility is also among the network analytics methods (Zellner, 2003; Ackers, 2005). This method enables tracing an individual's affiliations through the years and look at his/her expertise building throughout a career. It is assumed that moving from one institution to another throughout different stages of one's career helps in expertise building and can result in high productivity. Of course, there are many challenges to this approach and its value is still examined. One of the main challenges resides in the fact that affiliations' names as mentioned in the published papers are not always standardized, thus making it difficult to trace. Another factor is that education in a different country which might not have resulted in a publication cannot be measured, thus making this particular expertise building impossible to trace (Moed & Halevi, 2014).

Cross-disciplinary research is research involving multiple scientific-scholarly disciplines. The AUBR Expert Group defines it as "research that employs the knowledge structures and characteristic behaviours of more than one discipline and may interrogate, critique and integrate specific disciplines and disciplinarity" (AUBR, 2010, p. 68). Mono-disciplinary research is conducted within the boundaries of a specific discipline, contributing primarily to the advancement of knowledge in that discipline; Trans- or multi-disciplinarity bring together two or more disciplines without integration, while inter-disciplinarity blends the approaches of two or more disciplines often leading to the creation of a new discipline (Bordons, Morillo, & Gomez, 2004). Wagner et al. (2011) give a review of the various approaches to the study of inter-disciplinary research, and Rafols et al. (2011) analyze the position of inter-disciplinary research in bibliometric rankings of journals.

An important indicator of cross-disciplinarity using citation data measures the disciplinary *breadth* of the impact of generated by a particular paper, group or institution, or the relevance of a piece of research for surrounding disciplines. It must be noted that an assessment of cross-disciplinarity assumes the availability of a valid, operational classification of research into scientific-scholarly disciplines.

4.2.11 *Indicators of Research Infrastructure*

As outlined in Sect. 3.4 in Part I, indicators of research infrastructure are *not* primarily performance indicators, but focus on *preconditions* to performance. They measure basic facilities that support research, the scale of the research activities, and their sustainability. They are extensively discussed in a report published by an Expert Group on University Based Research, installed by the European Commission (AUBR, 2010).

Important indicators of *research intensity*, indicating the *scale* of the research, are: the ratio of research students (or Ph.D. students) per academic staff (or per

'research active' staff); the number or percentage of early career researchers involved in research; and the number of Ph.D. and research Master degree *completions*. Also the total investment in R&D from all funding sources is an indicator of scale. A count of national and international collaboration with other universities and/or with public-private and NGOs is also seen as an indicator of research involvement and scale of activity, as research is increasingly conducted in collaborative teams, nationally and internationally. But research practices differ across disciplines. Large research teams or laboratories are found mostly in the natural and life sciences.

The number of laboratories, their total research funding, the number of books in the library and number of journals that are accessible online, the available super computing power, and the total value of facilities expressed in money are typical indicators of research infrastructure. The AUBR Report underlines that it is difficult to obtain reliable, comparable institutional data, as there is no agreement on the basis on which full cost of research investment should be calculated.

The number or equivalent full-time (FTE) of research active academics employed by a university, and the ratio of the number of research active academics per total academic staff are key indicators of research capability and intensity. 'Research active' is established by setting threshold levels of performance for a specific period. But there is no clear, generally accepted definition of research activeness across universities, countries and disciplines.

4.3 Big Informetric Data

Big data refers to data sets that are so large and complex that it becomes difficult to process them using on-hand database management tools or traditional data processing applications. The advent of super computers and cloud computing able to process, analyze and visualize these datasets has its effect also on assessment methods and models. While a decade ago, scientific evaluation relied mainly on citations and publications counts, most of which were even collected manually, today this data is not only available digitally but can also be triangulated with other data types. Table 4.1 presents examples of big datasets that can be combined in informetric studies to investigate various phenomena related to scholarly outputs and their impacts. These examples were extracted from Moed (2012).

Thus, for example, publications and citation counts can be triangulated with collaborative indicators, text analysis and econometric measures producing a multi-level view of an institution, program or an individual. The combination on an article-by-article basis of citation indexes and usage log files of full text publication archives, enables analysts to investigate the relationships between downloads and citations, and develop ways for evaluators to generate a more comprehensive, multi-dimensional view of the impact of publications than each of the sources can achieve individually. Combining citation indexes and full text databases enables one to study the citation context.

Table 4.1 The combination of large informetric datasets

Combined datasets	Studied phenomena	Typical research questions
Citation indexes and usage log files of full text publication archives	Downloads versus citations; distinct phases in the process of processing scientific information	What do downloads of full text articles measure? To what extent do downloads and citations correlate?
Citation indexes and patent databases	Linkages between science and technology (the science–technology interface)	What is the technological impact of a scientific research finding or field?
Citation indexes and scholarly book indexes	The role of books in scholarly communication; research productivity taking scholarly book output into account	How important are books in the various scientific disciplines, how do journals and books interrelate, and what are the most important books publishers?
Citation indexes (or publication databases) and OECD national statistics	Research input or capacity; evolution of the number of active researchers in a country and the phase of their career	How many researchers enter and/or move out of a national research system in a particular year?
Citation indexes and full text article databases	The context of citations; sentiment analysis of the scientific-scholarly literature	In which ways can one objectively characterize citation contexts, and identify implicit citations to documents or concepts?
Citation indexes and altmetrics data sets	The use of reference manager platforms such as Mendeley as data sources of bibliometric analysis	To which extent are readership counts in reference managers a good predictor of citation impact?

Yet, the availability and processing capabilities of these large dataset does not necessarily mean that evaluation becomes simple or easy to communicate. The fact of the matter is that as they become more complex, both administrators and evaluators find it difficult to reach consensus as to which model best depicts productivity and impact of scientific activities. These technological abilities are becoming breeding grounds to more indices, models and measures and while each may be valid and grounded in research they present a challenge in deciding which is best to use and in what setting.

Combining informetric datasets covering distinct dimensions is not new. In the 1960 Francis Narin and co-workers at Computer Horizons (later CHI Research) systematically linked a dataset with citations in patents processed for the US Patent Office with a large file of articles in journals indexed for the Science Citation Index (currently Web of Science). Chapter 15 in Part V of this book dedicates attention to their pioneering work that could be denoted as big data science ‘avant la lettre’.

4.4 Science Maps

Mapping of science and technology can be defined as the development and application of computational techniques for the visualization, analysis, and modeling of a broad range of scientific and technological activities as a whole. There is an increasing interest in mapping techniques, and science mapping is to be qualified as one of the most important domains of informetrics as a big data science. Figure 4.1 presents a typical example of a comprehensive map of science and scholarship as a whole, visualizing the cognitive linkages between the various research specialties and disciplines.

Figure 4.1 is based on a citation-based technique denoted as co-citation analysis. In a first step, a large dataset is compiled of research articles covering a specific subject (in this case science as a whole). Next, the most highly cited documents in the set are identified, and the relatedness between any pair of them is defined on the basis of their co-citation strength, i.e., the number of times they are co-cited in the same source article. Next, the highly cited documents are clustered on the basis of

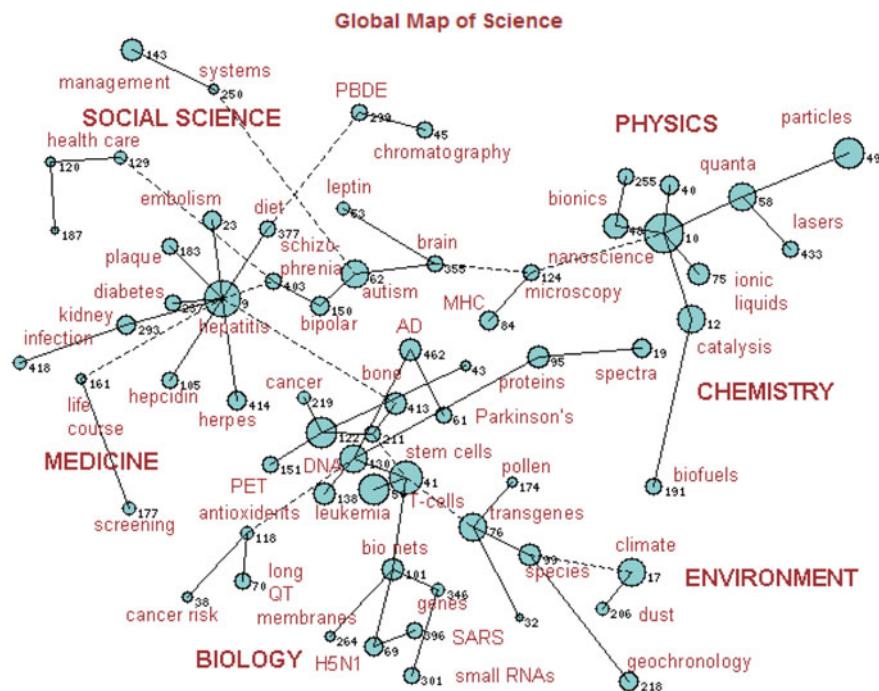


Fig. 4.1 Co-citation based research fronts (2004–2009). Based on Web of Science; Source ScienceWatch (2010), Essential Science Indicators, used by permission of Clarivate Analytics

their co-citation strength. Such clusters are denoted as research fronts. A typical example in Fig. 4.1 is the front *nanoscience*, located somewhere between the physics and the chemistry discipline. Finally, the relatedness between clusters is calculated. Chapter 13 in Part V presents a series of science maps and mapping software packages, and introduces them as the continuation of the creative work of one of the founding fathers of the ‘science of science’, Derek de Solla Price.

Chapter 5

Statistical Aspects

Abstract This chapter dedicates attention to three important statistical issues in applied informetrics: Are journal impact factors good predictors of the actual citation rates of individual articles? To what extent do errors or biases cancel out in large datasets? And how should one interpret linear or rank correlation coefficients?

Keywords Large samples · Pearson correlation coefficient · Skewed distribution · Spearman rank correlation coefficient

5.1 Introduction

As outlined in Chaps. 1 and 4, this book does *not* discuss technical details of indicators, their formal definition and their statistical properties. Other handbooks and literature reviews, cited in Sect. 4.1, do provide this information. The current chapter focuses on more general issues of interpretation of the outcomes of statistical analyses. Section 5.2 relates to journal impact factors, perhaps the most widely used informetric measure. If the values of the impact factors of two journals A and B are, for instance, 4 and 2, can it be safely assumed that a randomly selected article from A is more often cited than an arbitrary article from B?

Next, in Sect. 5.3 all users of informetric data are aware that the data may suffer from inaccuracies or biases. What are the implications for the usefulness of indicators based on such data? Does it mean that they are always inaccurate and therefore of little use? Can we safely assume that if the dataset is large enough, errors cancel out? Finally, in many statistical analyses, correlation coefficients between two or more variables are calculated, and their values may play an important role in debates about the interpretation of the outcomes of these analyses. Section 5.4 raises the question: what does a correlation coefficient mean?

5.2 Journal Impact Factors Are no Good Predictors of Citation Rates of Individual Articles

As a journal impact factor represents an average citation rate for articles published in a journal, an often heard claim is that it is a good predictor of the citation count of an individual paper in the journal. For instance, it is sometimes believed that if a journal's impact factor amounts to 4.0, it can be expected that papers published in the journal are cited about 4 times, and that most of them will have a higher score than articles published in a journal with an impact factor of 2.0. This section shows why these statements are invalid, following a line of argument presented in a report by the International Mathematical Union (Adler, Ewing, & Taylor, 2008).

Figure 5.1 gives the citation distribution for two journals in the field of mathematics, Proceedings of the American Mathematical Society (PAMS) and the Transactions of the American Mathematical Society (TAMS). The impact factor of the latter is about twice that of the former: 0.85 against 0.43. Both journals show a skewed citation distribution. Figure 5.1 shows for instance that about 70% of papers in PAMS receive no citations (during the impact factor window); for TAMS this percentage is about 50.

For comparison, Fig. 5.2 relates to the pairs of persons consisting of a soccer player and a junior kid (male or female) entering hand-in-hand the playing field prior to the start of a soccer match. It shows the distribution of the lengths of the players and the kids. Data are imaginary. It is assumed that the average length of a player is about twice that of an accompanying kid, and that the two distributions do not overlap.

Fig. 5.1 Citation distribution for two mathematical journals

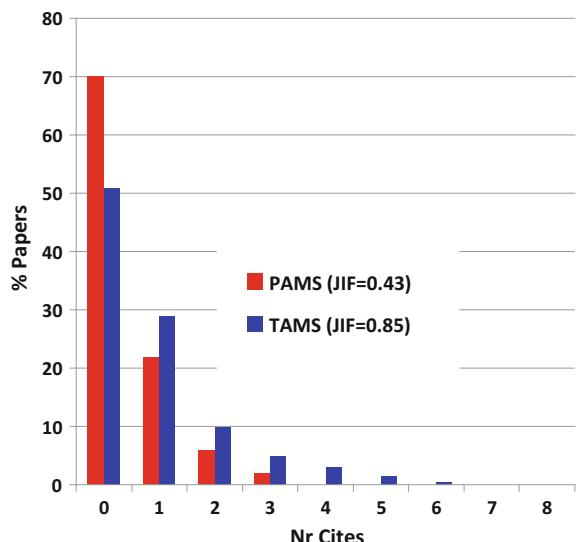
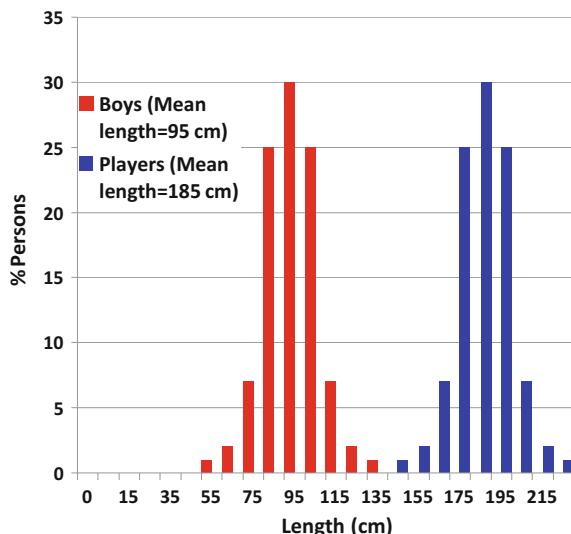


Fig. 5.2 Distribution of the length of kids and adults in pairs entering the playing field before the start of a soccer game



A first question relates to data on the soccer players and their kids. What is the probability that a randomly selected kid is at least as tall as a randomly selected adult player? The answer is: zero, as in this imaginary example no kid is taller than any of the players. The standard deviations are such that there is no overlap at all between the two distributions. But even if there were some overlap between the two, the probability at stake would be rather low. A more complex exercise should take into account this case.

A similar question is addressed for the citation count of articles published in the two journals, bearing in mind that, as in the case of the soccer game, the average of the TAMS distribution is about twice that of PAMS. However, the probability that a paper selected at random from PAMS (with impact factor 0.43) is cited at least as often as a randomly selected TAMS article (with impact factor 0.85) is 0.62. This means that, when picking up at random one paper from PAMS and one from TAMS, and comparing their citation rates, in 62 out of 100 cases the PAMS paper is cited at least as often as the TAMS paper.

The probability that a randomly selected paper in PAMS is cited at least as often as a randomly selected article in TAMS is calculated by summing up a series of probabilities that the various combinations of scores occur, namely, the probability that a PAMS paper is cited 0 times and at the same time the probability that a TAMS paper is cited 0 times, plus the probability that a PAMS paper has one citation and a TAMS paper zero citations or one citation, plus the probability that a PAMS paper has 2 citations and a TAMS article 0, 1 or 2 citations, and so forth.

The outcome presented above relates to one single case: a comparison of PAMS and TAMS. Figure 5.3 presents results related to all journal pairs for which the ratio of the impact factors of the constituent journals is around 2.0—as in the TAMS-PAMS comparison described above. As in the TAMS-PAMS case, for all

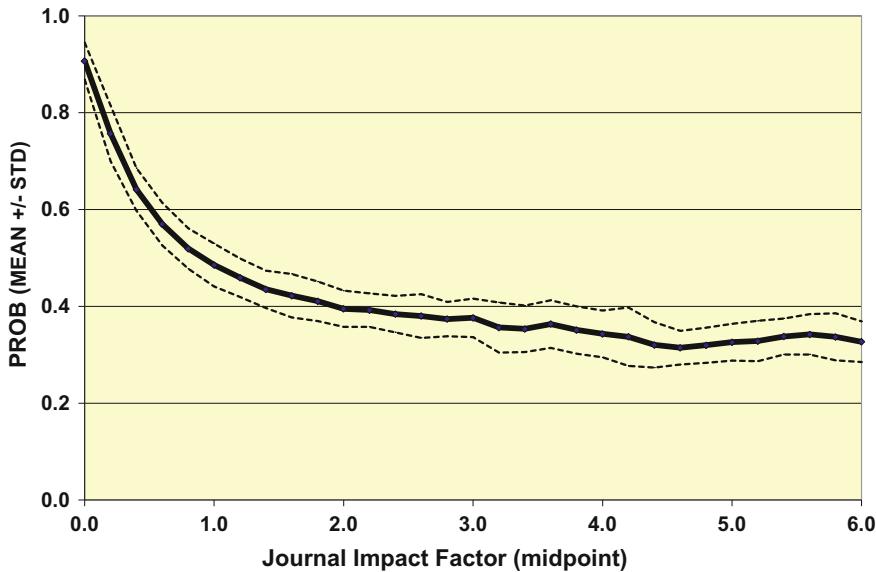


Fig. 5.3 Results similar to those presented in the TAMS-PAMS case for a large cross section of journals. Data were derived from Web of Science 2006–2008. The graph relates to journal pairs for which the ratio of the impact factors of the constituent journals is around 2.0 (as in the TAMS-PAMS comparison described above). The horizontal axis gives the midpoint of the impact factor of the journal with the lower score in a pair. For instance, midpoint 1.0 relates to impact factor values between 0.5 and 1.5. On the vertical axis, PROB indicates the probability that a randomly selected paper from the lower impact journal is cited at least as frequently as a random article from the higher impact outlet. The *solid line* gives the mean over all pairs, while the *dashed lines* indicate its standard error

such pairs the probability is calculated that the citation count of a randomly selected paper from the journal with the lower impact is at least as large as that of a random article from the journal with the higher impact. Figure 5.3 shows that this probability declines as a function of the impact factors of the journals, but obtains even for journals with an impact of 6.0 a level of around 0.33.

5.2.1 Aftermath

While preparing the final version of this book, a paper was published by Waltman and Traag (2017), presenting “a theoretical analysis of statistical arguments against the use of the impact factor at the level of individual articles”. The authors summarize their findings as follows.

Our analysis shows that these arguments do not support the conclusion that the impact factor should not be used for assessing individual articles. In fact, our computer simulations

demonstrate the possibility that the impact factor is a more accurate indicator of the value of an article than the number of citations the article has received. (Waltman & Traag, 2017, quote from the abstract)

They claim that, to justify the conclusion that the JIF should not be used at the level of individual articles, additional assumptions have to be made, for instance, “the assumption that citations accurately reflect the value of an article or the assumption that journals are very heterogeneous in terms of the values of the articles they publish”.

5.3 Errors or Biases in Data Samples¹

This section discusses two claims that are often heard in discussions about the statistical analysis of informetric data. The first is: informetric data tend to be so incomplete and affected by so many factors that have little to do with research performance, that it can never be used in a proper manner in research assessment. The second statement holds: enlarging the number of cases in an analysis dataset, for instance, by increasing the sample size but keeping the sampling technique constant, data errors and biases tend to cancel out; the larger the dataset one analyzes, the less the results are affected by errors. This section argues that both statements are incorrect, and finishes with a general warning against drawing invalid statistical conclusions based on informetric indicators.

Following an argument made by Harriet Zuckerman (1987) in her discussion with MacRoberts and MacRoberts (1987), “the presence of error does not preclude the possibility of precise measurement”, and that the issue at stake is whether or not biases *systematically* affect certain subgroups (Zuckerman, 1987, p. 331). For instance, life time citation counts are known to be biased in favor of researchers with long careers. Comparing the counts of two groups of researchers, the effect of *this* bias can be at least partly neutralized if it can be demonstrated that the age distributions of the members of the two groups are statistically similar.

The statement that enlargement of data samples tends to neutralize errors is statistically only valid for *random* errors, and if the sampling technique does not change. The notion that in large datasets random errors are to some extent neutralized is clearly reflected in Howard White’s interesting metaphor on the validity of co-citation analysis. He stated that co-citation maps provide an *aerial* view and measure a historical consensus as to important authors and works.

“When one sees that scores, hundreds, and even thousands of citations have accrued to a work, an author, a set of authors, it is [...] difficult not to believe that individual vagaries of citing behavior cancel each other out, corrected by the sheer numbers of persons citing. [...] Why not believe that there is a norm in citing – straightforward acknowledgement of related documents – and that the great majority of citations conform to it? (White, 1990, p. 91)”

¹This section re-uses with permission selected paragraphs from Moed (2005a).

But *systematic* errors or biases do *not* necessarily cancel out. Thus, it cannot a priori be assumed that any deviations of the norm cancel out when data samples are sufficiently large and the sampling method is kept constant. Using White's metaphor on the aerial view, one may ask whether the methodology generating a proper aerial view of a town also provides sufficiently detailed and valid information to describe and 'evaluate' an individual living in the town. This issue is particularly relevant in the use of citation indicators in research evaluation of individual entities such as authors, research groups or institutions. For instance, Waltman, van Eck and Wouters (2013) found systematic biases at the level of individual authors.

5.4 How to Interpret Correlation Coefficients?

The *Pearson* correlation coefficient is a measure of the strength of a *linear association* between two variables. If one draws a line of best fit through the data of two variables, the Pearson coefficient expresses how far away the data points lie from this line. The Spearman's rank correlation coefficient measures the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. If, as the one variable increases, the other decreases, the rank correlation coefficients will be negative. Spearman's Rho is often used to make the coefficient less sensitive to non-normality in the underlying distributions.

Figure 5.4 gives for three simple data patterns the corresponding Pearson (linear) and Spearman (rank) correlation coefficients. These two measures are often used coefficients of statistical correlation between two variables, especially in bibliometrics. X and Y are two arbitrary random variables. Their values are indicated in three scatterplots. The figure shows how radically the values of the correlation coefficients change by adding data points. In the upper figure the value of both coefficients is zero. There is no tendency that higher values of X are associated with higher—or lower—values of Y. The middle plot illustrates that even by adding one single data point, the linear correlation coefficient may substantially increase, in this case to a value of 0.85 which in this particular case appears to be statistically significant.

In the bottom graph five more data points are added. This graph shows two regions of data points, one at the bottom left, and one at the top right. Interestingly, calculating correlation coefficients *per data region*, the values zero would be obtained, as in the upper graph, for each region. But calculating measures for the entire set of data points, the Spearman rank correlation coefficient is now above 0.7 and statistically significant, while the Pearson coefficient now obtains a value of 0.94.

In more technical terms, the distributions of the variables X and Y in the two lower plots are far from normal. In the middle graph, extreme scores are added that can be denoted as outliers. In the lower plot both X and Y have a bi-modal distribution. This case then shows that it is important to take into account the shape of the underlying distribution. Focusing on the bottom graph, and supposing that X and Y relate to a particular object, say, a journal, and that X indicates a journal's

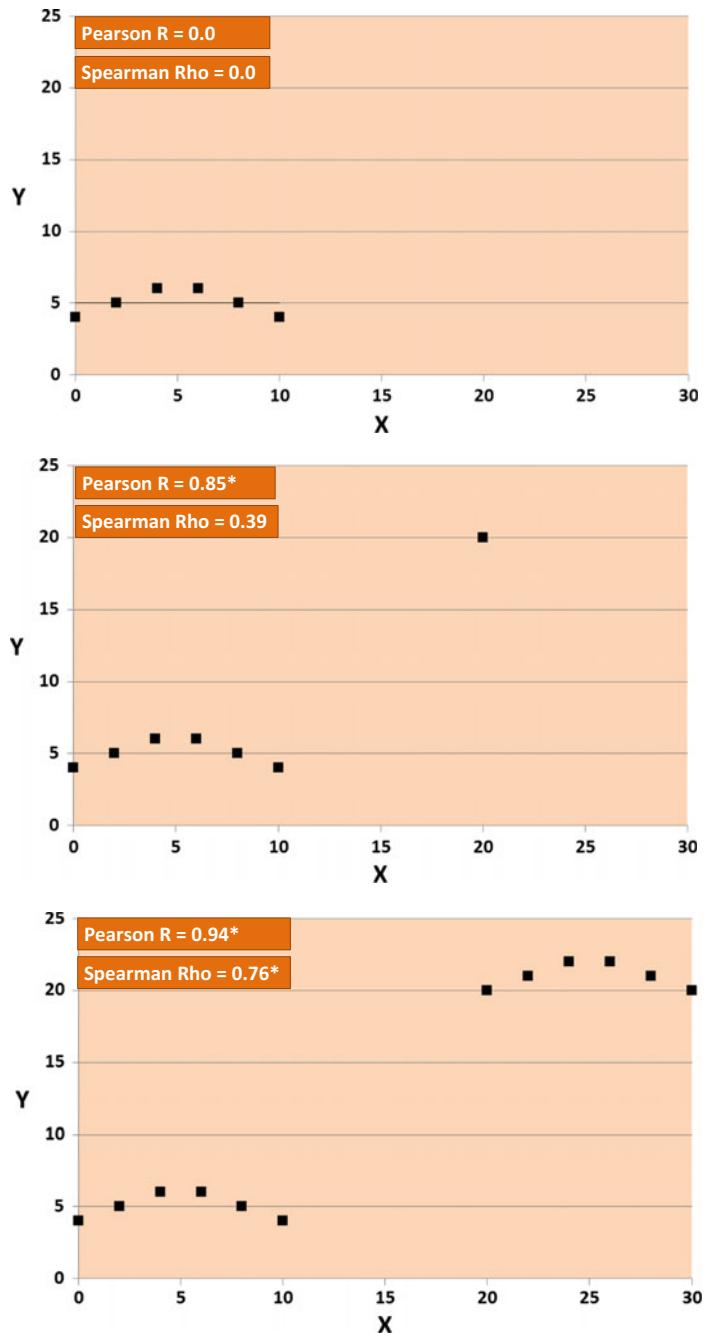


Fig. 5.4 Correlation coefficients for three sets of data points

impact factor, and Y the manuscript rejection rate in its peer review process, one would like to know in the case that journal J_1 has a higher impact factor than J_2 , whether it is probable that J_1 also has a higher rejection rate. Considering only the value of the correlation coefficient (Pearson: 0.94; Spearman: 0.76), one would perhaps conclude that this is highly probable. But this is only true if J_1 and J_2 belong to *different* data regions, one to the bottom-left, and the other one to the top-right region. If they are from the *same* region, the two variables do *not* correlate at all.

Apart from the shape of the underlying distributions, also the data range taken into account in the calculation of the coefficients is a crucial factor. An interesting example is given by Sugimoto et al. (2013). As outlined in Sect. 8.3, these authors examined the relationship between journal manuscript acceptance rates and journal impact factors, and found a statistically significant linear correlation coefficient between these two measures; but when journals were divided into quartiles on the basis of their impact factors, the correlation coefficients *per quartile* between acceptance rates and impact factors were much lower and not significant.

A rule with which probably all theoretical and applied statisticians agree is: never take a correlation coefficient for granted and never consider merely the numerical outcome. Always take into account a scatterplot of the underlying data, so that amongst other things the effect of non-normality of the distributions and the data ranges can be taken into account in the interpretation of the coefficient's numerical value. Also be aware that a lack of *linear* correlation does not necessarily mean a lack of correlation *as such*—as a correlation between two variables can be non-linear, and that statistical association of correlation does not necessarily mean *causation*. If X and Y correlate, it is logically possible that X causes Y , or Y causes X , or that there is a third variable, Z that influences both X and Y in the same direction.

Part III

The Application Context

Chapter 6

Research Assessment as an Evaluation Science

Abstract This chapter introduces evaluation science as a discipline, and explores its relationship with the field of quantitative research assessment. It presents an overview of major theoretical concepts and distinctions. The chapter continues with distinguishing four domains of intellectual activity in research assessment, namely the domains of policy, evaluation, analytics and data collection. It clarifies the relationships between these domains. Finally, it presents an overview of assessment models and strategies.

Keywords Analytical domain · Assessment strategies · Becker model · Comprehensive theory of change · Criterion-based reference framework · Data collection · Empowerment evaluation · Evaluation science · Evaluative domain · Evaluative judgments · Formative evaluation · Management tools · Normative reference framework · Operationalization · Performance monitoring · Policy domain · Quickies · Research tools · Summative evaluation

6.1 Introduction

Evaluation science is a multi-disciplinary subject field dealing with evaluation or assessment of complex systems and interventions. Its feeding disciplines are educational science focusing on teaching performance, business studies and management science dealing with business performance, and the sub-domains of medicine involved in the evaluation of medical treatments. The field publishes a series of specialist journals, some of which contain the terms evaluation or assessment in their titles. One of the aims of the current chapter is to illustrate how analytical distinctions and approaches in evaluation science field are potentially useful in research assessment as well, not only in the development and application of informetric research assessment tools, but also in building up a theoretical foundation of quantitative research assessment, and in the communication with users.

This does *not* mean that these approaches from evaluation science are superior to those currently developed in research assessment, or that they should become a

norm for methodology development in the latter domain. One reason for this is that the concepts of economic or teaching performance are perhaps more apt for quantification than the concept of research performance or quality. Perhaps the two disciplines could learn more from one another than they have done in the past. Section 6.2 gives an overview of a number of useful definitions of major concepts and analytical distinctions from the domain of evaluation science.

In Sect. 6.3 the field of quantitative research assessment is characterized as an evaluation science, using the notions and distinctions outlined in Sect. 6.2. It presents an analytical distinction into four domains of intellectual activity in an assessment process, and shows how they are related. It argues that while infor-metrics may provide useful analytical insights as inputs into an assessment process, it does itself *not* evaluate. Section 6.4 deals with assessment models. It presents details on three assessment models, one combining peer review and indicators, a second having its roots in psychology, and a third based on bibliometric model applied to medical research. Finally, Sect. 6.5 deals with the *costs* of assessment.

6.2 Evaluation Science

Table 6.1 presents a glossary of important terms and distinctions related to assessment.

6.2.1 *Research Versus Management Tools*

The distinction between *research tools and management or assessment tools* is well known in the field of quantitative science studies. This distinction clearly emerges from the following statements by Stephen Cole:

A crucial distinction must be made between using citations as a rough indicator of quality among a relatively large sample of scientists and in using citations to measure the quality of a particular individual's work (Cole, 1989, pp. 9, 11). In sociological studies our goal is not to examine individuals but to examine the relationships among variables (*ibid.*, p. 11). Citations are a very good measure of the quality of scientific work for use in sociological studies of science; but because the measure is far from perfect it would be an error to reify it and use it to make individual decisions. (*ibid.*, p. 12)

Citation indicators in *a scholarly research context* are used as tools in testing hypotheses or examining universal relationships among variables within a theoretical framework. It is the validity of a particular hypothesis that is at stake. In a *policy or management context*, citation indicators may be used in reaching some type of policy decision.

Table 6.1 Glossary of terms and distinctions related to assessment

Term or distinction	Definition
Monitoring	A continuous process of collecting and analyzing data in real time to understand how well an intervention, program or organization is executing against expected results in order to make both tactical and strategic adjustments
Performance measurement	Data collected against a system of indicators about aspects such as costs, inputs, activities, quality, outputs, and outcomes
Performance monitoring ← → evaluation	Performance monitoring investigates present operations with a forward-looking view toward bettering quality, enhancing efficiency, and improving sustainability and results using small feedback loops. Evaluation is concerned with the backward-facing perspective of understanding the worth of what has been accomplished
Evaluation	Entails the systematic assessment of a planned, ongoing, or completed intervention's design, implementation, and results, to determine the fulfilment of goals and objectives, efficiency, effectiveness, impact, and sustainability
Evaluative criteria	The values (i.e. principles, attributes or qualities held to be intrinsically good, desirable, important and of general worth) which will be used in an evaluation to judge the merit of an intervention
Comprehensive theory of change	Specifies the domain[s] of the organization's focus, intended outcomes, and codified activities to produce them—along with the capacities and competencies needed to work accordingly
Research ← → management tools	Research tools have the primary aim to generate robust knowledge; management tools aim to be useful pragmatic tools for management purposes in organizations
Attribution ← → contribution	Attribution identifies a factor as being solely or mainly responsible for a particular outcome; contribution highlights a factor's positive effect among other influential factors
Summative ← → formative evaluation	In summative evaluation the focus is on the outcome of a process or program. Formative evaluation assesses the unit's development at a particular time
Normative ← → criterion reference framework	Normative: comparing each unit against all others; Criterion evaluating units according to the same criteria

Table 6.1 is based on various sources, mostly on the chapter Performance management and evaluation: Exploring complementarities (Nielsen & Hunter, 2013) in the edited work *Performance management and evaluation* (Nielsen & Hunter, 2013)

6.2.2 Comprehensive Theory of Change

A comprehensive theory of change is a key concept in evaluation science (Hunter, 2006). Its articulation is a sine-qua-non in any evaluation process. Without it there is nothing to evaluate. In the case of a research group it would include the group's research programme, objectives and milestones. For an academic institution it

specifies its profile, short and longer term objectives. Hunter and Nielsen define the concept as follows.

Rarely is measuring itself the critical challenge. Much more fundamentally, an organization needs a comprehensive theory of change (specifying the domain[s] of its focus, intended outcomes, and codified activities to produce them—along with the capacities and competencies needed to work accordingly). Such a theory of change must, in effect, be the blueprint adopted by the organization to manage toward success [...] and as such it will specify what indicators the organization must track to manage program costs, quality, and effectiveness. (Hunter & Nielsen, 2013, pp. 13–14)

Principal components of a theory of change are: inputs, activities, outputs and outcomes or impacts, both at a short and a longer term. Some authors relate a “theory of change” explicitly to the planning and evaluation of programs promoting social change in the philanthropy, not-for-profit and government sectors (e.g., Theory of Change, n.d.).

6.2.3 Performance Management Versus Evaluation

Hunter & Nielsen define the concepts of performance management and evaluation in the following manner.

Performance management, then, is the set of self-correcting processes, grounded in real-time data measuring, monitoring, and analysis, that an organization uses to learn from its work and to make tactical (front line, quotidian) and strategic adjustments to achieve its goals and objectives” (Hunter & Nielsen, 2013., p. 10). Evaluation [...] entails the systematic assessment of a planned, ongoing, or completed intervention’s design, implementation, and results, to determine the fulfilment of goals and objectives, efficiency, effectiveness, impact, and sustainability. (*ibid.*, p. 14)

According to Hunter and Nielsen, a first key difference between performance monitoring and evaluation lies in the direction of the perspective: performance monitoring has a forward-looking, and evaluation a backward-looking view. A second difference relates to the role of evaluative criteria. In evaluation they play a central role, while in the process of performance monitoring through real-time data collection and analysis of indicators these criteria remain implicit, even though they are embodied in the implementation of a monitoring system and its indicators.

This interpretation is in line with the definition of evaluative criteria in the UNICEF Glossary of terms in evaluation, namely as “the values (i.e. principles, attributes or qualities held to be intrinsically good, desirable, important and of general worth) which will be used in an evaluation to judge the merit of an intervention” (UNICEF Glossary, n.d.).

6.2.4 Summative Versus Formative Evaluation

Table 6.2 presents key characteristics of summative and formative assessment in educational science. In discussions about informetric assessment of research performance, especially about the role of citation analysis, it is not always clear which of the two type of assessment one is aiming at. Informetric indicators can be used either way. For instance, when they are used by researchers as *self-assessment* tools, enabling one to assess the effectiveness of one's publication strategy, the type of assessment is *formative*. When they are used to decide about selection or promotion of individuals, the approach is *summative*.

6.2.5 Normative Versus Criterion Based Reference Framework

A distinction between a normative and a criterion-based reference framework is most relevant in informetrics. For instance, in rankings of world universities on the basis of a particular indicator, the reference framework is *normative*, as the score of one university is compared to all others. On the other hand, the use of a minimum threshold for publication output of individual researchers for all researchers in a university or country, is a typical example of *criterion-based* assessment, as there is one single yardstick against which each individual is compared.

6.2.6 Evaluation Versus Assessment

In many contexts, the terms assessment and evaluation are considered synonyms. In fact, in the New Oxford American Dictionary the definition of ‘assessment’ uses the word ‘evaluation’ and vice versa. In the Cambridge English Dictionary, the definitions are almost identical: evaluate means “to judge or calculate the quality, importance, amount, or value of something” and assess “to judge or decide the amount, value, quality, or importance of something”.

Table 6.2 Main differences between summative and formative assessment

	Summative assessment	Formative assessment
Time	At the end of a learning activity	During a learning activity
Goal	To make a decision	To improve learning
Feedback	Final judgement	Return to material
Frame of reference	Sometimes normative (comparing each student against all others); sometimes criterion	Always criterion (evaluating students according to the same criteria)

Source https://en.wikipedia.org/wiki/Formative_assessment

But in specific cases only one of the words may be used. The meanings of the terms are differentiated in research disciplines (e.g., educational research, management science, quantitative science studies), English language version (English vs. American) and professional settings (e.g., teachers vs. examination boards).

In the literature on evaluation or assessment, the term evaluation is often defined as the process of making *judgments* based on criteria and evidence of the ‘worth’, ‘value’ or ‘merit’ of an activity. This convention is in line with a definition found in Wikipedia: “Evaluation is a systematic determination of a subject’s merit, worth and significance, using criteria governed by a set of standards” (Evaluation, n.d.). But in the OECD-DAC Glossary of Key Terms and Concepts, the word evaluation has a much broader meaning, and is used as an *overarching* term:

Evaluation: The systematic and objective assessment of an on-going or completed project, programme or policy, its design, implementation and results. The aim is to determine the relevance and fulfilment of objectives, development efficiency, effectiveness, impact and sustainability. [...] Evaluation also refers to the process of determining the worth or significance of an activity, policy or program. An assessment, as systematic and objective as possible, of a planned, on-going, or completed development intervention. (OECD Glossary, n.d.)

The Research Methods Knowledge Base (n.d.), a web-based textbook available at the Web Center for Social Research Methods, criticizes the definition of evaluation as ‘the systematic assessment of the worth or merit of some object’, as many types of evaluations exist that do not necessarily result in an assessment of worth or merit, for instance, ‘descriptive studies’.

In the current book the term *assessment* is used as an *overarching* concept, denoting the total of activities in assessment or evaluation processes, or the act of evaluating or assessing *in general*. Table 6.3 in Sect. 6.3 distinguishes *four* types of intellectual activity—denoted as domains—in the assessment process. It shows that the term *evaluation* is reserved to indicate one of these domains only, namely the activity dealing with the definition of evaluation criteria and the formation of *judgments* of the ‘worth’ of a subject.

6.3 Types of Intellectual Activity in Assessment

To illuminate the differences between the various definitions of the terms evaluation and assessment, and to create a basis for a further theoretical development in evaluation science, the current author proposes a framework that distinguishes four domains of intellectual activity in an assessment process, and that enables one to clarify the relationships between these domains. The distinction is *analytical*. It should *not* be interpreted in terms of a *chronology* of phases in an assessment process. In addition, this table is *not* a sociological categorization of *actors*—individuals or their organizations. One and the same actor may be active in multiple domains. For instance, an analyst may both collect data, develop an analytical

Table 6.3 Four domains of intellectual activity in research assessment

Domain	Policy or management	Evaluation	Analytics	Data collection
Orientation	Management; Decision making	Evaluative: defining and evaluating worth	Empirical	Technical
Activity	Formulation of a constituent policy issue and assessment objectives	Specification of the evaluative framework, i.e., a set of evaluation criteria in agreement with the constituent policy issue and assessment objectives	Specification of an assessment model	Collection of relevant data for analytical purposes, as specified in the analytical model
	Decision on the assessment's organizational aspects and budget		Collecting, analyzing, and reporting empirical evidence	Data can be either quantitative or qualitative
Outcome	Making a policy decision based on the outcomes from the evaluative domain	Making judgments on the basis of the evaluative framework and the empirical evidence collected	Writing an analytical report as input for the evaluative domain	Creation of a dataset with underlying data for all indicators specified in the analytical model

framework and calculate indicators. A manager may both formulate the key issue to be assessed, and specify the evaluative framework.

In the *policy or management* domain of an organization a constituent issue emerges. Constituent means that the assessment process is centred around this issue. For instance, the director of a funding agency may aim to assess the funding policies of his organization. Once an evaluation is completed, its policy implications have to be established in this domain, possibly leading to the implementation of new policy measures. Most importantly from a practical point of view, in this domain decisions are made on organizational aspects of the assessment process, including its budget.

A next domain is denoted as the evaluative domain, in which the specification is made of the *evaluative framework*. It specifies the ‘values’ (i.e. principles, attributes or qualities held to be intrinsically good, desirable, important and of general worth), as well as the standards against which they are to be evaluated. Such values are linked to the (longer term) goals and (short term) objectives of the subject of assessment. It specifies for instance the relative weights of attributes, and defines proper benchmarks. When the analytical work is completed, *evaluative judgments*

are made on the basis of the evaluative framework and the empirical evidence collected. Such judgments are the input for decision making in the policy domain.

Next, the *analytical domain*, aims to collect empirical knowledge on the subjects under assessment. It specifies the assessment *model or strategy* that determines how empirical knowledge is being collected, one of the core activities of evaluation science. It is further discussed in Sect. 6.4. Next, the analytical domain collects, analyses, and reports empirical evidence on the basis of an analytical model that embodies the transition from raw data to indicators. A key task is to develop methodologically valid *operationalizations* of the qualities and benchmarks specified in the evaluative framework. The definition of ‘indicator’ follows Gerald Holton who defined this concept as follows:

I propose that the term indicator is properly reserved for a measure that explicitly tests some assumption, hypothesis or theory; for mere data, these underlying assumptions, hypotheses or theories usually remain implicit (Holton, 1978, p. 53). The indicators cannot be thought of given from ‘above’, or detached from the theoretical framework, or as unable to undergo changes in actual use. They should preferably be developed in response to and as aids in the solution of interesting questions and problems. (*ibid.*, p. 55)

From the analytical model follows a specification of the data that are needed to calculate the indicators specified in the analytical model. This data can be quantitative or qualitative, derived for instance from existing big databases such as the Web of Science, Scopus or Google Scholar, or, for instance, from questionnaire-based surveys or interviews.

As indicated above, an actual evaluation process may not develop chronologically along the analytical distinctions made above. An evaluation process may not be a linear process, and feedback loops play a key role. It may occur that a policy agency starts with preliminary articulation of a specific need, which is re-formulated in interaction with evaluators, evaluated subjects and informetricians. Also, an indicator considered useful in the development of an analytical model may, after consultation with a data expert, be replaced by another proxy since the underlying data needed to calculate it is not available.

A main objective of Table 6.3 and its analytical distinctions is to distinguish between scientific-methodological principles and considerations on the one hand, and policy-related, political or managerial considerations on the other. It enables one to formulate more precisely the statement that ‘informetricians should not sit in the chairs of managers or politicians’ and vice versa, and to further characterize the nature of the intellectual activities of the various participants in a research assessment process.

The current author defends the position that what is of worth, good, desirable, or important in relation to the functioning of a subject under assessment, *cannot* be established in informetric, or, more general, in quantitative-empirical research. The principal reason is that one cannot infer what *ought to be* from what actually *is*. What informetric investigators *can* do is empirically examine value *perceptions* of researchers, the conditions under which they were formed and the functions they fulfil, but the foundation of the validity of a value is *not* a task of

quantitative-empirical, informetric research. The same is true for the formation of *evaluative judgements*. The latter two activities belong to the domain of evaluation.

A key task of informetricians is to examine *indicator validity*, defined as the degree at which an indicator measures the aspect it claims to measure. For instance, are citation counts a valid indicator of research quality? Is co-authorship a valid indicator of scientific collaboration? These are *scientific-methodological* issues to be investigated in the domain of quantitative-empirical S&T studies, and decided on the basis of scientific-scholarly criteria.

Informetricians should maintain a *methodologically* neutral position towards the constituent policy issue, the criteria specified in the evaluative framework, and the goals and objectives of the assessed subject. As professional experts, their competence lies *primarily* in the development and application of analytical models *given* the established evaluative framework. To the extent that informetricians make ‘normative’ assumptions while assessing a unit’s worth, they should make these explicit and give them methodologically speaking a hypothetical status.

This methodological requirement does *not* imply that informetricians are not allowed to have normative views on the assessment exercise itself, but rather that they should make these views explicit as being their own, make clear that these normative views do not logically follow from their informetric research, and in this way put brackets around these views. Full neutrality is impossible; after all, informetricians are scholars themselves, and any assessment methodology could in principle be applied to their *own* performance as well. The complex status of informetricians as developers of analytical models and indicators *and at the same time also* as potential subjects of an assessment process, is further discussed below in Sect. 7.3.

Above it was stated that informetricians’ competence lies *primarily* in the development and application of analytical models embedded within a given evaluative framework. The word ‘primarily’ should still be explained. Informetricians have another competence, namely to critically discuss the constituent policy issue *itself* as well as the proposed evaluative framework, develop alternative ones, even suggest alternative policy questions, and in this way *enlighten* the debate about the assessment process as a whole, and inform policy makers or managers about alternative set-ups. In Chap. 10 the current author makes a series of suggestions for new elements in an assessment process, focusing on the *informetric feasibility* of such elements.

If the formation of an evaluative judgment is not a task for informetric research, the question arises how this type of intellectual activity should be characterized. If it is not an act of quantitative-empirical reasoning, what is it then? It integrates knowledge—arguments and facts—from a variety of disciplinary backgrounds and insights collected in practical experiences within an evaluative framework. It can be further characterised by taking into account another relevant distinction, namely between evaluation activity and *performance measurement*.

Table 6.1 defined performance measurement as data collected against a given system of indicators about aspects such as costs, inputs, activities, quality, outputs, and outcomes. This activity is best positioned in the analytical domain, powered by

the analytical model. In other words, it is one out of a series of possible components in an assessment process. Obviously, this model is related to an evaluative framework, but the relationship to this framework is loose or implicit. Performance measurement produces information, but as a stand-alone activity it does *not* evaluate in the sense of making evaluative judgements on the basis of an explicated evaluative framework.

A general conclusion is that although performance measurement, or informetrics can be applied in an evaluative context, it does *not* evaluate worth itself. It provides empirical evidence using analytical models. The current author takes the position that an assessment process should in principle be conducted within an explicit, evaluative framework. It should state in advance what is being assessed and why, which values are at stake, how they are assessed, which types of judgments are formed and on the basis of which grounds.

6.4 Assessment Models and Strategies¹

Informetric indicators, including the important group of citation-based measures, have become widely available, both in scholarly literature retrieval tools such Web of Science, Scopus and Google Scholar, but also in specialized indicator products such as Thomson Reuters' Incites, Elsevier's Scival, Scimago Journal and Institute Rank, and numerous World University Rankings. Also, there are many publications by experts in the field presenting new indicators, and also general principles of good practice in the use of these indicators (e.g., Hicks et al., 2015). An extensive review of the role of metrics in research assessment was published by an independent review group in 2015 (Wilsdon et al., 2015).

6.4.1 Base Distinctions

The AUBR Report (AUBR, 2010) makes two base distinctions regarding the types of assessment models. The first relates to the assessment method itself, namely, between *peer review*, providing a judgment based on expert knowledge, and a *metrics-based assessment*, using various types of indicators, including bibliometric and econometric indicators, but also end user reviews measuring, for instance, customer satisfaction. A second distinction relates to the *role of the research unit under assessment* in the evaluation process. It differentiates between self-evaluation—defined as a form of self-reflection which involves critically reviewing the quality of one's own performance—and external evaluation, conducted by an external

¹Sections 6.4 and 6.5 re-use with permission selected paragraphs from Moed and Halevi (2015).

evaluation agency. Methods are often combined, and are then to be viewed as components of an integral research assessment methodology.

A recent textbook on evaluation (Webcenter for Social Research Methods, n.d.) introduces the term ‘evaluation science’, and qualifies it as ‘a specific form of social research’. It distinguishes four classes of *evaluation strategies*, namely scientific-experimental, management-oriented systems, qualitative anthropological, and participant oriented strategies. The first take their values and methods from the sciences, and focus on impartiality, accuracy, objectivity and the validity of the information; the second class comprises management-oriented models from systems theory. An anthropological approach emphasises the importance of observation and of space for subjective judgment, while participant-oriented strategies underline the central role of evaluation participants, such as clients or consumers.

The challenges mentioned earlier, including the evaluation method, data, and analytics selection, have brought forth the development of hybrid assessment models. These models aim to build a modular method combining different measures and approaches depending on the field and target of assessment. These models were built for specific disciplines or areas of investigation answering challenges arising there. However, their modular nature and comprehensive approach demonstrate the importance of utilizing a variety of measures, models, and methods in order to accurately capture the impact and productivity of institutions, programs, and individuals. Three models adopting different approaches are described below. These examples are *illustrations of possible* assessment models. A complete overview of assessment models and a discussion of their pros and cons fall beyond the scope of this book.

6.4.2 Assessment of Basic Science Combining Peer Review and Bibliometric Indicators

A typical example of assessments combining peer review and bibliometric methods is the system of assessments of academic research conducted in the Netherlands during the 1990s and later, established by the Organisation of Universities in the Netherlands (VSNU) and the Netherlands Organisation for Scientific Research (NWO). Within the VSNU framework, peer review committees consisting of about 10 experts assessed all academic research groups in all disciplines. In several assessments of physics, chemistry or biology research, one of the information sources used by the Peer Review Group was a report presenting bibliometric indicators produced by the Leiden Centre for Science and Technology Studies.

Basic features of this procedure were that the research groups under assessment were aware from the very beginning that peer review committees responsible for the evaluation would use bibliometric indicators. They had the opportunity to verify the underlying data, and received the bibliometric outcomes related to their own activities. In addition, members of each group obtained an anonymous overview of

the scores of all groups subjected to the analysis, enabling them to position their own group. Slide 17.4 in Chap. 17 shows a typical example of a graph giving such an overview. A group had the opportunity to comment on the bibliometric outcomes, and its comments subsequently constituted a distinct source of information in the peer review process. Finally, bibliometric investigators had the opportunity to present their study to members of the review panel, to underline potentialities and limitations, and indicate possible pitfalls (see Moed, 2005a, pp. 239–245).

The current author considers this approach as a good example of how peer review and informetric indicators can be combined. A systematic evaluation of this approach falls beyond the scope of this book. One key issue would be the extent to which the formal use of indicators has increased the transparency of the process, especially of the motivation of the review groups' quality ratings.

6.4.3 Program Assessment: Empowerment Evaluation (EE)

Empowerment Evaluation is a flexible, goal-oriented approach to evaluation that puts an emphasis on the people involved. It places both evaluators and those being evaluated on the same levels of involvement and commitment to the success of individuals, programs, and institutions. This method was conceived in 1992 (Fetterman, 1994) and has been continuously developed since. It can be applied to a variety of activities and performances, not merely to scientific research. Empowerment Evaluation works on several principles which mainly aim to have all involved (evaluators included) as stakeholders in the evaluation process. Table 6.4 summarizes the EE principles.

The unique attributes of this model are in its capability to combine social, humanistic, and traditional evaluative approaches in a holistic manner. In addition, this model can be easily implemented in different areas; from scientific to social programs to industry performance and more (Wandersman & Snell-Johns, 2004; 2005; Fetterman & Wandersman, 2007). The stakeholders are responsible for the selection of methods and metrics appropriate to the purpose of assessment and take an active part in not only collecting and analyzing related data but also understanding it and implementing improvements processes.

6.4.4 Field-Specific Evaluation: The Becker Model

This model was developed by Cathy Sarli and Kristi Holmes (Sarla & Holmes, n.d.) at the Bernard Becker Medical Library at Washington University. It aims to provide a framework for tracking diffusion of research outputs and activities and identify indicators that demonstrate evidence of biomedical research impact. It is intended to be used as a supplement to publication analysis. The model consists of four dimensions within which a variety of indicators are utilized: (i) research output;

Table 6.4 Summary of Empowerment Evaluation (EE) principles

Category	Principle
Core Values	EE aims to influence the quality of programs
	Power and responsibility for evaluation lies with the program stakeholders
	EE adheres to the evaluation standards
Improvement-oriented culture	Empowerment evaluators demystify evaluation
	Empowerment evaluators emphasize collaboration with program stakeholders
	Empowerment evaluators build stakeholders capacity to conduct evaluation and use results effectively
	Empowerment evaluators use evaluation results in the spirit of continuous improvement
Developmental process	EE is helpful in any stage of program development
	EE influences program planning
	EE institutionalizes self-evaluation

Source Wandersman et al. (2004). Empowerment evaluation: Principles and action. Participatory community research: Theories and methods in action. Washington, DC: American Psychological Association, p. 141

(ii) knowledge transfer; (iii) clinical implementation; and (iv) community benefit. Indicators that demonstrate research impact are grouped at the appropriate stages along with specific criteria that serve as evidence of research impact for each indicator. By using a multilevel approach to evaluating biomedical research impact, the model aims to be able and assist scientists and program managers identify return on investment, quality of publications, collaboration opportunities, to name a few (Becker Model, n.d.).

6.5 Costs of Research Assessment

A well-constructed and executed research evaluation process incurs significant costs regardless of the methodology used. Hicks (2012), reviews some of the costs related to the implementation and use of performance-based university research funding systems (PRFSs) which are national systems built to evaluate a country's scientific output in order to better allocate funding based on performance. Since it is difficult to estimate the cost of each research evaluation method (Hicks, 2012, p. 256), taking the PRFS and their related costs, provides an understanding of the investment needed in order to conduct a sound scientific evolution as these systems include different data and analytical methodologies (Hicks, 2012, p. 255).

Peer review-based PRFSs incur indirect costs which include large panels of experts' time needed for the compilation and review of a university's output and faculty time needed for preparing submissions. However, indicators-based systems

also incur costs mainly due to the need of building, cleaning and maintaining a documentation system, purchasing citations data from vendors and developing calculation algorithms (Hicks, 2012, p. 258). This is true for any data-based evaluation method. Procuring the data, cleaning it and embedding it in a sound system are only part of the costs involved. Developing advanced algorithmic calculations of the data that will provide a true view of a country or an institution scientific output require expert opinion and know-how which come at a cost as well.

These expenses, as well as locating and engaging with expert reviewers, resulted in what is referred to by van Raan (2005) as “quickies”, i.e. rapid, cheap evaluations based on basic documents and citations counts with the help of standard journal impact factors (van Raan, 2005, p. 140). As van Raan notes:

Quite often I am confronted with the situation that responsible science administrators in national governments and in institutions request the application of bibliometric indicators that are not advanced enough...the real problem is not the use of bibliometric indicators as such, but the application of less-developed bibliometric measures. (pp. 140–141)

Therefore, when considering an evaluative method and especially one that requires a combination of more than one methodology or data type, one has to carefully estimate and calculate the costs involved. From data purchasing to systems development to expert reviewers; all involved will require appropriate funding in order to avoid a ‘quick and cheap’ evaluation exercise that might hinder an adequate assessment.

Chapter 7

Non-informetric Factors Influencing Indicator Development

Abstract This chapter illuminates non-informetric or extra-informetric factors that influence the development of indicators. It illustrates how under the surface of a technical-statistical debate about performance indicators a confrontation takes place between distinct views on what constitutes research performance. It also shows how the development of indicators fits into the wider socio-political context, even though in recent years this development is more and more influenced by internal dynamics.

Keywords Academia.edu · Age normalization · Altmetric.com · Benchmark set · Business interests · Clarivate analytics · Context of discovery · Context of justification · Declining groups · Elsevier · Emerging groups · Extra-informetric factors · Facebook · Field normalization · Fractional scientific strength · Google books · Google+ · Harvard level · Holtzbrinck publishing group · ImpactStory.com · Incites · Long term impact · Plum analytics · Research excellence · Research front · Research group · RG score · Saturation · Scival · Short term impact · Size normalization · Sleeping beauty · Thomson reuters · Top citation percentiles

7.1 Introduction

Section 6.3 presented a definition of indicators proposed by Gerald Holton (Holton, 1978), underlining that an indicator cannot be detached from a theoretical framework, and should preferably be developed as a tool in response to the solution of particular problems. The current chapter further elaborates on this notion along two lines. Section 7.2 illustrates how in seemingly technical discussions on the construction and statistical properties of science indicators, '*evaluative*', *theoretical assumptions on what constitutes research performance* play an important, though often implicit, role. Such values are denoted as extra-informetric, as their validity cannot be

grounded in informetric research. In this way, Sect. 7.2 *disentangles* informetric arguments and evaluative principles, one of the key objectives of this book.

Section 7.3 focuses on the *wider socio-political context* in which indicators were developed. It seeks to identify the context of their launch, not so much in terms of the *intentions* of the developers, but, at a higher analytical level, according to how they fit into—or are the expression of—a more general tendency in the policy, political or even cultural environment in which they were developed. This does *not* mean that such tendencies are the cause of their development, and that developers are merely instruments of an ‘external’ historical change. On the contrary, the developers actively contribute to these tendencies and make these real.

This section presents a brief history of some of the main lines in bibliometric indicator development from the early 1960s up to date. It argues that in the early decades newly proposed indicators aimed to solve specific policy problems and fitted in with specific national or institutional policy contexts, but during the past 10–15 years two tendencies emerged: on the one hand, previously developed indicators were used in more and more policy contexts, including contexts in which they are only partially or hardly fit-for-purpose; on the other hand, indicator development was more and more driven by internal dynamics powered by mathematical-statistical considerations.

7.2 How Evaluative Assumptions Shape Indicators

7.2.1 *Size Dependent Versus Size Independent Indicators*

A typical example of a size-dependent indicator is the number of articles published by a journal, or the total number of citations received by these articles. Some scientific journals publish several hundreds of articles per year, and others a few dozen. To be able to compare big and small journals, the journal impact factor was developed, calculating an average number of received citations per article published in a journal. But ‘size’ can also be a manifestation of performance. In the case of journals, it can be maintained that successful journals of high quality tend to attract more submissions and hence may publish more papers than their less successful counterparts do. According to this view, the size normalization applied by calculating an average ratio, in a sense ‘normalizes away’ a part of the journal’s performance.

This may also happen in the assessment of research groups of departments. By calculating the average number of received citation per published article, or alternatively, the number of citations per full time equivalent (FTE) research staff, a correction is made for differences in size. But the number of academic staff in a group, and the number of articles it published, depends upon the amount of research funding obtained, which in its turn can be assumed to depend at least partially upon a group’s research quality. Again, a part of what one aims to measure is normalized away.

How to deal with ‘size’ is one of the key issues in informetric indicator development. During the past years, an intensive debate has taken place among indicator developers about which indicators are the most appropriate for measuring research performance. While focusing on statistical aspects, this debate confronts at the same time distinct, partially conflicting, theoretical notions of what constitutes research performance, and distinct application contexts. Table 7.1 gives an overview of the position of some of the participants. More precise definitions of the indicators in this table can be found in Sect. 17.1 in Part V.

Table 7.1 shows differences in statistical assumptions. While Glanzel states that the use of the mean to characterize the citation distribution is statistically valid provided that the sample size is sufficiently large (above 50), Leydesdorff and Bornmann reject this claim. Important work by Waltman et al. (2013) advocate on statistical grounds the use of a *harmonic* rather than an *arithmetic* mean. But Table 7.1 also illustrates that the various indicator concepts have distinct evaluative, theoretical assumptions on what constitutes research performance.

All proposals seek an ‘optimal’ combination of publication and citation counts. Leydesdorff and Bornmann (2011) claim that a high *publication productivity* should be rewarded rather than disadvantaged. This is why citation curves must be integrated, in a way that corrects for differences in citation rates among subject fields. The Leiden group focuses on ‘*saturation*’, a phenomenon that occurs if a group increases its annual number of published papers while its citation impact per paper declines. Hirsch’s h-index tends to disregard papers in the bottom of the citation distribution, because in his view these do not reflect genuine performance. It is the *broadness* of a scientist’s *best* output in terms of citations that determines the value of his index. Of course, this assumption does not explain or justify why the cut-off citation frequency for defining broadness should be linked to the citation-based ranking of publications in the way expressed in the h-index. Finally, Abramo and d’Angelo adopt an econometric point of view: their quintessential indicator of research performance is a productivity or *efficiency* measure relating publication output or citation impact to research *input* in terms of cost of salaries of research staff. They essentially claim that the only good performance indicator is an efficiency indicator. By contrast, the comments by Glanzel are based on purely *statistical* considerations. His proposals are *neutral* in terms of evaluative assumptions.

Interestingly, both Leydesdorff and Bornmann, and Abramo and d’Angelo claim that their proposals mark a *paradigm switch* in citation analysis, abandoning the ‘old’ concept of a citation-per-article indicator, embodied both in the classical journal impact factor and in relative citation rates. Use of a citation-per-dollar type of measure recently advocated by Abramo and D’Angelo is not new. In the groundbreaking study on radio astronomy institutes by Martin and Irvine (1983) that played an important role in the development of scientometrics, at least in Europe, this type of indicator was calculated as well. Also, a percentile-based citation impact indicator defined as the number of publications among the top 10 per cent most frequently cited publications in a subject field, is already used by Francis Narin (1976) in the 1970s, although not in the integrated manner proposed by Leydesdorff and Bornmann (2011).

Table 7.1 Statistical claims, theoretical notions and application context of selected indicators

Article	Statistical claim	Basic notions of research performance	Preferred application context
Garfield, (1972) : Journal Impact Factor (JIF)	"In view of the relation between size and citation frequency, it would seem desirable to discount the effect of size when using citation data to assess a journal's importance. We have attempted to do this by calculating a relative impact factor—that is, by dividing the number of times a journal has been cited by the number of articles it has published during some specific period of time. The journal impact factor will thus reflect an average citation rate per published article (p. 477)"	"[In principle, bibliometric analyses provide a meaningful basis for discussion of the research performance of research groups with scientists in the field, and possibly also with members of the monitored research groups themselves. The analyses enable research policy-makers to ask relevant questions about research groups (p. 147)]"	"[...] this index may provide a useful yardstick to compare different individuals competing for the same resource when an important evaluation criterion is scientific achievement, in an unbiased way."
Leiden Group (Moed et al., 1985): Trend in short term impact	Citation distributions are skewed. This requires a very high level (almost 100%) of completeness of publication and citation data, and calculation of indicators at the level of research groups rather than for individuals	"[If the number of (short-term) citations is increasing, but the number of citations-per-publication is decreasing [...]], this indicates that the research group concerned is reaching some saturation level; such a group continues to publish more and more articles, but the impact of its performance as a whole does not increase proportionally (p. 136)"	"[...] for faculty at major research universities $h \sim 10$ to 12 might be a typical value for advancement to tenure (associate professor), and $h \sim 18$ for advancement to full professor"
Hirsch (2005): h-index	h-index is a stable and consistent estimator of scientific achievement	Performance must reflect both publication productivity and citation impact Publication counts alone "do not measure importance nor impact of papers"; total citations "may be inflated by a small number of big hits", which may not be representative of the individual if/he/she is coauthor"; citations per paper "rewards low publication productivity, penalizes high productivity"	(continued)

Table 7.1 (continued)

Leydesdorff and Bornmann (2011): Integrated Impact Indicator (I3)	<p>“The number of citations can be highly skewed and in this situation any measure of central tendency is theoretically meaningless” (p. 4)</p> <p>“[...] citation curves have to be integrated instead of averaged, [...], transforming them first into curves of hundred percentiles (p. 5)”</p>	<p>“The common assumption in citation impact analysis hitherto has been normalization to the mean. In our opinion, the results are then necessarily flawed because the citation distributions are often highly-skewed. Highly productive units can then be disadvantaged because they publish often in addition to higher-cited papers also a number of less-cited ones which depress their average performance.” (p. 34).</p> <p>“[...] differently from the h-index, the tails of the distributions are not discarded as irrelevant.” (p. 16)</p>	<p>The proposed measure can be applied to all common units of assessment (individuals, groups, institutions, countries, journals), and in various policy contexts, including in funding formula for principal investigators</p>
Glanzel (2010)	<p>In Scientometrics, most distributions are discrete and extremely skewed. “Nevertheless, the central limit theory guarantees the asymptotic normality of sample means even if the underlying distribution is discrete and skewed, provided the distribution belongs to the domain of attraction of the Gaussian distribution.” (p. 315). “Mean values and relative frequencies are unbiased estimators for the expectation and the corresponding probabilities (p. 316)”</p>	<p>“[...] The assumption of one single free parameter does not reflect reality of bibliometric practice. Based on the experience, at least two parameters are needed to model publication activity or citation impact.” Glanzel & Moed, 2013, p. 385, see also Glanzel, 2009</p>	
Abramo and D'Angelo (2014; 2016): Fractional Scientific Strength (FSS)		<p>“Productivity is the quintessential indicator of efficiency in any production system. (p. 1129)” Key aspect of research performance is productivity, expressed as a ratio of ‘output’ and ‘input’, specifically the ratio of citation impact and researchers’ salaries; in other words, (normalized) citation per Euro or Dollar spent</p>	<p>Ranking of individual researchers, departments and universities; national research assessment exercises such as the VQR assessment in Italy</p>

The application contexts differ substantially among the proposals, and reflect the changing role of citation-based indicators in research assessment, from monitoring tools to ‘hard’ components in research funding formula. Eugene Garfield needed ‘objective’ tools for expanding the source journal coverage of the Science Citation Index, and the work of the Leiden group in the early 1980s focused upon tools for identifying research groups that were losing connection with the international research front. But the proposals by Hirsch, Leydesdorff and Bornmann and Abramo and d’Angelo are linked to the assessment of individual researchers in matters of promotion and funding.

This case illustrates how in seemingly technical discussions on the construction and statistical properties of science indicators, *‘evaluative’, theoretical assumptions on what constitutes research performance* play an important, though often implicit, role. Below follow other examples illustrating how notions of what constitutes performance, in combination with particular policy objectives, influence the choice for indicators to be used in an assessment process.

7.2.2 Focus on the Top or the Bottom of a Performance Distribution?

In the first step of a standard citation analysis, a unit of assessment’s (UoA) publications are identified, and their citation counts are being collected within a citation database (e.g., Thomson Reuters’ Web of Science, Elsevier’s Scopus, or Google Scholar). As a rule, the distribution of citations among the UoA’s publications is *skewed*.

Analyzing this distribution, an important question is: *In which segment of the distribution is one interested: in the top, the bottom, or in the full distribution?* If one is interested in the *bottom*, because one wishes for instance to assess the broadness of the citation impact of an entity’s articles, the percentage of *uncited* publications is a useful measure. If one assumes that research *excellence* is best reflected in the extent to which a unit publishes highly cited (‘top’) publications, it is better to focus on the *top* of the distribution, and calculate the number or percentage of articles cited more than a specific threshold value. What constitutes genuine research quality? Different researchers may give different answers to this question. The Leiden study (Moed et al., 1985) also found differences between scholarly disciplines.

The example above relates to citation distributions of articles published by an individual researcher or a research department. But it can be further generalized as follows. If one aims to assess a large research institution by calculating a particular performance indicator for all research groups within the institution, the same question arises as above: Is one interested in the top or in the bottom of the quality distribution? Policy considerations determine which parameters of the citation distribution are relevant. This issue is closely related to the objectives of the assessment process, and the definition of the evaluation criteria: Is one interested to

identify groups that do not meet certain minimum performance criteria, or in identifying (potential) world-leading departments? This question is picked up again in Sect. 10.4 in Part IV.

7.2.3 Which Indicator Normalizations Should Be Implemented?

Indicators may be affected by distorting factors that have little to do with the concept they aim to measure. An approach often applied in informetrics to handle such biases is to construct new indicators by operationalizing specific distorting factors and including them as correction terms in the mathematical formula specifying the indicators.

A typical example of this approach is a so called relative citation rate that takes into account differences in citation practices among subject fields (see Sect. 17.1 in Part V). In biochemistry, citation levels are much higher than they are in mathematics. Hence, calculating simple citation rates, groups in the first field tend to rank higher than those active in the second. One way to overcome this bias is to normalize a unit's citation rate to the world citation average in the subject fields in which the unit is active. A second example is the construction of an indicator defined as the ratio of a scientist's h index and his/her academic age.

Including correctors for these 'other' factors in an indicator definition is not a theoretically neutral act, as the developer expresses that these factors *distort* the indicator, so that it does not properly measure what it is supposed to measure. Chap. 18 in Part VI shows how indicator normalizations influence the ranking positions of given institutions in world university rankings.

7.2.4 How to Define a Proper Reference Framework?

In informetrics, performance is normally viewed as a relative concept; there is no absolute scale available to measure it, but it is assessed by comparing the score of a unit of assessment within a particular reference framework. A distinction can be made between a *normative* and a *criterion-based reference framework* (see Table 6.1 in Chap. 6). In the first, each unit is compared against all others, while in the second all units are evaluated according to the same criterion.

When applying a normative framework, it is crucial on the basis of which grounds the 'benchmark' set of units is defined. For instance, in calculating rankings of world universities, some systems adopt a *global* perspective and compare a given institution with all other institutions in the world, whereas other systems first apply a *regional* normalization, by ranking universities by geographical region, and in a second step merge the regional rankings and create a global set in such a way that the top segment in the newly formed ranking includes institutions from all

geographical regions. Such a normalization decision has a large influence on *which* institutions appear in the top of the world ranking. This is clearly illustrated in Chap. 18 in Part VI. Regardless of the intention the informetric analysts, such normalization is *not* neutral from a policy perspective.

7.2.5 *Short Term Versus Long Term Perspective*

In any impact assessment a key question is: which time period does one take into account? How long does one wait and give a unit the time to generate a demonstrable impact? This depends first of all on the rapidity at which impact-generating processes take place, but also policy considerations play an important role. On the one hand, evaluating the significance of a scientific-scholarly piece of work is an intellectual activity that takes time. New ideas need to mature, to be considered from various points of view, and their implications understood. But how long should one wait with assessing the impact of a particular event?

On the other hand, an assessment of publications published a decade ago or more may be of little value for current research policy. In any practical study a balance must be found between scientific-methodological and political-managerial requirements. But the choice of a time window in an impact analysis has implications for how impact should be interpreted. In the literature on citation analysis a distinction is made between short-term impact (typically, measured during the first 3 years after publication date) and longer-term impact (assessed during a time period of at least 10 years). Short term impact relates to visibility at the research front, whereas longer term impact is related by some authors to ‘durability’, or, ‘sustainability’ of research results. (Moed et al., 1985; Leydesdorff et al., 2016), although there are also articles that generate a large citation impact only many years after publication, denoted by van Raan (2004b) as “sleeping beauties”.

With the development of altmetrics and usage measures this discussion gains even more relevance. For instance, there may be a delay of several months between the submission date of a manuscript and its online publication date. Once an article is published online, full text downloads may start almost instantaneously, while citations may show a delay of one year or more. Altmetrics relate to social media tend to have short response times in terms of hours if not minutes. Some experts denote altmetrics related to social media as indicators of *attention*. (e.g. Kurtz & Bollen, 2010). The implications of the use of short term impact indicators is further discussed in Sect. 10.2 in Part V.

7.3 The Influence of the Wider Context on Indicator Development

A base assumption of this section is that knowledge of the wider context in which specific indicators were developed contributes to a better understanding of how and under which conditions they can be properly applied. This section aims to illustrate, by means of a series of examples, how the interplay between the various participants may influence indicator development, and how indicators are developed in a particular socio-economic or cultural context and reflect scientific-scholarly, political and economic factors. These examples aim to contribute to a better understanding of current and future indicator development. In terms of the distinction between the *context of discovery* and the *context of justification*, the current section deals with the first.

Those who propose new performance indicators are researchers themselves. They are part of a research community, and are subjected to assessment as well. Their experiences in research assessment practices (both as assessors and as objects of assessment) are a part of their ‘intellectual base’, and may influence the underlying notions and objectives of their newly proposed indicators. This section does *not* claim that the relationship between intellectual base and indicator concept is causal, as if a particular experience of circumstance ‘causes’ a specific indicator functionality, nor that the context of discovery may provide a justification of the validity of an empirical knowledge claim.

An indicator concept can be assumed to also being the result of a reflection upon the assessment experiences of the developers and the wider research and policy communities to which they belong. It is essential that such a reflection takes place and influences new concepts. This illustrates that indicator development, although mostly of a quantitative nature, is more a social science or humanities-like activity than a natural science activity. Unlike the situation in, for instance, physics, the developer does not play the role of an inter-changeable observer, but as a fully engaged person, aware of the fact that the community in which he takes part, including himself, will be subjected to the new indicators as well.

Following the above lines of argument, this section presents a first step towards a brief history of indicator development during the past 50 years. The overview is far from comprehensive, and focuses on indicators with which the current author is most familiar. It does not discuss technical details of the various types of indicators. For a thorough overview of these aspects the reader is referred to Waltman (2015). It focuses on three ‘normalizations’, namely size, subject field, and age normalization. The relatively large attention given to the Centre for Science and Technology Studies (CWTS) at Leiden University reflects the circumstance that the current author has been a CWTS staff member for almost 30 years and knows much more about this group than any other institute, and does *not* emerge from a notion of superiority of CWTS or depreciation of the excellent and important work of many of informetric colleagues.

The history starts with the journal impact factor (JIF), a citation-per-article indicator, calculated at a level of a scientific journal, with the purpose of assessing a journal's information utility, correcting for the size of its annual volume, and used by Eugene Garfield as a tool to monitor the coverage of his Science Citation Index (Garfield, 1972). It is noteworthy that it is implicitly assumed that 'difference in size' is a factor that an information utility indicator should correct for. At this point the question arises whether a journal publishing 100 papers cited in total 200 times is from an informational point of view less useful than a journal publishing 20 papers cited 50 times. Hence, it is perhaps more appropriate to speak of the JIF as an indicator of the information utility *per published article*.

This issue of *size normalization* is one of the central issues in bibliometric indicator development, and will appear several times below. A second notion that emerged immediately after analyzing lists of JIF values of journals was that there appeared to be large differences across subject fields. A common notion is that *subject field normalization* should correct for such differences. A third key issue, of a *statistical* nature, is how to characterise the *distribution* of citations among target articles published by a particular unit of assessment. Main approaches focus on: (a) the *mean* of the distribution (arithmetic or harmonic mean); (b) the *top percentiles* of the citation distribution, e.g., calculating the number of published articles among the most highly cited in a subject field; (c) the *h-index* (Hirsch, 2005). It focuses on the top of the citation distribution as well.

Several authors started applying the citation-based indicators also at other levels of aggregation, especially that of countries and research institutions. In the USA, in studies mainly commissioned by US institutions, *Francis Narin* and colleagues at Computer Horizons Inc. (later renamed to *CHI Research*) used as a key indicator the number of published articles in top citation percentiles, calculated by subject field, for US institutions and their international competitors (Narin, 1976). These studies identified in most cases the USA as top country, followed by the UK. The use of an indicator based on the number of articles published in the top citation percentiles in a specific subject field underwent a revival in the past 15 years (Tijssen, Visser, & van Leeuwen; 2002; Leydesdorff, Bornmann, Mutz, & Ophof, 2011; Bornmann, de Moya-Anegon & Leydesdorff, 2011) when major producers as Scimago, CWTS and Evidence Ltd. included particular variants in their bibliometric profiles.

The pioneering work of *ISSRU* at the Library of the Hungarian Academy of Sciences calculated for all countries in the world mean citation rates of articles published by country (Schubert, Glanzel & Braun, 1989). It became clear that a ranking of countries based on this type of indicator was strongly biased in favour of a selected group of Western, developed countries, in which the USA and UK ranked on top. Especially Central or East European countries tended to occupy much lower positions in these rankings. Hence, users from a large group of other countries would perceive the results as being too negative, and of little policy relevance. At the same time, the experts at ISSRU were very well aware that subject field normalization should be applied as well. By calculating a 'relative' indicator dividing a country's citations-per-article ratio by the(weighted) average impact of

the *journals* in which the country's articles were published, an indicator was constructed for which the bias towards the selected group of Western countries was substantially reduced.

In its early work, CWTS also calculated in its assessment methodology a citation rate relative to the average journal impact, similar to the one proposed by ISSRU. But in a validation round this measure was strongly criticized by researchers from the University of Leiden (Moed et al., 1985). Critics claimed that in this indicator a substantial portion of genuine impact was 'normalized away'. It led to the problematic, and for many researchers unacceptable, outcome that a group publishing in the very best international journals and having a citation rate equal to the high journal average would have the same score as group publishing in mediocre or more nationally oriented journals and cited on average as often as the papers in these second-tier periodicals. While this indicator could still be most useful in a macro study when comparing large sets of countries of the type conducted by ISSRU, in the context of an assessment of research departments within a Dutch university it was considered inappropriate.

The CWTS approach was based on the notion that the base unit of activity in science—and often the 'business unit'—is the *research group*, consisting of a group leader, one or more senior staff members, post docs and a group of Ph.D students. Although researchers from social sciences and humanities claimed that this model is not valid in their subject fields, there was in the research community a general agreement that the bibliometric method should focus on groups, and *not* on *individuals*, based on the following reasoning: academic researchers are not rich; their only possession is their reputation. Unjustly harming their reputation may easily create un-repairable damage. Bibliometric indicators at the level of individuals could be harmful. Hence, it was considered appropriate to focus on groups rather than individuals.

But in the policy domain, practitioners claimed that in research groups specific individuals make a difference, and that information on individual performance is needed even in the assessment of groups and departments. They also argued that, if a group leader is a co-author of all articles published from a group—a phenomenon that occurred often—, individual and group performance coincide. In this way, a tension was being created between what policy officials and managers needed on the one hand, and what responsible bibliometric developers felt they could (or could not) provide, on the other. This tension still exists today.

The calculation of 'size normalized' indicators based on the citation-per-article concept was essential in the CWTS approach, as it aimed to accommodate a particular policy need, namely the development of tools for identifying scientifically *emerging* and *declining* research groups. Especially in the Dutch context, 30–40 years after the end of the Second World War, when many senior academic staff was about to retire, and the growth rate in academic funding was about to decline, this was viewed essential. Hence, a key indicator should, for instance, be able to assign to a small emerging group with a relatively short track record and a still limited level of external, competitive funding, but already rapidly increasing its international visibility, a higher score than a large department that had consolidated its status for many years,

following research lines in which international colleagues in the field were losing interest.

CWTS proposed a new indicator, sometimes denoted as ‘crown’ indicator, in which the citation rate of a group’s articles was divided by the word citation average in the subject fields (defined as journal categories) in which the group was active (Moed, De Bruin & van Leeuwen, 1995). This indicator ranked The Netherlands in the top of a global ranking, jointly with Scandinavian countries, and above USA, UK, Germany and France. American top universities obtained a score of about 3.0, which was denoted as ‘Harvard level’.

The crown indicator has never become popular in the USA and the UK. While Francis Narin and co-workers at CHI Research considered the crown indicator as an improvement compared to the indicator normalized by a journal citation average, they continued using in their own work the absolute number of ‘top’ publications (in the top citation percentiles) as the key indicator. In the UK there was a decline in interest for bibliometric indicators during the 1990s, followed by a revival in 2004, when an article published in Nature (King, 2004), based on a bibliometric analysis by Evidence Ltd., using the absolute number of citations and highly cited publications as key indicators, positioned UK at the *second* position in a global ranking, directly after USA.

In 2005 a creative physicist and a relative outsider to the field of scientometrics proposed the so called *h-index*, a bibliometric indicator measuring the research performance of an individual researcher (Hirsch, 2005). Shortly after its publication, an editorial published in the journal Nature brought it to the attention of its readers and asked them to provide feedback, also in view of an upcoming research assessment exercise in the UK (Nature Anonymous, 2005). The physical dimension of this index is ‘number of published articles’. The h-index counts the number of articles published by an individual that meet a particular statistical criterion. For instance, a h-index value of 10 means that there are 10 published articles with at least 10 citations. It follows immediately that an individual’s h index does never exceed the number of articles he or she has published.

Hirsch positioned his indicator in a particular application context, namely the promotion of academic researchers to various positions in the US academic system, including tenured professorships, and memberships of scientific societies and academies. He proposed a series of h-index threshold values above which researchers could be considered serious candidates for particular positions. Although the indicator was criticized—for instance, a report by the International Mathematical Union claimed that it lacks common sense (Adler, Ewing & Taylor, 2008; see also van Raan, 2006)—, the three large multi-disciplinary citation databases, Thomson Reuters’ Web of Science, Elsevier’s Scopus and Google Scholar, enable users to collect h-index values of individual authors. The last two databases even calculate h-indices themselves.

After the launch of the h-index, a series of new indicators was proposed, many obviously inspired by the h-index, but based on slightly different statistical assumptions, seeking to take into account various distorting factors. In this way, during the past ten years, indicator development had an internal dynamics, strongly

driven by mathematical-statistical considerations, in which validation, user acceptance tests and fine-tuning in a concrete application context did not play an important role. As Wildgaard observed, “many author-level indicators are created as statistical exploration of distributions rather than advancing knowledge about the performance of researchers” (Wildgaard 2015, p. 115).

It is perhaps no surprise that during the past 10 years large groups of especially young researchers lost interest in the ‘classical’ bibliometrics based on publication- and citation-based indicators derived from Web of Science, Scopus, and, to some extent, also from Google Scholar. The h-index is obviously of little use in demonstrating the emerging status of young, upcoming researchers. Also, in order to gain visibility via indicators derived from Web of Science or Scopus, one needs to have published already a number of publications in established international journals, indexed in these citation databases, while young researchers often start their publishing career with contributions in other publication outlets, such as conference proceedings or research reports.

On the other hand, Google Scholar does cover a large number of sources (journals, books, conference proceedings, disciplinary preprint archives or institutional repositories) that are not indexed in WoS or Scopus, and thus has a much wider coverage (e.g., Moed, Bar-Ilan & Halevi, 2016); its surplus is especially relevant for young researchers. Last but not least, the notion emerged that there are other traces of research impact which could manifest themselves with a much shorter time delay than those reflected in the classical metrics, and in other types of sources, especially in social media and scholarly reference managers. These traces were labelled as altmetrics (Plem, Taraborelli, Groth & Neylon, 2010). They are further discussed in Chap. 11 in Part IV.

7.4 Indicator Development and Business Interests¹

Development and delivery of informetric indicators does not only take place in an academic environment, but also in private companies or in hybrid, partly publicly and partly privately funded institutions. Information companies may have multiple types of relationships with informetric indicators. Firstly, they may produce and sell *indicators*, based on the elaboration of databases created and owned by themselves. Typical examples are the indicator product Incites based on the bibliographic database Web of Science, both owned by Thomson Reuters (currently Clarivate Analytics); SciVal indicators derived from Scopus, both owned by Elsevier; and Google Scholar Metrics, based on the bibliographic database Google Scholar. Though separate products, the indicator databases provide visibility to the underlying databases and vice versa.

¹This section re-uses selected paragraphs from Moed (2016e).

Secondly, *as publishers or web service providers*, information companies may also deliver a part of the *content* of the databases from which the indicator products are derived. Thomson Reuters, Elsevier and Google not only offer large bibliographic databases and indicator products derived from them, they are also publishers of a part of the sources processed for these databases. Though Thomson Reuters is a minor publisher of books and journals, Elsevier is considered the largest scientific publisher in the world. Around 15% of the sources processed for Scopus are published by Elsevier itself. Google produces *Google Books*, a web service providing extracts from the contents of millions of books scanned by Google, which constitute a substantial part of the source coverage of Google Scholar.

A third type of relationship between information industry and indicators occurs when companies provide in their indicator products metrics related to sources that they themselves create and sell. A typical example is that Elsevier offers metrics for all journals covered in Scopus, including those published by Elsevier itself. It must be noted that the two Scopus-based journal indicators SJR and SNIP available since 2010 were invented and calculated by two independent academic groups, using in-house versions of Scopus (González-Pereira, Guerrero-Bote & Moya-Anegón, 2010; Moed, 2010).

The idea of an information company producing both sources and indicators to offer indicators developed by independent research groups, is appropriate in view of a possible conflict of interest, but has its limits as well. It is conceivable that, as academic groups strive at technically realizing their indicator concepts, such a company may be in the position to choose between competing concepts, and use the indicator ratings of their own sources as an important, if not decisive criterion.

During the past decade a series of social networking sites have been launched, some for the general public, but often used by researchers—e.g., Facebook, Twitter and Google+ –, and some especially for researchers, including ResearchGate, Academia.edu and Mendeley. (see for instance Rasmussen Neal ed., 2012). Many of these sites generate informetric indicators on activity and visibility. Typical examples are ResearchGate and Mendeley. Also, intermediary companies such as Altmetric.com, owned by Digital Science, a technology company operated by Holtzbrinck Publishing Group, and such as ImpactStory and Plum Analytics (the latter currently owned by Elsevier), collect scientific-communication related data from general and scholarly networking files, and calculate a composite metric of online visibility in social media.

ResearchGate calculates an indicator for its subscribers, denoted as RG Score, based on a proprietary algorithm. Several researchers have attempted to shed light on RG score by correlating it with other, better known indicators (e.g., Thelwall & Kousha, 2014). It is hypothesised that impact factors of journals play an important role, and it is plausible to assume that the degree of activity and visibility within ResearchGate itself is an important factor as well. But there is no solid evidence for this. Mendeley provides users with information on the number of ‘readers’ in Mendeley, i.e., the number of Mendeley users who have added a particular article into their personal library.

Since Eugene Garfield introduced the JIF as an ‘objective’ tool to expand the journal coverage of his citation index independently of journal publishers, the landscape of scientific information providers and users has changed significantly. While, on the one hand, politicians and research managers at various institutional levels need valid and reliable fit-for-purpose metrics in the assessment of publicly funded research, there is, on the other hand, a tendency that indicators increasingly become a tool in the business strategy of companies with product portfolios that may include underlying databases, social networking sites, or even metrics products. This may be true both for ‘classical’ bibliometric indicators and for alternative metrics.

Chapter 8

The Policy Context

Abstract This chapter discusses the policy context of the use of informetric indicators in research assessment. It introduces the notion of a multi-dimensional research assessment matrix, representing the multi-dimensionality of research performance, and the multi-faceted domain of science policy and research management. Next, it illustrates the influence of the policy context upon the choice of indicators. Finally, it introduces the notion of meta-studies that inform policy makers and evaluators about how to design an assessment process.

Keywords Attraction capacity · Hiring · Journal manuscript acceptance rates · Meta-analyses · Minimum performance standards · Promotion · Quality of academic staff · Research assesment matrix · Research income · Systemic characteristics

8.1 Introduction

In 2010 the Expert Group on the Assessment of University-Based Research, installed by the European Commission, published a report introducing the concept of a multi-dimensional research assessment matrix, built upon the notion of multi-dimensionality of research as outlined in Chap. 3 in Part II of this book. Rather than focusing on one single output dimension and underlining the need to obtain convergence among a set of different indicators in order to produce valid and useful outcomes, the report aimed at proposing “a consolidated multidimensional methodological approach addressing the various user needs, interests and purposes, and identifying data and indicator requirements” (AUBR, 2010, p. 10).

It is based on the notion that “indicators designed to meet a particular objective or inform one target group may not be adequate for other purposes or target groups”. Diverse institutional missions, and different policy environments and objectives require different assessment processes and indicators. In addition, the

This chapter re-uses with permission selected paragraphs from Moed & Halevi (2015).

range of people and organizations requiring information about university based research is growing. Each group has specific but also overlapping requirements (AUBR, 2010, p. 51).

The aim of the current chapter is to further develop the notion of the multi-dimensional research assessment matrix. It focuses on the purpose, objectives and policy context of research assessments, and demonstrates how these characteristics determine the methodology and metrics to be applied. For instance, publication counts are useful instruments for discriminating between those staff members who are research active, and those who are not, but are of little value if research active scientists with publication output exceeding a certain minimum threshold are to be compared one with another according to their research performance.

In addition, it introduces the concept of a ‘meta-analysis’ of the units under assessment in which metrics are not used as tools to evaluate individual units, but to reach policy decisions regarding the overall objective and general set-up of an assessment process. In line with the distinction made in Chap. 6 between the evaluative and the analytical domain in an assessment process, and with the implications this has for the professional behavior of informetricians, the current author maintains a neutral position towards political objectives and strategies, and examines whether particular approaches or indicators are *defensible* from an informetric viewpoint.

8.2 The Multi-dimensional Research Assessment Matrix

When designing a research assessment process, one has to decide which methodology should be used, which indicators to calculate, and which data to collect. To make proper decisions about these matters, one should address the following key questions, each of which relates to a particular aspect of the research assessment process.

- What is the *unit* of the assessment? A country, an institution, a research group, an individual, or a research field or an international network? In which discipline(s) is it active?
- Which *dimension* of the research process must be assessed? Scientific-scholarly impact? Social benefit? Multi-disciplinarity? Participation in international networks?
- What are the *purpose* and the *objectives* of the assessment? Allocate funding? Improve performance? Increase regional engagement?
- What are relevant ‘*systemic*’ characteristics of the *units of assessment*? For instance, to which extent are researchers in a country oriented towards the international research front?

The answers to these questions determine which indicators are the most appropriate in a particular assessment process. Indicators that are useful in one context, may be less so in another. The aim of this section is to further develop this principle by taking into account new bibliometric and non-bibliometric indicators, a series of aggregation levels, impact sub-dimensions, and by focusing on the objectives and policy background of the assessment.

8.2.1 Units of Assessment

Two characteristics of the unit under assessment must be underlined, as they determine the type of measures to be used in the assessment. Firstly, the discipline(s) in which the unit under evaluation must be taken into consideration. There are several disciplines that are difficult to assess mainly because they are geographically or culturally specific. Among these one can identify linguistics, language-specific literature, law, and others, especially in the humanities (e.g., Nederhof, Luwel & Moed, 2001). Secondly, the mission of the research unit under assessment is relevant as well. To the extent that it is taken into account in the assessment process, it determines the indicators that have to be applied.

8.2.2 Objectives and Performance Dimensions

A distinction can be made between purpose and objective of an assessment. A purpose has a more general nature, and tends to be grounded in general notions (e.g., “increase research performance”), whereas objectives are more specific, more formulated in operational terms (e.g., “stimulate international publishing”). Objectives are grounded in assumptions on how they are related to the general purpose (e.g., “it is assumed that by stimulating international publishing, research performance increases, at least in the longer run”).

Table 8.1 presents a list of three main assessment objectives highlighted in the AUBR Report. For each objective it gives the most useful indicators from the set of 28 principal indicators or indicator families included in Table 3.1 in Sect. 3.3.

8.3 Systemic Characteristics of the Units of Assessment

Meta-assumptions on the state or condition of the system of units of assessment play an important role in the formulation of the objectives of an assessment. Unlike an assessment of these units on a one-by-one basis, such assumptions do not focus on *individual* units, but relate to more general or systemic characteristics of these

Table 8.1 The most important indicators informing three main assessment objectives

Policy issue	Performance aspect	Key indicators (families)
Quality, sustainability, relevance and impact of research activity	Quality of academic staff and Ph.D. students	Citation impact indicators; readership measures
	Societal impact	Mentions in social media
	Economic and technological impact	Number of licenses, spin-offs, patents
	Sustainability	Ratio research students/staff; Percentage early stage researchers involved in research
	Research funding	External research income; Number and percentage competitive grants won
Investor confidence/value-for-money and efficiency	Quality of academic staff and Ph.D. students	Citation impact indicators
	Attraction capacity	Recruitment of academic staff and students from abroad
	Efficiency level	Economic efficiency indicators (output per input)
	Benchmarking nationally and world wide	Position in world university rankings
	Sustainability and scale	Total investment in research; research students/staff; Value of research infrastructure
Research strategy development/management	Attraction capacity	Recruitment of academic staff and students from abroad
	Benchmarking against peer institutions, nationally and worldwide	Position in university rankings
	Research intensity, quality of staff per subject field	Citation measures per department; ratio research students/staff

units *as a group*. Therefore, they can be denoted as ‘meta’ assumptions, and illuminate the assessment’s *policy context*. Typical *examples* are as follows.

- A substantial part of professors in this country is not research-active. For instance, they may be too much engaged in teaching.
- Through self-selection applicants for a vacant research position are research active; their quality level tends to be high.
- Researchers in this country are not well integrated into the international community and publish mainly in national journals.
- Young research groups have no good chances to develop in this funding system.
- Decisions on hiring or granting of proposals tend to be made on the basis of political considerations.

For instance, “stimulating international publishing” as an objective in a national research assessment exercise makes sense from a policy viewpoint only if there are

good reasons to believe that the level of international publishing among a country's researchers is relatively low compared to their international counterparts. Similarly, assessing whether an academic staff member is 'research active' or not, makes sense only if there is evidence that a non-negligible part of staff hardly carries out research.

The claim that the selection of indicators is influenced by the policy objectives and meta assumptions on the state of the system of units of assessment is illustrated by means of two examples that relate to the assessment of individuals: One relates to the use of journal metrics, and a second to the application of publication counts. The policy relevance of these examples is that managers at the departmental and central level in academic institutions are confronted with the necessity to evaluate researchers for promotion or hiring on a daily basis. *Section 8.5* below further discusses the notion of meta-analyses.

8.3.1 *The Use of Journal Impact Factors for Measuring International Orientation*

A *first* example to clarify the influence of the policy context upon the choice of indicators in a research assessment process relates to international publishing. This term may refer to the level of the *quality criteria* applied by editors and referees in the review of submitted manuscripts, or to the *geographical location* of authors, members of the editorial or referee board, and/or readers of a journal. The following definition would include both dimensions: international publishing is publishing in outlets that have: (1) rigorous, high-standard manuscript peer review; and (2) international publishing and reading audiences.

If an analysis of the state of a country's science provides evidence that a substantial group of researchers tends to publish predominantly in national journals that are hardly read outside the country's borders and do not have severe rigorous peer review, it is informetrically *defensible* to use the number of publications in the top quartile of journals according to citation impact as an indicator of research performance. In this manner one is able to discriminate between those researchers whose research quality is sufficiently high to publish in international, peer reviewed journals, and those who are less capable of doing so.

But if in internationally oriented, leading universities one has to assess candidates submitting their job application, it is questionable whether it makes sense comparing them according to the average citation impact of the journals in which they published their papers, using journal impact factors or other journal indicators. Due to *self-selection*, the applicants will probably publish at least a large part of the papers in good, international journals. Other characteristics of the published articles, especially their actual citation impact, are probably more informative as to the candidates' past research performance and future potential than indicators based on journal metrics are.

What is the *empirical, informetric evidence* that makes the use of the journal impact factor along the lines sketched in this example *defensible*? Bibliometric

studies found that the journal impact factor is a proxy of a journal's international status. Sugimoto et al. (2013) examined the relationship between journal manuscript acceptance rates and 5 year journal impact factors, and found in a sample of 1325 journals a statistically significant, negative, linear correlation coefficient between these two measures, suggesting that journals with rigorous referee systems—and hence, with a high international status—tend to generate higher impact than others.

But Sect. 5.4 warned that if in a particular study a positive (linear or rank) correlation is found to hold between two variables, it does not follow that it holds for all sub-ranges of values of these variables. Whether or not a sample of the two variables can be expected to correlate in a particular study, very much depends upon the range of values obtained by the units in the sample. In fact, the study by Sugimoto et al. (2013) also found that, when dividing journals into *quartiles* according to their acceptance rates and analyzing correlation coefficients *within quartiles*, the correlation coefficients between acceptance rates and impact factors were much closer to zero and *not significant*.

This case shows that the application of journal metrics or publication counts to assess the comparative performance of researchers who publish on a regular basis in international journals cannot be sufficiently justified by referring merely to earlier studies reporting on observed positive correlation between these measures and peer ratings of research performance. For *this* group of researchers, the two measures do *not* correlate significantly at all.

The current author agrees with the critique made by The San Francisco Declaration on Research Assessment (DORA, 2009) and many others that the role of journal metrics in the assessment of individual researchers has become far too dominant. There is no solid empirical, informetric support for an assessment practice aiming to discriminate in a group of research active researchers publishing in good journals between high and low performers on the basis of weighted impact factors of the journals in which they published their articles.

On the other hand, it does *not* follow that the use of this type of indicator is invalid under all circumstances. If in a country international publishing is not a common practice, and if the policy sphere decides it should be stimulated, use of journal impact factors in the way suggested above is informetrically *defensible*, not so much as an assessment *tool*, but rather as an instrument to *change* researchers' *communication practices*. As outlined in Sect. 9.1, a full assessment of individual research performance requires much more information than journal impact factors, and more than informetric indicators alone.

8.3.2 *The Use of Publication Counts in the Assessment of Being Research Active*

A *second* example relates to the use of *publication counts*. In order to identify academic staff that is not research active, it is informetrically defensible to consider the publication output of the staff under assessment, and identify those whose

output is below a certain—subject field dependent—minimum. But if one has to assess candidates submitting their job application to a leading research university, it hardly makes sense to compare them according to their publication counts. Due to self-selection, they will probably all meet a minimum threshold. In other words, while there are good reasons to believe that journal metrics or publication counts are appropriate indicators to identify the bottom of the quality distribution of research staff, they have a limited value if one aims to discriminate in the top of that distribution.

Informetric evidence for the proposed type of use of publication counts stems from several studies, including one mentioned in Sect. 4.2 by a peer review committee assessing the performance of academic research groups in the Netherlands in the field of biochemistry, that concluded that while publication counts are useful to identify research units of which the output is below a *minimum level* of scientific production, they are *unapt* to discriminate in terms of performance between groups with counts *above* this threshold (Survey Committee Biochemistry, 1982). The idea to use informetric indicators to define minimum performance standards is further discussed in Sect. 10.4.

8.4 Meta-Analyses

It was stated in the previous section that a meta-analysis of systemic characteristics of the units of assessment' has an influence on the selection of the methodology and indicators to be applied in an assessment process. It must be noted that informetric indicators may—and actually do—play an important role in the empirical foundation of such a meta view. Indicators are essential tools on two levels: in the assessment process itself, and on the meta level aimed to design that process. Yet, their function in these two levels is different. At the first level they are tools in the assessment of a particular unit, e.g., a particular individual researcher, or department, and may provide relevant empirical evidence about their performance. At a second level, they provide insight into the research system as a whole, and help draw general conclusions about its state and about the objectives and general set-up of an assessment process.

A meta level analysis can also provide a clue as to how peer review and quantitative approaches might be combined. For instance, the complexity of finding appropriate peers to assess all research groups in a broad science discipline in a national research assessment exercise, may urge the organizers of that exercise to carry out a bibliometric study first, and decide on the basis of its outcomes in which specialized fields or for which groups a thorough peer assessment is likely to be necessary. One important element of the meta-analysis is a systematic investigation of the effects of the assessment process, both the intended and the unintended ones.

8.5 Policy Considerations

Research assessment methodologies cannot be introduced in practice at any point in time, and do not have eternal lives. In the previous section it was argued that under certain conditions it is defensible to use publication counts and journal metrics as one of the sources of information in individual assessments. But one may argue that it is fair to maintain a time delay of several years between the moment it is decided to use a particular assessment method or indicator on the one hand, and the time at which it is actually used, on the other. In this way, the researchers under assessment have the opportunity to change their publication behavior—to the extent that they are capable of doing that. It must be noted that fairness is *not* an informetric category, but relates to what is *good governance*.

In recent years there have been several discussions that challenge the common practice of research evaluation using, for example, journal impact factors (Alberts, 2013; van Noorden, 2013). The San Francisco Declaration on Research Assessment (DORA, 2009) is one of these manifestations, calling for improvements that need to be made to ways in which research is evaluated and especially challenging the impact factor as a tool in such evaluations. One should consider changing an assessment method radically every 5–10 years. Two considerations in which informetric arguments play an important role may lead to such a decision. First, a meta-analysis may reveal that the overall state of the units of assessment has changed in such a manner, that the old methodology is either irrelevant, or even invalid due to strategic behavior. Secondly, any use of assessment methodologies and indicators must be thoroughly monitored in terms of its effects, especially the unintended ones. Severe negative effects such as indicator manipulation may lead to the decision to abandon a method, and establish a new one, even though one can to some extent detect and correct for such behavior in adapted indicators. This issue is further discussed in *Chapter 9*.

Regarding the—either negative or positive—effects of the use of metrics or any other methodology in research assessment, one may distinguish two points of view. One may focus on its consequences for an *individual entity*, such as an individual scholar, a research group or institution, or on the effects it has upon scholarly activity and progress in general. A methodology, even if it provides invalid outcomes in individual cases, may be beneficial to the scholarly system as a whole. Cole and Cole expressed this notion several decades ago in their study of chance and consensus in peer review of proposals submitted to the National Science Foundation (Cole, Cole & Simon, 1981).

It must be noted that this statement is also subjected to criticism. One should take into account *the effects* of cases in which a method generates invalid outcomes. One cannot *a priori* assume that these effects are small merely because the sheer number of such cases is small. Single events or small disturbances in a system can have enormous consequences for a system as a whole. This is especially true in so called non-linear systems. To give an example, one may try to imagine what would have happened in the science system if key papers by Albert Einstein would have been systematically rejected for publication in leading physics journals.

Each methodology has its strengths and limitations, and is associated with a certain risk of arriving at invalid outcomes. As Martin (1996) rightly pointed out, this is true not only for metrics but also for peer review. It is the task of members from the scholarly community and the domain of research policy, and not of informetricians to decide what are acceptable “error rates” and whether its benefits prevail, based on a notion of what is a fair assessment process. Informetricians and other analysts of science and technology should *qualitate qua* provide insight into the uses and limits of the various types of metrics, in order to help scholars and policy makers to carry out such a delicate task.

Part IV

The Way Forward

Chapter 9

Major Problems in the Use of Informetric Indicators

Abstract This chapter discusses a series of main problems in the use of informetric indicators for evaluative purposes: the problem of assessing individual scholars; the effect of a limited time horizon; the problem of assessing societal impact; the effects of the use of indicators upon authors and editors; constitutive effects of indicators; and the need for an evaluative framework and an assessment model.

Keywords ACUMEN portfolio · Editorial self-citations · Evaluative bibliometrics · Indicator manipulation · Individual research performance · Professional contacts · Publication repositories · Research assessment exercise · Research excellence framework · Science-technology helix · Social merit

9.1 The Problem of Assessing Individual Scholars

Section 1.1 in Part I states that performance indicators may not only be influenced by the various aspects of performance, but also by factors that have little to do with performance. They are partial or imperfect, and this makes their use a permanent quest for bias. This section focuses on the problem of assessing individual researchers and on the interpretation of author level metrics. It underlies the need to be cautious when interpreting informetric indicators calculated at the level of an individual researcher.

To what extent can informetric indicators assess the research performance of an individual scientist? Performance of an individual and citation impact of the papers he or she (co-)authored relate to two distinct levels of aggregation. Many bibliometric studies have shown that multiple co-authorship is a rule rather than an exception, especially in the natural and life sciences. As a consequence, publications (co-)authored by an individual researcher are often, if not always, the result of research to which other scientists have contributed as well, sometimes even dozens of them. The crucial issue is how one should relate the citation impact of a *team's* papers to the performance of an *individual* working in that team.

Bibliometrists have experimented with technical-statistical solutions, by assigning weights to co-authors that measure the contribution of each author, on the basis of his or her position in the author list. A first solution, denoted as integer or full counting, assigns a multi-authored paper fully to each contributing author. Uniform fractional counting assigns to each author of a publication the same fraction, such that the weights sum up to one, while first author counts assign a paper fully to the first author only. But there are also more sophisticated counting schemes, based on assumptions about authoring conventions in particular scientific subfields. One typical example in case of a paper with 5 authors is assigning a weight of 0.25 to the first and to the last author, and distributing the remaining half equally among the other co-authors. A special solution is the allocation of a paper fully to the (team of the) reprint author, under the assumption that the latter represents the research guarantor (Moya-Anagon et al., 2013).

Despite these technical solutions, the author of this book defends the position that a valid assessment of individual research performance can be done properly only on the basis of sufficient background knowledge on the particular role they played in the research presented in their publications, for instance, on whether this role has been leading, instrumental, or technical. In addition, other manifestations of research performance should be taken into account as well. Calculating indicators at the level of an individual and claiming they measure *by themselves* an individual's performance suggests an accuracy of measurement that cannot be justified. In other words, it suggests a façade of exactness.¹ This is why it is an excellent idea to ask individual scholars under assessment to provide 'narratives', and enable them to contribute various types of quantitative and qualitative information about their performance to the assessment process, as proposed by the team creating the ACUMEN portfolio (Bar-Ilan, 2014) or by Benedictus and Miedema (2016).

The recent feature in several scientific journals to invite authors to specify their contribution to a manuscript is helpful and provides relevant information, but it currently is not available for all journals. In addition, it is based on self-reported data, which has its limitations. This information is useful, but cannot replace the background knowledge indicated above. In Sect. 12.4 it is argued that an online bibliometric tool for self-assessment based on sufficiently accurate data and flexible, adequate benchmarking, is still lacking, but it is urgently needed and, most importantly, technically feasible.

9.2 The Effect of a Limited Time Horizon

Section 1.1 underlines that most informetric assessments adopt a limited time horizon. The choice of a time horizon depends on the speed of the communication process, for instance, on the publication delay in a field, defined as the average time

¹David Pendlebury informed the current author that in his communication with Robert K. Merton the latter used the term 'false precision' (Pendlebury, 2017).

interval between the date a manuscript is submitted for publication, and the date it is published. But policy considerations play an important role as well. An assessment of publications published more than a decade ago may be of little immediate value for current research policy, except in special historical studies.

In any practical study a balance must be found between scientific-methodological and political-managerial requirements. As a result, in most studies using publication and citation-based indicators the time period considered is at most ten years. The journal impact factor, still a key indicator in informetrics, is based on a citation time window of at most three years. In the analysis of download counts and altmetrics, especially the appearance in social media, the time horizon may be even shorter than that.

The current author agrees with the proposition that usage-based indicators and altmetrics primarily reflect *attention* rather than *influence* (Kurtz & Bollen, 2010). To assess the value of a scientific result one needs a certain time to reflect upon its significance. It takes time for a new idea to mature and establish itself among colleagues. Twitter counts and full text downloads in the first few months after its publication tend to indicate the amount of attention a piece of research collects, and represent a document's face value. Since the number of cited references in a research article is limited, forcing authors to make a selection of what they cite, citations tend to be more strongly based upon a reflection on the relevance of articles for the prepared text than article downloads or reads are. Thus, the length of the time window considered influences the interpretation of the figures.

But the view on performance offered by citations is also limited. The claim that new ideas need time to mature is also made in relation to citation-based indicators, especially those measuring 'short term' impact by counting citations during the first few years of a cited publication's life time. There is an increasing interest in studying what Anthony van Raan has termed as 'sleeping beauties', i.e., publications that during the first years of their existence are hardly cited or not cited at all, but that from a certain moment onwards become highly cited (van Raan, 2004b). Also, contributions that generated a high citation impact in the short term not necessarily have impact in the longer term as well. An early key paper by the Leiden group stated:

We assume that there is a research front in every scientific field. At this front, scientists develop theories about the structure of reality and these theories are confronted with each other through experimental performance research. In the end certain theories will triumph—temporarily or otherwise—and be added to the basic knowledge in the field. The short-term impact indicates how groups maintain themselves at the research front, the long-term impact indicates to what extent they eventually succeed in scoring 'triumphs'" (Moed et al, 1985, p. 133/134).

Recently, Leydesdorff et al. (2016) have further developed these ideas. But also a citation analysis during a longer time period has its limits. One of the reasons is the phenomenon of obliteration by incorporation, a term coined by Robert K. Merton, and studied by Katherine McCain (McCain, 2012) and other colleagues. Ideas may have become so influential and accepted, that their contributors are no longer cited in articles using these ideas—even though they may be mentioned in the full text.

The current author defends the position that the notion of research performance, and of making a contribution to scientific-scholarly progress, does have a basis in reality, that can best be illustrated by referring to an *historical* viewpoint. *History will show* which contributions to scholarly knowledge are valuable and sustainable. Informetric indicators do *not* measure contribution to scientific-scholarly progress in this sense, but tend to indicate attention, visibility or short term impact.

9.3 The Problem of Assessing Societal Impact

Two crucial problems in the assessment of societal impact relate to complexity to assess societal value, and to the time delay with which a scientific-scholarly achievement may generate societal impact, respectively. As argued in Chap. 8 in Part III, the choice of informetric tools in an assessment strongly depends upon the application context, namely, what is being assessed, how, and why. In itself it is defensible when a university's management team gives a large weight to societal aspects in defining an institution's profile and in internal research assessment processes. But Sect. 1.1 underlined that the assessment of societal value cannot be carried out in a politically neutral manner, and Sect. 6.3 argued that informetric experts should not sit in the chairs of politicians or policy makers.

The current author wishes to argue that this latter principle is especially relevant in the assessment of societal value. This is *not* to say that researchers are not allowed to have an opinion on what has societal merit, and express their views to colleagues or to the wider public. On the contrary, they do not only have the right to do so, but it may also be informative and helpful for policy makers or managers to see how experts in a field integrate scientific-scholarly and societal or political aspects in an issue as complex as the multi-dimensional assessment of research and setting its priorities. But the foundation of the criteria for assessing societal value is not a matter in which scientific experts have *qualitate qua* a preferred status, but should eventually take place in the policy domain.

The current author questions the appropriateness of a peer review-based research assessment exercise in which peers are requested to evaluate not only the scientific-scholarly merits but also the societal value of their research. This is also true for assessment in the medical sciences. For instance, the idea to measure the societal value of a clinical medical research activity aimed to develop a treatment against a particular disease on the basis of the number of patients that could potentially profit from a successful treatment, has certainly its merits, and such an indicator could provide one of the relevant pieces of information in setting priorities in medical research. But it is not merely—and perhaps, not primarily—the task of a review group of medical researchers to evaluate social merit on the basis of such essentially political criteria.

One possible approach to the assessment of societal impact would be to move away from the objective to evaluate an activity's societal *value*, and measure in a neutral manner researchers' *orientation* towards any articulated, lawful need in

society, as reflected in *professional contacts* in the broadest sense with subjects or institutions outside the academy interested in the knowledge generated inside her walls. Also, indicators of the amount of funding from external organizations, or measures of customer satisfaction can be used as they maintain a neutral position towards the objectives of these organizations or their reasons for being satisfied.

Any approach aimed to measure the notion of societal impact of research should take into account that a scientific finding that eventually enables a useful innovation may show this type of impact with a delay of many years. Chapter 15 in Part V of this book dedicates attention to the ‘science-technology helix’ developed by the Dutch physicist Hendrik Casimir, who estimated that the time lag between a fundamental result and its practical application is typically 15 years. Several studies have shown that the impact of new concepts and methods developed in the scientific domain of information science appeared in patents from the information or computer industry with a delay of 10 years or more (see for instance Halevi & Moed, 2014). This is why the emphasis on societal impact on the one hand, and an assessment focus on *recent* past performance on the other, are at least partially conflicting policy incentives.

9.4 The Effects of the Use of Indicators upon Authors and Editors

More and more studies are conducted on the effects of the use of informetric indicators on the practices of authors and editors. One group of studies focuses on the journal impact factor, and a second on changes in publication practices under the influence of large research assessment exercises.

Reedijk and Moed (2008) described mechanisms that may affect the numerical values of journal impact factors. While the authors underlined that they cannot make any statements about the *intentions* of journal editors, they illustrated how effective particular mechanisms *can be* in raising impact factors. Two important strategies that editors may adopt to raise the impact factor of their journals are as follows (for a definition of the journal impact factor the reader is referred to Chap. 16 in Part V).

- Publish expected top-cited papers as early as possible in a year, so that their contribution to the numerical value of the publishing journal’s impact factor is maximized. The authors found that papers in chemical journal issues published at the beginning of a year do indeed tend to have a slightly higher average citation impact than those papers in issues published at the end of a year, and concluded that, although the differences are small but, they are of interest when realizing that the underlying editorial behavior certainly is not (yet) general.
- Write “editorials”, preferentially in the last months of a year, citing many, if not all articles published in a particular journal in the two preceding years, so that the contribution of these citations to the impact factor is maximized. If this is done for

a recently launched journal, an even greater effect can be achieved by publishing this editorial as an introduction to the new journal also in other serials of the same publishing house, authored by the editors of these other serials. Reedijk and Moed described a number of such cases in detail. They also found in a set of 6000 journals that for 3% of journals the percentage of editorial self-citations was above 5. The data presented in this paper relate to the time period up until 2004. More recent data is not available. It must be noted that Thomson Reuters, the producer of the journal impact factor, currently monitors journals more carefully, and if certain behaviors are detected that strongly suggest indicator manipulation, a journal may even be removed from the Web of Science.

During the past 20 years, a series of research assessment exercises (currently denoted as Research Excellence Framework, or REF) have been conducted in the UK. The system gradually developed over time. From the very start the effect of the use of bibliometric indicators upon authors behavior has been intensively debated and investigated. Figure 9.1 gives a typical example. It presents a longitudinal analysis of UK science covering almost 20 years, analyzing trends in the annual growth rates of the percentage of UK *authors* and the percentage of *papers* published by UK authors, relative to the total population of authors and papers from all over the world and indexed in the Science Citation Index (currently Web of Science).

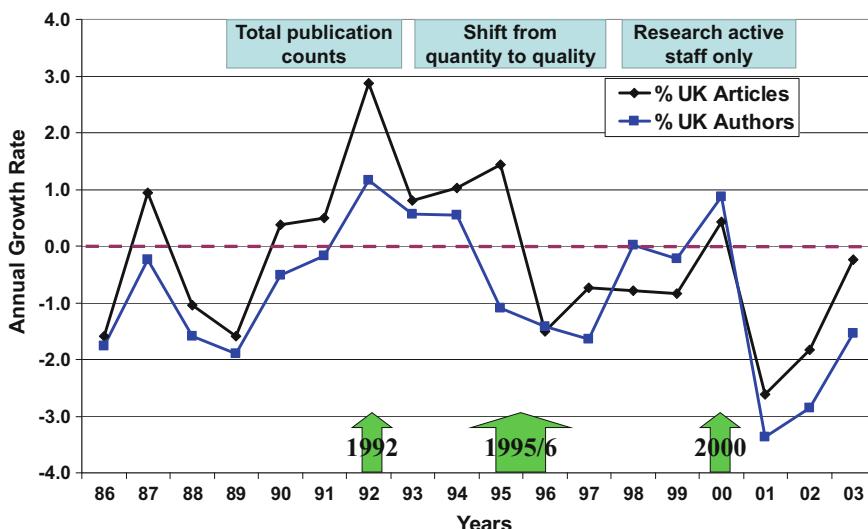


Fig. 9.1 Trends in the annual growth rates of the percentage share of UK authors and papers. Data relate to the time period 1986–2003 and were extracted from the Science Citation Index (SCI). The blue line represents the annual growth rate in the percentage share of publishing authors with affiliations in the UK relative to the total number of authors indexed in the SCI, and the black line the same rate for the share of publications with at least one UK affiliation, relative to the global number of publications. The text boxes at the top of the figure indicate the principal evaluation criteria applied in the RAE of 1992, 1996 and 2000, respectively. Copied with permission from Moed (2008)

Figure 9.1 reveals in the years prior to a Research Assessment Exercise (RAE 1992, 1996 and 2001) three distinct bibliometric patterns, that can be interpreted in terms of scientists' responses to the principal evaluation criteria applied in the RAE of a particular year. When in the RAE 1992 total publications counts were requested, UK scientists substantially increased their article production. When a shift in evaluation criteria in the RAE 1996 was announced from 'quantity' to 'quality', UK authors gradually increased their number of papers in journals with a relatively high citation impact. And during 1997–2000, institutions raised their number of active research staff by stimulating their staff members to collaborate more intensively, or at least to co-author more intensively, although their joint publication productivity increased less strongly.

The current author defends the proposition that in the assessment of possible effects of the use of indicators upon authors' or editors' practices, the crucial issue at stake is *not* whether scholars' practices *change* under the influence of the use of bibliometric indicators, but rather whether or not the application of such measures as a research evaluation tool *enhances research performance* and scholarly progress in general. In the cases shown above related to journal impact factors it is difficult to maintain that the observed changes indicate an increase in journal performance. Interpretation of the findings related to the RAE in the UK is more complex. One should not *a priori* exclude the possibility that the use of all three criteria have stimulated universities to increase the volume and the significance of their research efforts. On the other hand, one cannot claim *a priori* that this use indeed did have a positive effect upon performance. Without additional qualitative and quantitative evidence, it is impossible to draw any solid conclusions about this matter.

9.5 Constitutive Effects of Indicators and Magical Thinking About Research Quality

Those who show strategic behavior aimed to obtain the highest possible scores on particular indicators do not necessarily agree with the claim that these indicators are valid measures of research performance. They may disagree with it, but 'play the game' anyway. Also, policy makers using specific indicators may very well realize that these are only proxies, but consider them useful as they have certain intended effects upon research practices.

The notion of constitutive effects of indicators questions the validity of the distinction between intended and unintended effects, and claims that there is a tendency that the use of indicators of a theoretical concept such as research quality creates a reality in which the concept and the indicator tend to coincide (see for instance, Dahlen-Larsen, 2013). Thus, research quality would be more and more defined as to what citations measure. The author of this book believes that this critique should be taken very seriously, and would like to make the following comments.

If the tendency to replace reality with symbols, and to conceive these symbols as an even a higher form of reality, are typical characteristics of *magical* thinking, jointly with the belief to be able to change reality by acting upon the symbol, one could argue that the un-reflected, unconditional belief in indicators shows rather strong similarities with *magical* thinking. If such a belief is dominant, a situation would have emerged that, while modern science surpassed a magical attitude towards *nature*, such an attitude has emerged again, but now *towards science*, namely in research assessment and science policy.

A second comment questions the *empirical basis* of the hypothesis of the constitutive role of research performance indicators. This basis is in the perception of the current author still rather weak, and not seldom based on personal considerations and informal evidences. Relevant as these may be, if sound empirical evidence is lacking, the expression of the ‘constitutive-ness’ hypothesis itself could even become a self-fulfilling prophecy. Systematic, impartial empirical research is needed to further empirically test this hypothesis. In combination with a reflection upon whether indicators still measure what they are supposed to measure, and whether their use serves the formulated policy objectives, such an analysis may prevent researchers and policy makers to narrow the concept to one particular operationalization, and thus reduce their constitutive effects.

A complicating factor is that research evaluation processes and policy decision making processes are difficult to study, as they take often place in the sphere of confidentiality. Argumentation structures used in these processes to reach an evaluation or decision may not always be formally recorded, and, to the extent that they are, these recordings may not be available for research purposes by independent researchers. This issue is further discussed in Sect. 12.2 related to studies of the peer review system.

A third comment holds that it is important to consider a possible constitutive effect of informetric indicators in an *historical* perspective. In the 1960s and 1970s peer review was a dominant assessment methodology; peer review processes were less transparent than they are today. In quality assessment, a dominant view was that quality is what peers say it is. Following the argument of the constitutive-ness hypothesis, one could maintain that peer judgements had a constitutive effect, and that it is in reaction to their dominant position that indicator methodologies were developed, not to replace peer review, but to make it more transparent and to provide its judgements with a more solid basis.

If there is a genuine constitutive effect of informetric indicators in quality assessment at all, one should not point the critique on current assessment practices merely towards informetric indicators as such, but rather towards any claim for an absolute status of a particular way to assess quality. If the role of informetric indicators has become too dominant, it does not follow that the idea to intelligently combine peer judgements and indicators is fundamentally flawed and that indicators should be banned from the assessment arena. But it does show the combination of the two methodologies has to be organized in a more balanced manner.

One important, practical solution, discussed in the next section, is to enforce an assessment’s evaluative framework. A second solution is to dedicate in doctoral

programs more attention to the ins and outs, potential and limits of the various assessment methodologies. Research assessment is an activity one can learn.

9.6 The Need for an Evaluative Framework and an Assessment Model

Section 6.3 underlined the need to define an evaluative framework and an assessment model. To the extent that in a practical application an evaluative framework is absent or *implicit*, there is so to speak a *vacuum*, that may perhaps be too easily filled either with ad-hoc arguments of evaluators and policy makers, or with un-reflected assumptions underlying informetric tools. Such assumptions may have a normative character, but their validity cannot be established in quantitative-empirical, informetric research. Perhaps the role of such ad-hoc or un-reflected assumptions in research assessment has nowadays become too dominant. Their role can be reduced and kept within adequate boundaries only if evaluative frameworks becomes stronger, and more actively determine which tools are to be used, and how. This in its turn can only be achieved if informetricians make the normative assumptions of their tools explicit, so that policy makers and evaluators obtain a better understanding of the potential and the limits of these tools.

The previous paragraph stated that perhaps the role of ad-hoc and un-reflected assumptions in research assessment has become too dominant. The report by Benedictus and Miedema (2016) on the “obsession for metrics” mentioned in Sect. 1.2 is most informative and provides a confirmation of this statement. The qualification ‘perhaps’ is added because in the view of the current author the empirical basis of this statement is still limited. More studies on the effects of the use of performance indicators on research and assessment practices need to be conducted in order to have a more complete picture. But this does not mean that the current author believes that there are no problems at all and that all criticism is exaggerated.

Chapters 10, 11 and 12 in Part IV present a series of proposals for moving forward, both in the application of informetric indicators in research assessment, and in the development of new informetric approaches and indicators. The assumptions underlying current approaches are critically discussed and the pros and cons of the new approaches are discussed.

Chapter 10

The Way Forward in Quantitative Research Assessment

Abstract This chapter reflects on the assumptions underlying current practices in the use of informetric indicators in research assessment, and proposes a series of alternative approaches, indicating their pros and cons. The proposals relate to: measurement of communication effectiveness; new indicators of multi-dimensional research output and impact; the definition and application of minimum performance standards; academic policies towards the world university ranking systems; and an alternative approach to performance-based funding.

Keywords Journal internationality · Accreditation · Actual influence · Book publishers · Communication effectiveness · Journal functions · Leiden ranking · Libcitations · Magical thinking · Matthew effect · National journals · Open access · Potential influence · QS world university rankings · Research training · Scientifically developing countries · Specialist journals · Target audience · THE world university rankings · U-Multirank · World university rankings

10.1 Introduction

The aim of this chapter is to critically reflect on the assumptions underlying current practices in the use of informetric indicators in research assessment, and to propose a series of alternative approaches, indicating their pros and cons. It builds further upon the discussion of major problems in the use of informetric indicators presented in Chap. 9 in Part III. A base notion in this chapter is the proposition defended in Chap. 8, namely that the choice of an assessment methodology and its informetric indicators depends upon the subject of an assessment, the performance dimensions to be taken into account, and upon the assessment's objectives. The following issues are considered:

- Assessing *communication effectiveness* as a *precondition* for performance.
- The application of *minimum performance criteria* in the assessment of individual researchers.

- Subdividing scientific publications into subgroups on the basis of their *function* in the communication process and their *target* audiences, and calculating indicators per function rather than an aggregate score.
- Alternative ways to measure *journal internationality* other than calculating journal impact factors.
- Options in an academic policy towards the numerous world university rankings that are currently being published.
- A performance-based funding system that does *not* require a full assessment of all research activities.

10.2 Communication Effectiveness as a Precondition for Performance

As outlined in Sect. 1.2, a seminal article by Ben Martin and John Irvine (Martin & Irvine, 1983) defined ‘importance’ of a piece of research as its *potential* influence, and ‘impact’ as its *actual* influence, determined partly by its importance, but also by other factors, including communication practices, political pressures and visibility of authors. Of particular interest is their proposition as regards the *policy relevance* of these indicators. They claim that it is not the *potential* influence but the *actual* influence, not the importance but the impact that is most closely linked to the notion of scientific-scholarly progress, and that in an actual research assessment, it is not the *importance* but the *impact* that is of primary *interest to policy makers*. In the work of the Leiden Group in the early 1980s (Moed, Burger, Frankfort, & van Raan, 1985) a similar assumption was made. Martin and Irvine focused on relatively independent, targeted, *research institutes*. But is this claim a valid assumption of an assessment of an *academic institution*?

In *academic institutions*, especially research universities, it is generally considered appropriate to require academic staff to make contributions to scientific-scholarly progress. But is it defensible to require that they generate impact? What should be of primary interest to academic policy makers: importance (potential influence) or impact (actual influence)?

In current research assessment practices the measurement of citation impact plays an important role. But citation impact indicators are not only influenced by the importance of a piece of research, but also by the extent to which researchers bring their work to the attention of potentially interested audiences. The first relates to scientific-scholarly content, and the second to how this content is exposed in the market of scientific ideas. In citation impact measurement it is difficult if not impossible to separate these two factors. Also, as argued in Sect. 9.2, informetric indicators tend to measure attention, visibility or short term impact, rather than sustainable contributions to scientific-scholarly progress.

An academic assessment policy is conceivable that rejects the claim that impact rather than importance is the key aspect to be assessed, and discourages the use of citation data as a *principal* indicator of importance. It is based on the assumption

that, as importance can only be measured in a longer term perspective using a combination of peer review and informetric tools as the primary method, a key assessment criterion from a short term perspective should relate to researchers' *communication strategies*, namely to the extent to which researchers bring their work to the attention of a broad, potentially interested audience. The *effectiveness* with which researchers do this can in principle be measured with informetric tools.

Such an assessment process would *not* aim at measuring importance or *contribution to scientific-scholarly progress* as such, but rather *communication effectiveness*, a concept that relates to a *precondition* for performance rather than to performance itself. A base assumption in such a process is that bringing research work to the attention of an as large as possible audience is a valuable objective in an academic research policy, and deserves to be stimulated and rewarded.

This shift in perspective from research quality to communication effectiveness lays a focus on communication processes. At least the following three topics are of great interest, but the list is far from comprehensive. A *first* important topic is that of access modalities and publication business models. Although many interesting studies have been conducted on the 'effects' of the Open Access model (Bjork & Solomon, 2012), more research is needed that relates these studies to research assessment of researches and their institutions. An important study along this line is published by Torres-Salinas, Robinson-Garcia and Aguillo (2016) on the effects of publishing in Gold Open Access journals upon output and impact measures of Spanish researchers. A related issue is the definition of an informetrically appropriate basis to establish a fair market value and author royalties for publication products such as journal articles and books.

A *second* approach is to systematically study the role of book and other non-journal sources in scholarly communication, using informetric data from the large citation indexes Web of Science, Scopus or Google Scholar, as an extension of the work by, for instance, Kousha, Thelwall and Rezaie (2011) on Google Scholar citations to books by UK authors, Torres-Salinas and Moed (2009) and White et al. (2009) on inclusions of books in library catalogs or 'libcitations', and Zuccala, Guns, Cornacchia and Bod (2015) on ranking of book publishers.

A *third* approach would focus on scientific-scholarly journals and explore along the lines once initiated by Eugene Garfield, and picked up by Zitt and Bassecoulard (1998) and others, new indicators of the position of journals in the international communication network. A typical example of this approach is presented in Sect. 10.3.

10.3 Some New Indicators of Multi-dimensional Research Output

As argued in Sect. 10.1, in the development of new indicators the *functions* of publications and other forms of scientific-scholarly output, as well as their *target audiences* should be taken into account more explicitly than they have been in the past. The aim of this section is to illustrate by means of typical examples how this could be done.

10.3.1 Journal Functions and Target Audiences

Firstly, scientific-scholarly journals may have different functions and target audiences. The most important category relates to journals aimed towards the international research front, and reporting original contributions to scientific-scholarly progress. A second group is primarily directed towards a *national* audience, and aims to transfer knowledge on new research findings and techniques to interested researchers and practitioners, who are themselves less active at the international research front. For instance, in clinical medicine, national publishers in many countries publish journals informing medical practitioners about the most recent developments at the international research front. In humanities, journals may provide a national scholarly audience with information about interesting developments in other countries.

Journals with such an educative or enlightening function are important in scientific-scholarly communication, and tend to have a substantial *societal value*. Since their visibility at the international research front as reflected in citations and journal impact factors may be low, in a standard bibliometric analysis based on publication and citation counts this value may not be visible.

It could even happen that researchers who publish in such journals are ‘penalized’ for this in a citation impact analysis calculating citation-per-article ratios or measures derived from this average. Since publications in poorly cited journals tend to be poorly cited themselves, they may reduce an entity’s average citation rate. This may lead to the paradox that if a national journal from a particular country is removed as source from a citation index, so that the articles published in it are *not* taken into account in the calculation of the country’s average citation rate, this rate may *increase*. This is particularly true for journals publishing articles in a language other than English (van Leeuwen et al., 2001).

Journals could be systematically categorized according to their function and target audience, and separate indicators could be calculated for each category. In an analysis of research output in journals directed towards *national* audiences, citation-based indicators are *less relevant*. At the same time, in citation analyses based on the large international citation indexes focusing on the international research front, it would be appropriate to disregard such journals. The question as to which weights should be given to the various aspects in an assessment should be answered in the assessment’s evaluative framework.

It would be helpful to develop and publish for an as large as possible set of journals one or more indicators of journal *internationality*. An operational distinction can be made between nationally and internationally oriented publication outlets, based on the *geographical* distribution of their *publishing*, *citing* and/or *cited* authors. As a typical example of this approach, Fig. 10.1 presents for seven Italian journals their citation impact and their internationality. For a definition of these indicators the reader is referred to the legend to Fig. 10.1. It shows that journals with similar citation impact may have quite different values of internationality, and hence that journal impact and internationality are by no means identical concepts.

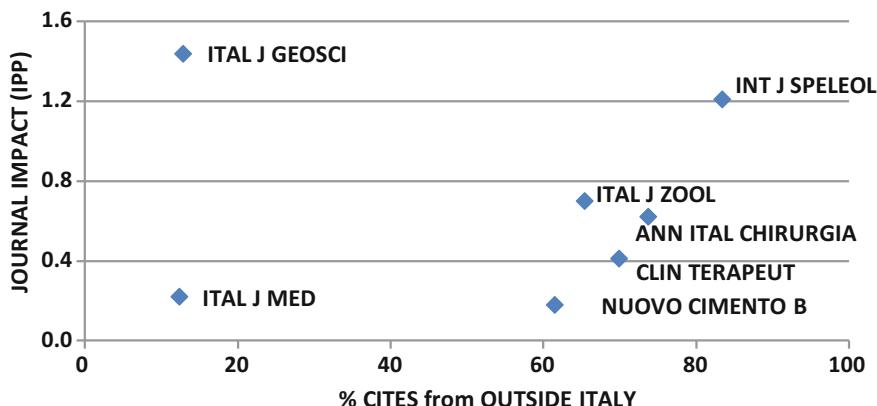


Fig. 10.1 Citation impact versus international orientation for 7 Italian journals. Data on both indicators are obtained from Scopus.com. The journal impact measure is the IPP (Journal impact per paper), the numerator of the SNIP (Source Normalized Impact per Paper) metric, available both via Scopus.com, in Elsevier's website for journal metrics (www.journalmetrics.com), and also in the website www.journalindicators.com created by the Centre for Science and Technology Studies (CWTS) at Leiden University. Data on internationality were extracted from Scopus.com in the following manner. Taking into account the IPP citation window, for each journal a set of all citing articles was created. Conducting additional queries, the number and percentage of citing articles was determined of which none of the authors has an Italian affiliation. The figure shows that for Italian Journal of Geosciences this percentage amounts to 12.9, and for International Journal of Speleology to 83.3%. Both journals have a relatively large citation impact (IPP)

10.3.2 A Note on Journal Coverage of the Citation Indexes

In some national research assessment exercises articles in journals indexed in the Web of Science or in Scopus have a special status. In the assessment of science fields only articles published in these journals are considered ‘valid’ or ‘genuine’ scientific publications, and taken into account in publication counts. Users of these counts as performance indicator should be warned against this type of use. It is based on the assumption that the current citation indexes are created on the basis of the principles applied by Eugene Garfield in the creation of the Science Citation Index, according to which the Index has a selective source coverage of journals with a relatively high citation impact.

But currently, database producers may apply other criteria for the selection of journals as well. This is especially true for Scopus, the source coverage of which is more comprehensive than that of Web of Science, as is clearly illustrated in the comparative studies of Lopez-Illescas et al. (2008) in the field oncology, and Gorraiz and Schloegl (2008) on pharmacological journals. For more details on the former study, the reader is referred to Sect. 14.2 in Part V. A secondary analysis of data from Scopus revealed for the (citation) year 2011, considering target articles published during the three preceding years, that 25% of journals in Scopus receives at most 15 citations (in a three-year time window) and 10% at most 3 citations. 25%

of journals has an impact-per-paper (IPP) below 0.21 (Moed, 2016b). Many of these sources are national journals, often from scientifically developing countries, or—from the point of view of the international research front—more peripheral journals in scientifically developed nations.

Section 10.3 further discusses the definition of indicators of communication effectiveness. An important aspect is that the *functions* of publications and other forms of scientific-scholarly output, as well as their *target audiences* are considered more explicitly than they have been in the past.

10.3.3 Research Training and Scientifically Developing Countries

A crucial societal function of academic institutions is *training* of researchers. There are several non-bibliometric indicators of academic educational performance, for instance, those based on the number of doctoral degrees attained. Publication counts per category of scientific staff would be useful to indicate the role of the various categories in the research process. This would require at the level of an academic institution the linking of a research output file with a dataset on scientific staff, in which the names of the latter could be merely intermediary variables, aimed to link a piece of output to a category of academic staff producing it.

One socially highly relevant aspect of research training is the contribution academic institutions make to the development of scientifically developing countries. If one agrees that carrying out this type of training is most valuable from a societal point of view and worthwhile being presented as a separate activity to the outside world, it would be feasible to make this contribution more visible in informetric indicators. Publications emerging from research training of researchers from scientifically developing countries could be labeled and counted separately; the same is true for scientific collaborations with research groups from these countries.

10.4 Definition of Minimum Performance Standards

Section 7.2 showed how evaluative assumptions shape informetric indicators and addressed the issue in which part of the quality distribution a research performance assessment is interested: the top or the bottom. Focusing on the top, questions arise such as: which groups are world leaders in the field? Which group has made the largest contribution to a field in the past 10 years? Who has obtained the largest h-index in the field? When outcomes of citation analyses are published in daily newspapers, they relate almost always to the top. The current section explores the use of indicators in the bottom segment of the quality distribution, namely in defining *minimum* performance standards.

The use in academic policies of minimum acceptable performance standards is not new. For example, in the 1970s and 1980s in universities in the Netherlands and many other countries, a debate took place on whether a permanent academic staff member should have a Ph.D. degree. One of the arguments against this proposal was that there were at that time among permanent staff many researchers who performed well in research, but who did not have a Ph.D. degree. The proponents replied that, given the need to establish an internal quality assurance system, having a doctoral degree is a fair and appropriate requirement, and that the performant staff without Ph.D. would probably not have great difficulty in attaining it.

The notion of minimum performance standards plays also a role in national research assessment exercises, in the definition of being *research active*. For instance, in the national assessments in the UK (currently REF, Research Excellence Framework) being research active is a necessary condition for an academic staff member to be allowed to submit research outputs to the assessment. Also in research assessment processes in Australia this concept played a role. For example, ‘research active’ might be defined as achieving at least three out of the following 4 average performance levels per year: one scholarly publication; one Ph.D. completion; one project funded from a competitive funding source; €50,000 income.

One possible approach to the use of informetric indicators in research assessment is a systematic exploration of indicators as tools to set minimum performance criteria and define in this way a performance baseline. Important considerations in favor of this approach are as follows. Firstly, a comparison between peer ratings of the performance research departments with indicators of citation impact of these departments found that citation rates are a good predictor of how peers discriminated between a ‘valuable’ and a ‘less valuable’ past performance, but do not properly predict within the class of ‘valuable’ performances, peers’ perception of ‘genuine excellence’ (Moed, 2005, p. 57). This outcome underlines the potential of informetric indicators in the assessment of the *lower part* of the quality distribution.

Secondly, Sects. 9.4 and 9.5 noted that the application of particular indicators in research assessment may change research and communication practices of researchers, and that indicators may have constitutive effects to the extent that a theoretical concept such as quality may in the minds of researchers and policy makers easily be narrowed to its operationalization (e.g., quality is what citations measure). Using indicators to define a baseline, researchers will most probably change their research practices as they are stimulated to meet the standards, but if the standards are appropriate and fair, this behavior will actually increase their performance and that of their institutions.

Focusing on minimum criteria involves a shift in perspective from measuring performance as such towards assessing the *preconditions* for performance. Expert opinion and background knowledge will play a crucial role not only in the definition of the standards themselves, but also in the assessment processes in which these are applied. As indicated in Sect. 4.2, there is currently no clear, generally accepted definition of being research active across universities, countries and disciplines. This means that a lot of work is to be done to define the performance standards in a proper way.

The standards should take into account differences in publication and communication practices among subject fields, and be in social science and humanities different from those in the natural and life sciences. Although it is conceivable that national accreditation organizations will assess the foundation of these criteria and the consistency with which they are applied in an institution,—a task that will lead to a certain degree of standardization—, it is not unthinkable that differences will emerge in the threshold levels among universities, and that these levels themselves start functioning as research quality markers of institutions.

The definition of minimum standards could also be applied to *journal impact measures*. Rather than focusing on the most highly cited journals and rewarding publications in this top set, it would be possible to discourage publication in the bottom set of journals with the lowest citation impact. Further empirical research should be conducted to set an appropriate minimum citation impact threshold. Putting too strong an emphasis on publication in highly cited journals may discourage researchers to publish in the good, specialist journals in their field, as such journals often do not have a citation impact as high as general or multi-disciplinary outlets. On the other hand, as argued in Sect. 10.2, using as quality criterion whether or not a journal is indexed in one of the citation indexes would be inappropriate.

It must be underlined that the current author does not claim that the use of indicators to assess the top of the quality distribution must be banned from the assessment arena. The following approach focusing on the upper part of the quality distribution could be considered: distinguishing in the top of the distribution entities that are '*hors catégorie*'—using a French term used to categorize the very highest mountains to be climbed in the bicycle race Tour de France, or '*at Nobel Prize level*'. Assessment processes focusing on the top of the quality distributions could further operationalize the criteria for this qualification.

For instance, during the 1970s and 1980s, a professor of immuno-hematology in the Department of Medicine at Leiden University made amongst other achievements major contributions to the development of organ transplantation techniques. He was chairperson of several global task forces and expert groups, and regular publisher of papers in the most impactful journals. It appeared that 15% of all citations to articles published from the Department of Medicine (that contained hundreds of staff members), were (co-)authored by this researcher, and that his performance was generally considered beyond compare or 'at Nobel Prize level'.

10.5 Policy Towards World University Rankings

University ranking systems have been intensely debated, for instance by van Raan (2005), Calero-Medina et al. (2008), Salmi (2009), Hazelkorn (2011), Rauhvargers (2011; n.d.) and Shin, Toutkoushian and Teichler (2011). A report from the European University Association concluded that despite their shortcomings, evident biases and flaws, rankings are here to stay.

For this reason it is important that universities are aware of the degree to which they are transparent, from a user's perspective, of the relationship between what it is stated is being measured and what is in fact being measured, how the scores are calculated and what they mean. (Rauhvargers, 2011, p. 7)

To provide users insight into the value and limits of world university rankings, Chap. 18 in Part VI of this book presents a comparative analysis of 5 World University Ranking Systems: ARWU, the Academic Ranking of World Universities, also indicated as 'Shanghai Ranking'; The Leiden Ranking created by the Centre for Science and Technology Studies (CWTS); The Times Higher Education (THE) World University Ranking; QS World University Rankings; and an information system denoted as U-Multirank created by a consortium of European research groups.

As all ranking systems claim to measure essentially academic excellence, one would expect to find a substantial degree of consistency among them. The overarching issue addressed in the paper presented in Chap. 6 is the assessment of this consistency-between-systems. To the extent that a lack of consistency is found, what are the main causes of the observed discrepancies? The following conclusions were drawn.

- Each ranking system has its proper orientation or 'profile'; there is no 'perfect' system. There is only a limited overlap between the top 100 segments of the 5 rankings.
- What appears in the top of a ranking depends to a large extent upon a system's geographical coverage, rating methodologies applied, indicators selected and indicator normalizations carried out.
- Current ranking systems are still one-dimensional in the sense that they provide finalized, seemingly unrelated indicator values rather than offer a dataset and tools to observe patterns in multi-faceted data.
- To enhance the level of understanding and adequacy of interpretation of a system's outcomes, more insight is to be provided to users into the methodological differences between the various systems, especially on how their orientations influence the ranking positions of given institutions.

University managers may use their institution's position in world university rankings primarily for marketing purposes, but they should not disregard the negative effects such use may have upon researchers' practices within their institution, and they should also critically address the validity of the methodological claims made by ranking producers.

The author of this book would like to suggest a strategy towards these ranking systems. Academic institutions could, individually or jointly, seek to influence the various systems by formally sending them a request to consider the implementation of the following new features.

- Offer more advanced analytical tools, enabling a user for instance to cross-tabulate indicators.
- Provide more insight into how the methodological decisions of their producers influence the ranking positions of given universities.

- Enhance the information in the system on additional factors, such as teaching course language.

In addition, academic institutions could proceed as follows.

- Create a special university webpage providing information on a university's internal assessment and funding policies; on its various types of performance; and giving comments on the methodologies or outcomes of ranking systems.
- Request ranking producers to make these pages directly accessible via their systems.

10.6 An Alternative Approach to Performance Based Funding

National research assessment exercises of the type conducted in the UK, Australia or Italy tend to require enormous efforts, both in the collection of data at the institutions and in the peer review process evaluating numerous submissions. One of the severe points of critique against such exercises is that they tend to increase disparities between institutions over time, to the extent that there is a tendency for institutions that have performed well in the past to obtain larger amount of funding than less performing institutions. The driving mechanism in such a funding system is denoted as 'success breeds success' or the 'Matthew Effect', a term coined by Harriet Zuckerman and Robert K. Merton.

Bishop (2013) created a simulation of the dynamics of a funding system based on such mechanism. In the system's initial state all institutions have the same number of research active staff, assigned at random from a population among which research quality is normally distributed. A key assumption is that a fixed percentage of staff moves from one institution to another, and that institutions with higher levels of funding can attract better replacement staff than institutions with lower funding. The new staff constitutes the basis for funding allocations in a new cycle. The distribution of funds across institutions is initially normal, but becomes more and more skewed, and results in a configuration of a limited number of elite universities and a large pool of poorly funded institutions.

A crucial issue is whether the goal of a funding formula should be to focus on elite institutions or distribute funds more widely. To inform this policy debate, informetric research on the effects of the various funding models upon national and institutional research performance is one of the very important research topics in the field of informetrics and research assessment. Adopting an informetric viewpoint, the current section addresses whether alternative assessment models are conceivable that would require less efforts and reduce to some extent the *Matthew Effect*.

It presents an approach to a national research assessment exercise that focuses on *emerging groups*, rather than on the total set of *individual* academic staff or the

research active part of it. An emerging group is a small research group that is expected to have a great scientific potential. It would normally consist of a young director who has performed very well during the Ph.D. phase and as post-doctoral student.

Acknowledged as a ‘rising star’, the director has developed a promising research program, and has already been able to establish a small research unit, typically consisting of a few Ph.D. students and one or two post docs, financed from funds allocated within the institution and partly from externally funded research projects. The profile of an emerging group described above should be further operationalized into a set of *minimum* quantitative criteria, taking into account the communication practices and funding opportunities in the group’s subject field.

In the assessment procedure, institutions submit *groups* rather than *individual* staff, and *emerging* rather than *established* groups. Submissions provide information on past performance and a future programme, which are evaluated in a peer review process, informed by appropriate informetric indicators. The primary aim of the peer review is to define the minimum standards in operational terms and assess whether the submitted groups comply with these standards.

The idea to combine metrics and peer review in this way is based on experiences collected in the ‘Leiden’ study in the 1980s, according to which the research group is the most appropriate unit of assessment in research at least in the natural and life sciences, and that time series of indicators can be constructed that aim to reflect the emergence of new, promising research groups.

Elements of this proposal are already included in existing grant application procedures of national research councils aimed to support young researchers. A key difference between these procedures and the proposed assessment is that the latter would focus on defining and applying *minimum criteria* as a precondition for future performance. There would be *no* need to rank groups, assign ratings, discriminate between ‘top’ and ‘almost top’ groups, or make funding decisions about them. The latter decisions take place *within* their institution.

A second difference is that in the current proposal the set of minimum quality criteria is defined in advance and known to all participants. Also, the availability of a certain amount of funding from *internal*, performance-based allocation processes could be posed as a necessary condition for being granted. In this way, the proposed funding procedure not only enhances the transparency of the process, but also stimulates the implementation of quality control processes *within* an institution.

A part of public funding (block grant) could be allocated to institutions as a lump sum on the basis of the number of acknowledged emerging groups. As an additional criterion, peer review could assess the degree to which the submitted group is well integrated into the research and teaching infrastructure within the submitting institution.

10.7 Concluding Remark

The practical realization of the suggestions and proposals made in this chapter requires a large amount of informetric research and development. They constitute important elements of a wider R&D program of *applied evaluative informetrics*. The further exploration of measures of communication effectiveness, minimum performance standards, new functionalities in research information systems, and tools to facilitate alternative funding formula, should be conducted in close collaboration between informetricians and external stakeholders, each with their own domain of expertise and responsibilities.

Chapter 11

A Perspective on Altmetrics

Abstract This chapter discusses the potential of altmetrics. It starts with analyzing what the main drivers of altmetrics are. A multi-dimensional conception of altmetrics is proposed, namely as traces of the computerization of the research process, and as a tool for the practical realization of the ethos of science and scholarship in a computerized or digital age. An attempt is made to provide a theoretical foundation of altmetrics, based on notions developed by Michael Nielsen in his monograph *Reinventing Discovery: The New Era of Networked Science* (Nielsen, 2011).

Keywords Architecture of attention · Computerization · Critical mass · Digitization · Ethos of science · Open science · Politically neutral · Reinventing discovery · Zotero

11.1 Introduction

In the Altmetrics Manifesto published on the Web in October 2010, the concept of “Altmetrics” is introduced as follows. It gives an overview of the increasingly important role of social media such as Twitter and Facebook, online reference managers such as Mendeley and Zotero, scholarly blogs and online repositories in scientific scholarly communication. It underlines that the activities in these online tools can be tracked: “This diverse group of activities forms a composite trace of impact far richer than any available before. We call the elements of this trace altmetrics” (Priem et al., 2010). More and more practitioners use the new term

This chapter re-uses with permission selected paragraphs from: Moed, H.F. (2016a). Altmetrics as traces of the computerization of the research process. In: C.R. Sugimoto (ed.), *Theories of Informetrics and Scholarly Communication (A Festschrift in honour of Blaise Cronin)*. ISBN 978-3-11-029803-1. Berlin/Boston: Walter de Gruyter. It is based on an earlier key note presentation ‘Altmetrics as traces of the computerization of the research process’ at the Altmetrics’14 Conference, Indiana University, Bloomington, USA, 23 June 2014; and on a key note ‘(Alt)Metrics-Based Research Assessment’ at the STI-ENID Conference, Leiden, 3–5 Sept. 2014.

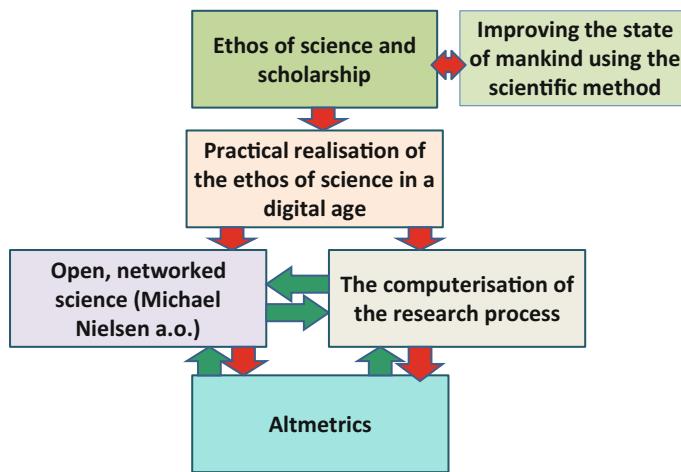


Fig. 11.1 Drivers of Altmetrics

“alternative metrics” rather than “altmetrics”. In this section the original term is used. Section 4.3 gives a concise overview of the various altmetric tools. (Fig. 11.1)

Three drivers of development of the field of altmetrics can be distinguished

- Firstly, in the policy or political domain, there is an increasing awareness of the multi-dimensionality of research performance, and an increasing emphasis on societal merit, an overview of which can be found in Chap. 3 in Part 2.
- In the domain of technology, a second driver is the development of information and communication technologies (ICTs), especially websites and software in order to support and foster social interaction. The technological inventions mentioned in the Altmetrics Manifesto are typical examples of this development.
- A third driver, primarily emerging from the scientific community itself, is the Open Science movement. Open Science is conceived as the movement to make scientific research, data and dissemination accessible to all levels of an inquiring society, amateur or professional (“Open Science”, n.d.).

For an expression of the ‘ethos of science’ one could perhaps best refer to Francis Bacon’s proposal “for a universal reform of knowledge into scientific methodology and the improvement of mankind’s state using the scientific method” (“Francis Bacon”, n.d.). The ethos captured in the current section does not merely relate to the positive, empirical sciences, but to science and scholarship in general, including, for instance, hermeneutic scholarship. Bacon’s proposal develops *two* base notions, namely the notion that science can be used to improve the state of mankind, *and* that it is governed by a strict scientific-scholarly methodology. Both dimensions, the *practical* and the *theoretical-methodological*, are essential in his idea.

A huge challenge nowadays is how the ethos of science and scholarship, admittedly outlined so vaguely above, must be realized in the modern, computerized, or digital age. The state of development of information and communication technology (ICTs) creates enormous possibilities for the organization of the research process, as well as for society as a whole. It is against this background that the emergence and potential of altmetrics should be considered.

In the next sub-sections, a multi-dimensional conception of altmetrics is proposed, namely as traces of the computerization of the research process. “Computerization” should be conceived in its broadest sense, including all recent developments in ICT and software, taking place in society as a whole. It is argued that altmetrics can provide tools not only to reflect this process passively, but, even more so, to design, monitor, improve, and actively facilitate it. From this perspective, altmetrics can be conceived as tools for the practical realization of the ethos of science and scholarship in a computerized or digital age. An attempt is made to provide a theoretical foundation of altmetrics, based on notions developed by Michael Nielsen in his monograph *Reinventing Discovery: The New Era of Networked Science* (Nielsen, 2011).

11.2 The Computerization of the Research Process

Figure 11.2 visualizes four aspects of the computerization of research process.

- Firstly, at the level of the everyday research practice, there is the collection of research data and the development of research methods. A “classical” citation analysis found that in many disciplines, computing-related articles are the most heavily cited (Halevi, 2014). This marks the key role of computerization in current research practices. Interestingly, the most frequently cited article in social sciences is about user acceptance of information technology.
- The second aspect relates to scientific information processing. A topic of rapidly increasing importance is the study of searching, browsing, and reading behavior of researchers, based on an analysis of the electronic log files recording the usage of publication archives such as Elsevier’s Science Direct or an Open Access archive such as arxiv.org. A comparison of citation counts and full text downloads is presented in Chap. 19 in Part 6.
- Communication and organization is a third group of aspects. These two elements are distinct, from an altmetric point of view, to the extent that the first takes place via blogs, Twitter and similar social media, whereas the second occurs for instance in scholarly tools as Mendeley or Zotero.
- The use of informetric indicators in research assessment is a fourth aspect of the computerization of the research process. Mentions of authors and their publications in social media like twitter, in scholarly blogs and in reference managers form the basis of the exploration of new impact measures. For a critical discussion of the role of altmetrics in research assessment the reader is referred to Sect. 3.2.

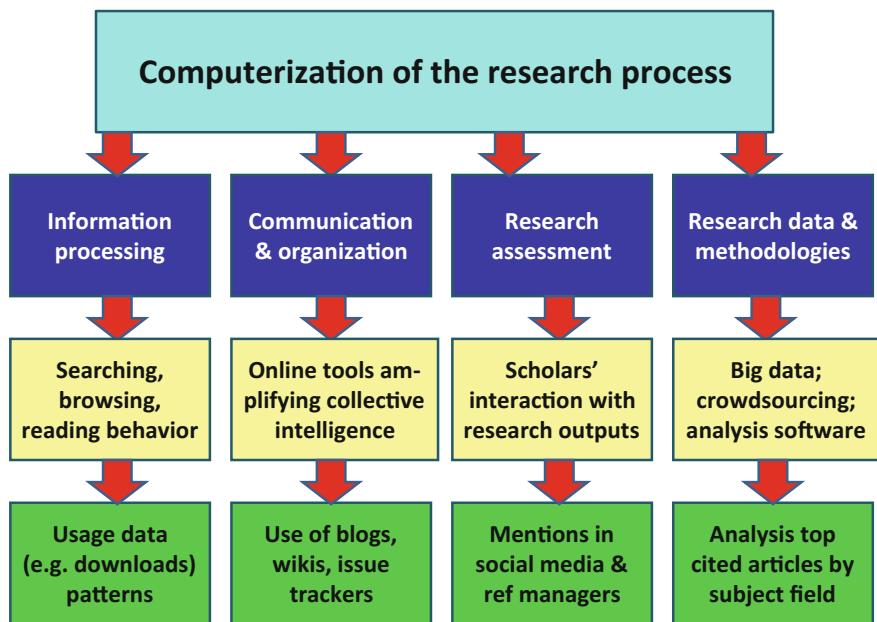


Fig. 11.2 Four aspects of the computerization of the research process

11.3 Michael Nielsen's "Reinventing Discovery"

Michael Nielsen's (2010) monograph presents a systematic, creative exploration of the actual and potential value of the new ICT for the organization of the research process. This sub-section summarizes some of the main features of this thinking.

In building up his ideas, Nielsen borrows concepts from several disciplines, and uses them as building blocks or models. A central thesis is that online tools can and should be used in science to amplify collective intelligence. Collective intelligence results from an appropriate organization of collaborative projects. In order to further explain this, he uses the concept of 'diversity', borrowed perhaps from biology, or its sub-branch, ecology, but in the sense of cognitive diversity, as he states: "To amplify cognitive intelligence, we should scale up collaborations, increasing cognitive diversity and the range of available expertise as much as possible" (Nielsen, 2010, p. 32).

As each participant can give only a limited amount of attention in a collaboration, there are inherent limits to size of the contributions that participants can make. At this point the genuine challenge of the new online tools comes into the picture: they should create an "architecture of attention," and in my view one of the most intriguing notions in Nielsen's work, "that directs each participant's attention where it is best suited—i.e., where they have maximal competitive advantage" (Nielsen, 2010, p. 33).

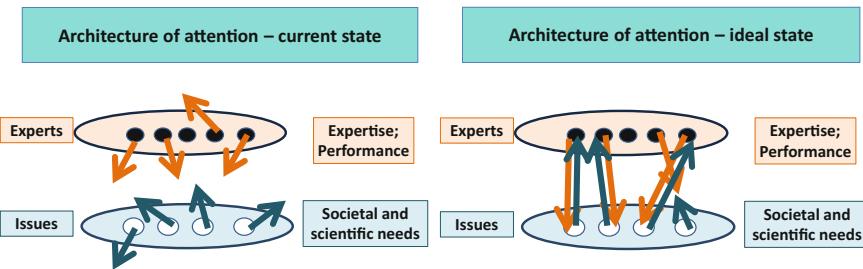


Fig. 11.3 Architecture of attention: current and ideal state

In the ideal case, scientific collaboration will achieve what he terms as “designed serendipity,” so that a problem posed by someone who cannot solve it finds its way to one with the right micro expertise. Using a concept stemming from statistical physics, namely, critical mass, he further explains that “conversational critical mass is achieved and the collaboration becomes self-stimulating, with new ideas constantly being explored” (Nielsen, 2010, p. 33). Important tools in the optimization of collaboration is ‘modularization’; the creation of a “data web”, defined as “a linked web of data that connects all parts of knowledge” and “an online network intended to be read by machines”; and crowd sourcing techniques.

Michael Nielsen’s set of creative ideas can be used as a framework in which altmetrics can be positioned. Their role would not merely be, rather passively, that of descriptors, but, actively, or proactively, as tools to establish and optimize Nielsen’s “architecture of attention”, a configuration that combines the efforts of researchers and technicians on the one hand, and the wider public and the policy domain on the other, as visualized in Fig. 11.3.

11.4 Useful Distinctions

To further explore the potential and limitations of altmetrics, Table 11.1 highlights a series of distinctions that are often made in the context of the use in research assessment of “classical” metrics and publishing, but that are in my view most relevant in connection with altmetrics as well.

It is essential to distinguish also in altmetrics between *scientific-scholarly* and *societal* impact along the lines discussed in Chap. 3, and to realize, as argued in Sect. 9.4, that societal merit cannot be assessed in a politically neutral manner. Speaking of the ethos of science above, *two* dimensions were highlighted: a practical and a theoretical-methodological: science potentially improves the state of mankind, *and* is governed by strict scientific-scholarly methodology. The position of the current author is that, in order to be successful, the project proposed by Bacon and so many others, requires a certain distance and independence from the

Table 11.1 Important distinctions and their implications for the use of altmetrics

Distinction	Implications for altmetrics
Scientific-scholarly ↔ Societal impact	Validity of knowledge claims is to be examined according to strict methodological rules; social media reflect primarily societal attention
Attention ↔ influence	Altmetrics tend to measure attention rather than influence. Reflection needed in order to assess significance
Opinion ↔ scientific finding or fact	Blogs tend to express opinions rather than scientific facts
Peer-reviewed ↔ non-peer reviewed	Social websites have no quality control mechanisms
Intended ↔ unintended effects; constitutive effects	Attitude towards social media determines visibility therein
Formal ↔ natural ↔ life ↔ social ↔ technical science	Differences in validity and utility of altmetrics among the sciences

political domain, and most of all, a strong, continuous defense of proper methodological rules when making knowledge claims and examining their validity.

It is important also in altmetrics to distinguish between *attention* and *influence*, as discussed in Sect. 9.3, and acknowledge the need to carry out a reflection upon research findings to assess their significance. Altmetrics tend to measure attention rather than influence. At the level of altmetrics indicators, scholarly reference managers reflect scientific scholarly, while Twitter and social media count primarily societal attention.

A next relevant distinction, though in practice perhaps difficult to make, is between scientific *opinion* and scientific *fact* or result. In journal publishing, many journals distinguish between research articles on the one hand, and opinion pieces, discussion papers, or editorials on the other. At least in the empirical sciences, the first type ideally reports on the outcomes of empirical research conducted along valid methodological lines, and discusses their theoretical implications. The second type is more informal, normally not peer-reviewed, and speculative. The two types have, from an epistemological point of view, a different status, it is crucial to keep this in mind when exploring the role of altmetric data sources containing scholarly commentaries, such scientific-scholarly blogs.

At this point, it is also important to distinguish between speculations or *opinion* pieces related to *scientific-scholarly* issues, and those primarily connected with *political* issues. It is in the interest of the ethos of science to be especially alert to a practice in which researchers make political statements using their authority as scientific-scholarly experts. Such practices should be rigorously unmasked whenever they are detected.

Intended versus *unintended* consequences of particular behavior is a next relevant distinction, along with the notion of *constitutive* effects of indicators, as discussed in Sects. 9.5 and 9.6. As indicated in Sect. 9.4, Thelwall warns that the problem of *manipulability* is much larger in case of altmetrics than it is in the application of citation indices (Thelwall, 2014). Finally, it is also crucial to

distinguish the various domains of science and scholarship, for instance, natural, technical, formal, biological, medical, social sciences, and humanities. Although such subject classifications suffer from a certain degree of arbitrariness, it is important to realize that the research process, including communication practices, reference practices, and orientation towards social media, may differ significantly between one discipline and another.

11.5 Concluding Remarks

One of the limitations of the model Michael Nielsen proposes in his monograph *Reinventing Discovery* should be highlighted: the use of the open source software development as a model of collaboration may fit the domain of the formal sciences rather well, but may be less appropriate in many subject fields in humanities and social sciences. In other passages in his monograph he is aware that this organizational model may not be appropriate in all domains of science and scholarship.

In the same way that classical citation metrics are often uniquely linked to the use of journal impact factors for assessing individual researchers—although so many other citation-based metrics and methodologies have been developed, applied to different aggregations and with different purposes—altmetrics runs perhaps a danger of being too closely linked with the notion of assessing individuals by counting mentions in Twitter and related social media, a practice that may provide a richer impression of impact than citation counts do, but that clearly has its limitations as well (e.g., Cronin, 2014).

Altmetrics, and science indicators in general, are much more than that. Apart from the fact that much more sophisticated indicators are available than journal impact factors or Twitter counts, these indicators do not have a function merely in the assessment of research performance of individuals and groups, but also in the study of the research process. In this way, in terms of the distinctions made in Sect. 3.2, these indicators are used as *input* (research infrastructure) and *process* (collaboration) indicators rather than as output or impact measures. Also, like science metrics in general, altmetrics does not merely provide reflections of the computerization of the research process, but can, in fact, develop into a set of tools to further shape, facilitate, design, and carry out this process.

Chapter 12

The Way Forward in Indicator Development

Abstract This chapter proposes a series of alternative approaches in the development of informetric indicators for research assessment. It starts with a proposal for new indicators of the manuscript peer review process. Next, it presents an approach towards an ontology-based, informetric data management system, and launches the idea of creating informetric self-assessment tools. Finally, it illustrates the important role of models in the analysis of informetric data, by presenting a case study on scientific development.

Keywords Computational linguistics · Digital humanities · Indicator dimensions · Internationally co-authored · OBDM · Ontology · Persian gulf · Referee report · Reviewers · Scientific development · Self-assessment · South-East asia

12.1 Towards New Indicators of the Manuscript Peer Review Process¹

12.1.1 Introduction

Any academic journal is only as good as its peer review. And yet, although past investigators have conducted valuable studies on biases in the outcomes of peer review (Cicchetti, 1991; Daniel, 2004; Bornmann, 2011), the process itself is still strikingly opaque. Journal publishers and editors-in-chief rightly acknowledge that reviewers' independence must be preserved. On the other hand, reviewers tend to receive little training in what is one of the key academic activities, and there is little evidence of any standardization in how review reports are composed. Most importantly, there is little systematic, objective information on the quality of the process across journals and subjects, and on its effect upon the quality of submitted papers.

¹This section re-uses selected paragraphs from Moed (2016d) and Moed (2017a).

Table 12.1 Two phases in the proposed manuscript peer review project

	Phase	Brief description
1	Exploration phase Classical-humanities approach	<ul style="list-style-type: none"> • Development of a conceptual model • Construction of referee report profiles • Analysis of communication modes between actors • Based on well-selected, small data samples
2	Data mining Digital humanities approach	<ul style="list-style-type: none"> • Use of computational linguistics tools • Natural language processing • Statistical analysis • Data mining of large data samples

Some gauge of journal quality is important to authors' decision on where to publish their articles—and from this, to decisions by libraries, research evaluators, appointment panels and so on. With peer review largely a black box, proxies for its quality have grown up, most notably the journal impact factor (JIF) based on citation counts. Its shortcomings are no secret, but its virtues—which include being highly visible, easily accessible, and relatively simple to understand—and the lack of alternatives, maintain its prominence.

This, though, need no longer be the case. The digitization of scientific information offers great potential for the development of tools to allow peer review to be analyzed directly. Computational linguistic analysis and text-mining, combined with more traditional techniques from the humanities, offer the prospect for a better understanding of the process and more insight into differences among disciplines.

This, in turn could lead to tools to help improve peer review for both reviewers and journal editors, and, ultimately, perhaps journal-level metrics that can supersede proxies such as journal impact factors. The analysis consists of two phases, an explorative phase in which a classical-humanities approach is dominant, and a data mining phase, applying techniques from digital humanities. Table 12.1 summarizes main tasks in each phase.

12.1.2 Analyses

At present, this project is at the stage of exploration rather than invention. Researchers are asking publishers to open up their reviews, suitably sampled and anonymized, for analysis. Some are already doing so; others can be expected to be more reluctant. It will be important to cover a range of different disciplines and journals, as reviewing practices vary widely. Once such data are available, one can begin to identify the different elements of a review report.

Taking into account a journal's scope and instructions to reviewers, the various elements of a review report should be analyzed. Statements are categorized in terms of aspect and modality. Standards applied by a reviewer are identified. Relevant concepts are developed, including the formative content of a review report, and the

Table 12.2 Illustrative quotes from referee reports

Nr	Quote
1	“The paper lacks originality as this formula was developed by author A in paper P [...]. Not surprisingly, the empirical findings in the paper, such as Finding F have been found before in several published papers”
2	“I have problems to identify the main purpose of your research [...]. I would suggest to formulate explicit research questions and also explain why you use data from dataset D which do not reflect up-to-date field F very well”
3	“I have severe difficulties with this paper. For me an article should start with defining a hypothesis. Next empirical research is conducted examining its validity, and conclusions are drawn. This paper does not have this structure. It is more exploratory, it discusses a series of properties of database D without explaining their significance”
4	“Although I keep my doubts about the reliability of method M, which I find extremely low, I find the current version interesting and informative, and balanced in its discussion”

distance a reviewer maintains towards his own methodological and theoretical views. Illustrative quotes from existing review reports are listed in Table 12.2.

The first quote in Table 12.2 relates to the manuscript’s originality, one of the key criteria to be assessed in a review report. It clearly states that relevant earlier work has been overlooked. Quote number 2 may refer to the set-up of the research or to the quality of the exposure of the results. It has a formative function, as it suggests how the manuscript can be improved. The referee in quote 3 defines what he/she believes to be a model of a good empirical research paper and applies it as a norm in his review. He states that the manuscript under review does not comply with this norm. Finally, the reviewer in quote 4 maintains a certain distance towards his or her own methodological views and preferences. He uses in his assessment the qualifications ‘informative’ and ‘balanced’.

This approach could reveal the standards that reviewers apply, especially how these general standards are expressed in the content of reviews. One could further develop and validate the hypothesis that review reports that apply vague standards, or fail to apply any assumed key standards, and that contain no reference to the text of the manuscript under review, are less informative and of lower quality than those adopting a series of clear assessment criteria backed up by citing text or tables from the manuscript.

The aim of the project should be to build up for each discipline an understanding of what is considered a reasonable quality threshold for publication, how it differs among journals, and what distinguishes an acceptable paper from one that is rejected. A practical outcome might be that editors set out the assessment criteria for their journals by presenting a list of evaluative statements—anonymous both in terms of reviewer and reviewed author—made by the journal’s referees. It would also be valuable to assess the degree of consensus among referees.

From here, one can begin to develop tentative guidelines and tools for journals and reviewers. This approach, of trying to identify the point at which a paper becomes acceptable for publication, will be more tractable, and useful, than trying to analyze the top end of publishing, where fashion and politics play much larger roles.

12.1.3 *Concluding Remarks*

Ultimately, these analyses might lead to new ways to measure journal quality directly from peer review. Such metrics would need to be complex enough to reflect the peer-review process, but also take into account the qualitative level of submitted manuscripts and authors' publication strategies—they may submit their best papers to the best journals—, and, last but not least, simple enough to be conceptually transparent and allow validation by users.

This is a formidable challenge, but, given the probable demand from many corners of the research world, not an insurmountable one. The ultimate aim of the project, though, should reach beyond comparing journals, to demonstrate the added value of the manuscript referee process, and to further enhance its transparency and efficiency.

12.2 Towards an Ontology-Based Informetric Data Management System²

12.2.1 *Introduction*

Section 1.3 highlighted a series of trends during the past decade in the domain of science policy: an increasing emphasis on societal value and value for money, performance-based funding and on globalization of academic research, and a growing need for internal research assessment and research information systems. Due to the computerization of the research process and the digitization of scholarly communication, research assessment is more and more becoming a 'big data' activity, involving multiple comprehensive citation indexes, electronic full text databases, large publication repositories, usage data from publishers' sites, and altmetric, webometric and other new data sources. The above trends generated an increasing interest in the development, availability and practical application of new indicators for research assessment. Many new indicators were developed, and more and more are available on a large scale. Desktop bibliometrics is becoming a common assessment practice.

These developments pose a series of challenges to the management—collection, handling, integration, analysis and maintenance—of informetric data, and in the design of S&T indicators. Without claiming to be fully comprehensive, the following problems should be mentioned. Many of these were already highlighted in a Special Session on standardization at the ISSI Conference in 1995 in Chicago and published in its proceedings (Glazebrook et al., eds., 1996; Glazebrook, 1996).

²This section re-uses with permission selected paragraphs from Daraio, C., Lenzerini, M., Leporelli, C., Moed, H.F., Naggar, P., Bonacorsi, A., & Bartolucci, A. (2016).

Data-Related Issues

- Data quality issues: completeness, validity, accuracy, consistency, availability and timeliness.
- Comparability problems between databases.
- A lack of standardization, interoperability and modularization of databases.
- Difficulties in the creation of concordance tables among different classification schemes.
- The difficult and costly extension and update of an integrated database.
- A lack of transparency of the database content.
- Difficulty to decompose aggregate indicators.
- Data sources may easily often become data structures coupled to a specific application, rather than application-*independent* databases.
- The data stored in different sources and the processes operating over them tend to be redundant, and mutually inconsistent, mainly because of the lack of central, coherent and unified coordination of data management tasks.

Concept-Related Issues

- Ambiguity of concepts- concepts with the same name may be defined in different manners, or one single operational concept may have different names.
- Informal (non-codified) or ad hoc definitions of indicators are used.
- Indicators used in one particular context (project, time) cannot be used in other contexts.

Maintenance-Related Issues

- Databases may be unavailable and forgotten, though potentially useful.
- Knowledge of how databases were created and how they are structured may get lost; this may make them useless.
- Although the initial design of a collection of data sources and services might be adequate, corrective maintenance actions tend to re-shape them into a form that often diverges from the original conceptual structure.

To solve these issues, it is proposed to develop an ontology-based data management system for research assessment, along the lines set out in Daraio et al., (2016).

12.2.2 An OBDM Approach

Recent studies have explored a promising methodology to deal with these issues, namely an Ontology-Based Data Management approach (OBDM). The notions of

OBDM were introduced by Poggi et al., (2008), Calvanese et al., (2007), Lenzerini, (2011). The approach originated from several disciplines, in particular, Information Integration, Knowledge Representation and Reasoning, and Incomplete and Deductive Databases.

The key idea of OBDM is to create a three-level architecture, constituted by a) the ontology; b) the data sources; and c) the mapping between the two. An *ontology* can be defined as a conceptual, formal description of the domain of interest to a given entity (e.g., organization or community of users), expressed in terms of relevant concepts, *attributes* of concepts, *relationships* between concepts, and *logical* assertions characterizing the domain knowledge. The *sources* are the data repositories accessible by the organization in which data concerning the domain are stored. The mapping is a precise specification of the correspondence between the data contained in the *data sources* on the one hand, and the elements of the *ontology* on the other.

The main advantages of an OBDM approach are as follows

- Users can access the data by using the elements of the ontology. A strict separation exists between the *conceptual* and the *logical-physical* level.
- By making the representation of the domain explicit, the acquired knowledge can be easily re-used.
- The mapping layer explicitly specifies the relationships between the domain concepts in the ontology and the data sources. It is useful also for documentation and standardization purposes.
- The system is more flexible. It is for instance not necessary to merge and integrate all the data sources at once, which could be extremely time consuming and costly.
- The system can be more easily extended. New elements in the ontology or data sources can be added incrementally when they become available. In this sense, the system is *dynamical* and develops over time.

The main purpose of an OBDM System is to allow information consumers to query the data using the elements in the ontology as predicates. It can be seen as a form of information *integration*, where the usual global scheme is replaced by the conceptual model of the application domain, formulated as an ontology expressed in a logic-based language.

12.2.3 Design of Indicators

Table 12.3 distinguishes four dimensions of indicators: their ontological, logical, functional and qualitative dimension. The traditional approach to indicators' design is based on informal definitions expressed in a natural language (English, typically). An indicator is then defined as a relationship between variables, e.g. a ratio between number of publications per academic staff, chosen among a predefined set of data collected and aggregated ad hoc, by a private or a public entity, according to the

Table 12.3 Four dimensions of indicators

Dimension	Specification
Ontological	Formal representation of a domain: objects, their properties and relationships
Logical	Data extracted from sources through mapping considering a query's logical specification
Functional	Mathematical expression to be applied to the results of the logical data extraction
Qualitative	Questions addressed to the ontology for the assessment of the indicator's meaningfulness

user needs, and hence not re-usable for future assessment and use. The OBDM approach permits a more advanced specification of an indicator according to the following indicator dimensions:

The *ontological* dimension represents the domain (portion) of the reality to be measured by the indicator; the *logical* dimension relates to the query applied to the ontological portion in order to retrieve all the information (data) needed for calculating the indicator value. The *functional* dimension indicates the mathematical expression that has to be applied to the result of the logical extraction of data in order to calculate the indicator value. Finally, the *qualitative* dimension specifies the questions that have to be asked to the ontological part in order to generate a list of problems affecting the meaningfulness of the calculated indicator. An indicator will be considered meaningful only if the list of its problems is empty.

12.2.4 Concluding Remarks

Due to the increasing interest in the development and use of research assessment methodologies during the past two decades, the need for more standardization in informetric research and applications need has grown stronger and stronger. Therefore, a systematic exploration of the use of new, ontology-based data management techniques and other approaches from the formal sciences is highly relevant. This does *not* mean that one could expect this approach to solve all standardization problems. A *theoretical* reflection on the concepts used in research assessment is essential as well. The development of an ontology of research assessment could provide a practical framework for such a reflection.

12.3 Towards Informetric Self-Assessment Tools³

12.3.1 Introduction

This section discusses the creation of an informetric self-assessment tool at the level of individual authors or small research groups. The current author defends the position that such an application would be highly useful, but is currently unavailable. Below, it is first argued why such an application is valuable, and what problems it would solve. A next sub-section shows what the tool could look like.

12.3.2 Why an Informetric Self-Assessment Tool Is Useful

Section 9.2 argued that an adequate assessment of individual research performance can take place only on the basis of sufficient background knowledge on the particular role they played in the research presented in their publications, and that calculating indicators at the level of an individual researcher and claiming they measure *by themselves* individual performance suggests a façade of exactness. This does not mean that bibliometric measures in the assessment of individuals are irrelevant. They may still provide relevant information when combined with other types of information, and can especially be useful in *self* assessments.

Despite the critique in the DORA Manifesto (DORA, 2009) published about 8 years ago, journal impact factors are still heavily used by research managers, evaluators and publishers in the assessment of individual researchers and scientific journals. Moreover, an indicator of individual research performance, h-index (Hirsch, 2005) has become a very popular bibliometric measure, despite serious critique on its validity and limited scope of application (e.g., Adler, Ewing & Taylor, 2008). For more information on the h-index, the reader is referred to Sect. 4.3 in *Part II* and Sect. 17.2 in *Part V*.

In their self-assessments, researchers may wish to calculate specific fit-for-purpose indicators that are *not* ‘standard’ and therefore unavailable at the websites of indicator producers. A bibliometric self-assessment tool could become a genuine alternative to using journal impact factors or h-indices in the assessment of research performance of individuals and groups. It could be fully in line with the notion that bibliometrics is not only a helpful evaluation instrument complementary to peer review system, but also for researchers to increase general visibility and optimize their publication strategies (Gorraiz, Wieland, & Gumpenberger, 2016).

It must also be noted that, regardless of their severe limitations, several bibliometric indicators at the author level are widely available. For instance, all three multi-disciplinary citation indexes Web of Science, Scopus and Google Scholar contain h-index values for individual authors. Even if one is against the use of such

³This section re-uses selected paragraphs from Moed (2016e).

indicators in individual assessments, one cannot ignore their availability to a wider public. Therefore, it would be useful if researchers had an online application to check the indicator data calculated about them, and to decompose the indicators' values.

Such a tool would also enable users to learn more about the ins and outs of informetric indicators in general and of the underlying data, and show them for instance how the outcomes of an assessment depend upon the way *benchmark sets* are being defined. The experiences and insights collected in this way would enable researchers subjected to assessment to critically follow assessment processes, and to defend themselves against inaccurate calculation or invalid interpretation of indicators. Distributing knowledge and experiences is also a necessary condition for achieving more standardization of bibliometric concepts and methods.

12.3.3 *What an Informetric Self-Assessment Tool Could Look like*

A challenge is to make optimal use of the potentialities of the current information and communication technologies and create an online application based on key notions expressed decades ago by Eugene Garfield about author benchmarking, and by Robert K. Merton about the formation of a *reference group*. In a first step, the application enables an author to define a set of his publications he/she wishes to take into account in the assessment. It is important that there is a proper data verification tool at hand.

In a next step, a benchmark set is created from researchers with whom the assessed author can best be compared. In his well-known articles about how to use citation analysis in faculty evaluation, Garfield (1983a; 1983b) proposed an algorithm for creating for a given author under assessment a set of 'candidate' benchmark authors who have bibliometric characteristics that are similar to those of the given author. His idea could be further developed by creating a *flexible* benchmarking feature as the practical realization of Merton's notion of a reference group, i.e., the group with which individuals compare themselves, but to which they do not necessarily belong but aspire to (see Holton, 2004).

The calculated indicators should be the result of simple statistical operations on absolute numbers. Not only the outcome, but also the underlying numbers themselves should be visible. In addition, researchers must have the opportunity to decompose and reconstruct an indicator. It should also be possible to insert particular data manually. A typical example is a specification of an author's academic age, as suggested in the ACUMEN portfolio (Bar-Ilan, 2014).

Beyond any doubt, indicators applied in assessment processes must have a sufficiently high level of accuracy, validity and methodological sophistication. And in this respect, much progress has been made during the past decennia. In the proposed tool, the trade-off between methodological sophistication and usability for

large user groups should be in favour of the latter. Sophisticated indicators are particularly useful as *research tools* in testing specific hypotheses in quantitative science and technology studies, but are not necessarily also useful tools for wide user groups. They can be used to validate simplified indicator variants derived from them, thus more easily intelligible and usable by large groups of users.

12.4 Towards Informetric Models of Scientific Development⁴

12.4.1 Introduction

Based on the notion that research and development is a key driver to innovation, growth and economic prosperity, the current section focuses on scientifically developing countries. They need tools to monitor the state of their scientific development and the effectiveness of their research policies in a framework that categorizes national research systems in terms of *the phase* of their scientific development. Analyzing bibliometric indicators at the country level, such a framework enables them to compare ‘like with like’ rather than on the basis of indicator values.

Such numbers are often of limited value and can even be discouraging as they are extracted from global ranking systems which tend to position scientifically developed countries in the *upper part* of a ranking. The framework should also acknowledge that the scientific and economic conditions in which a less-developed country finds itself at a certain moment are not static, but can be viewed as a phase in a process that more scientifically developed countries have already started previously with great success. The current section presents a model of a country’s scientific development using bibliometric indicators based on publications in international, peer-reviewed journals. For a life cycle based theoretical model for analyzing long-term development of research groups and institutes the reader is referred to Braam & van den Besselaar (2014).

12.4.2 A Model of Scientific Development

A simplified and experimental bibliometric model for different phases of development of a national research system distinguishes four phases: (i) a pre-development phase; (ii) building up; (iii) consolidation and expansion; and (iv) internationalization. It is presented in Table 12.4. The model assumes that

⁴This section re-uses with permission selected paragraphs from Halevi & Moed (2014a) and Moed (2016c).

during the various phases of a country's scientific development, the number of published articles in peer-reviewed journals shows a more or less continuous increase, although the rate of increase may vary substantially over the years and between countries. It is the share of a country's internationally co-authored articles that discriminates between the various phases in the development. The model also illustrates the *ambiguity* of this indicator, as a high percentage at a certain point in time may indicate that a country is either in the building up or in the internationalization phase.

Table 12.4 reflects the discontinuities that the model assumes that take place in the indicator values over time when moving from one phase into another.

The distinction into phases is purely *analytical*. The Listing of subsequent phases of development does *not* say anything about the *chronological order* in which they have taken place in a particular country. One would expect that under 'normal' conditions a developing country would subsequently move through the various phases. But if severe changes take place in a country, such as its involvement in a war with another nation for a certain time period, the chronological order of the various phases in that country may deviate from the 'normal' pattern, as it can, for instance, fall from a consolidation phase back into the building-up phase. It must also be noted that international scientific collaboration is not only a matter of development; cultural and historical factors play a role as well (e.g., De Bruin, Braam & Moed, 1991).

Table 12.4 A bibliometric model for capturing the state of a country's scientific development

Phase	Description	Trend in # published articles (*)	Trend in % internationally co-authored publications
Pre-development	Low research activity without clear policy of structural funding of research	~	~
Building up	Collaborations with developed countries are established; national researchers enter international scientific networks	+	++
Consolidation and expansion	The country develops its own infrastructure; the amount of funds available for research increases	++	-
Internationalization	Research institutions in the country start as fully-fledged partners, increasingly take the lead in international collaboration	+	+

(*). ~ denotes: no clear trend; +: increase; -: decline; ++: strong increase. Source UNESCO (2014)

12.4.3 Application to South-East Asian Countries

This sub-section presents an analysis of a series of Asian countries predominantly from South East Asia. The development process that the model seeks to capture takes place during a time span that is much longer than the time period analyzed in this study. Rather than following one particular country through all stages, this study tracks a series of countries during a ten-year period 2003–2012 and draws hypotheses on the current phase of a country's development on the basis of the assertions made in the development model.

Table copied from the UNESCO Report Higher Education in Asia: Expanding Out, Expanding Up. The rise of graduate education and university research (UNESCO, 2014, p.84). (a) ICAP means: Internationally Co-Authored Publications. Trends in a country's percentage share of ICAP are identified on the basis of a linear regression model with the publication year as an independent variable. A trend is labelled positive (negative) if the linear regression coefficient is significantly positive (negative) at the 95% confidence level. (b) Country by income level is based on the World Bank classification 2012. Bibliometric data extracted from Scopus.

Table 12.5 compares a classification of countries according to the model with a categorization based on income. It shows that none of the five high-income countries or territories show a significantly negative trend in international co-authorship, for three the trend is significantly positive. Among the nine countries revealing a significantly negative trend, seven are low-income and two

Table 12.5 Country classifications based on the trend in the percentage of international co-authored publications (ICAP) between 2003 and 2012

Trend in % ICAP (a)	Country or territory by income level (b)		
	High income	Low income	Low or lower-middle income
Positive	<ul style="list-style-type: none"> • Hong Kong (China) • Japan • Singapore 		<ul style="list-style-type: none"> • Cambodia • Nepal • Pakistan
Negative		<ul style="list-style-type: none"> • China • Indonesia • Iran, Islamic Rep. • Malaysia • Philippines • Thailand • Viet Nam 	<ul style="list-style-type: none"> • Bangladesh • Myanmar
No significant trend	<ul style="list-style-type: none"> • Brunei Darussalam • Korea, Rep. 	<ul style="list-style-type: none"> • India • Sri Lanka • Maldives 	<ul style="list-style-type: none"> • Afghanistan • Bhutan • Korea, DPR • Lao PDR

middle-income nations. A decline trend is a sign of the consolidation and expansion phase in scientific development which is apparently dominant in middle-income countries. The fact that low- or lower middle-income countries are spread over the three trend categories is partly due to the low annual publication counts for these countries.

12.4.4 Application to the Persian Gulf Region

Figure 12.1 applies the model to the scientific output of a group of major countries in the Persian Gulf Region and neighboring countries in the Middle East. (Moed, 2016c).

Large differences exist between Iran and Saudi Arabia with respect to the amount of foreign input needed to produce these papers. The percentage of internationally co-authored publications (ICAP) is in 2015 almost 80% for Saudi Arabia, but only around 20% for Iran. While Table 12.5 above indicated for Iran a negative trend in the percentage of ICAP during 2003–2012, Fig. 12.1, relating to the period 2003–2015, shows that this trend is converted into a positive one around the year 2012.

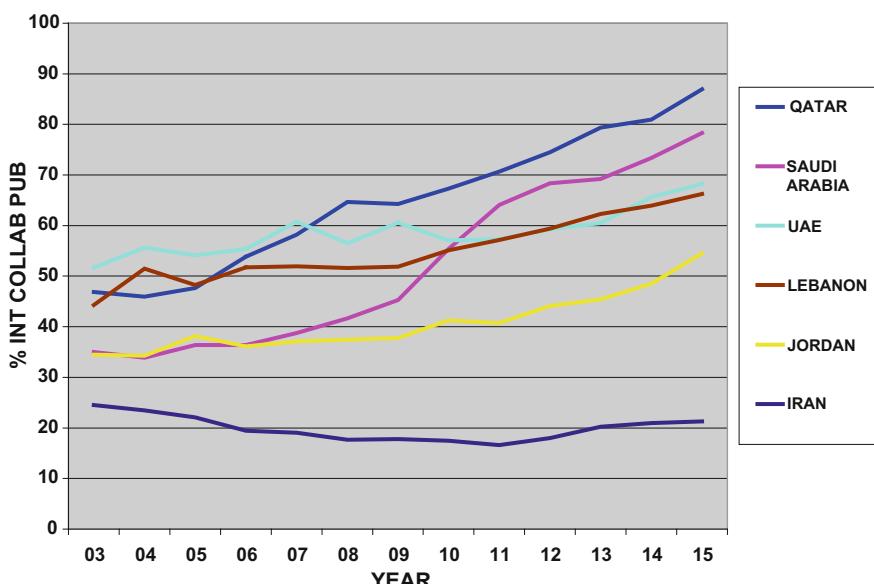


Fig. 12.1 Percentage of internationally co-authored publications (relative to total publication output) of major Gulf States and neighbouring Middle East nations during 2003–2015. Data from Elsevier's Scopus. Figure copied from Moed (2016c)

The results suggest that Iran and Saudi Arabia are in different phases of scientific development. While Saudi Arabia and most other Gulf States are still in the building up phase, Iran is currently moving from a consolidation and expansion into the internationalization phase. This trend can be expected to continue since the international boycotts were cancelled in 2016. The other Gulf States still depend in various degrees upon collaboration with external institutions, increase their international co-authorship rate, and are in a phase of building up a scientific infrastructure. This is especially true for the two countries showing the largest increase of their publication output, namely Qatar and United Arab Emirates.

12.4.5 Concluding Remarks

The model is experimental and needs to be further validated and empirically tested. From a bibliometric point of view, citation impact indicators should be added. It would be especially interesting to take into account an indicator calculated by Scimago.com denoted *as research guarantor* indicator (Moya et al., 2013). It provides for internationally co-authored publications an indication of the role of contributing countries in terms of leading or primary versus following or secondary, based on information of the country in which the *reprint* author is based. The base assumption is that in many research fields obtaining the reprint authorship is seen as a sign of prestige, and that it tends to be gained by the group that made the largest contribution.

Part V

Lectures

Chapter 13

From Derek Price's Network of Scientific Papers to Advanced Science Mapping

Abstract This chapter presents two visionary papers published by Derek de Solla Price, the founding father of the science of science. It presents his view on the scientific literature as a network of scientific papers, and introduces important informetric concepts, including ‘research front’ and ‘immediacy effect’. Next, the chapter shows how his pioneering work on modelling the relational structure of subject space evolved into the creation of a series of currently available, advanced science mapping tools.

Keywords Bibliographic coupling · Citation window · Co-word analysis · Immediacy effect · Information flow · Literature practices · Relational structure · Similarity matrix · Vosviewer

13.1 Networks of Scientific Papers

SLIDE 13.1

Science, 149(3683) : 510-515, July 30, 1965

Networks of Scientific Papers

The pattern of bibliographic references indicates
the nature of the scientific research front.

Derek J. de Solla Price

This article is an attempt to describe in the broadest outline the nature of the total world network of scientific papers. We shall try to picture the network which is obtained by linking each published paper to the other papers

chine-handled citation studies, of large and representative portions of literature, which are much more tractable for such analysis than any topical indexing known to me. It is from such studies, by Garfield (1, 2), Kessler (3), Tukey

This chapter is based on a lecture presented by the author in a doctoral course given in February 2015 at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome.

Slide 13.1 shows the front page of Derek de Solla Price's seminal article Networks of Scientific Papers (Price, 1965). It is based on the idea that "if the paper is an expression of a person or several persons working at the research front, we can tell something about the relations among the people from the papers themselves (Price, 1970, pp. 6-7)".

SLIDE 13.2

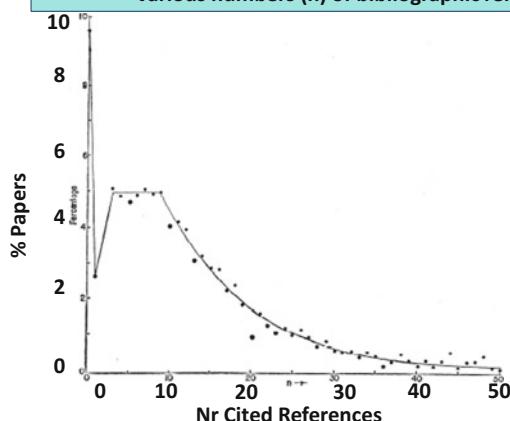
Definition of “reference” and “citation” [Price, 1970, p.3]

- “if Paper R contains a bibliographic footnote using and describing Paper C, then R contains a **reference to C**, and C has a **citation from R**.
- The number of **references** a paper has is measured by the number of items **in its bibliography** as endnotes and footnotes, etc.,
- ...while the number of **citations** a paper has is found by looking it up on some sort of **citation index** and seeing how many other papers mention it”.

In bibliometrics there is often confusion about the difference between the terms citation and reference. Slide 13.2 gives Price's definitions of these two terms (obtained from Price, 1970, p. 3)

SLIDE 13.3

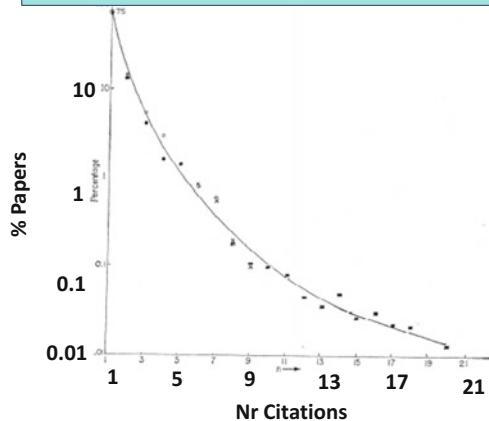
Fig. 1. Percentages of papers published in 1961 which contain various numbers (n) of bibliographic references



Slide 13.3 presents Figure 1 from Price's article Networks of Scientific Papers (Price, 1965)¹. For each source document (also denoted as paper) processed for the Science Citation Index in 1961 the number of *cited references* is determined. The figure presents the distribution of this number across papers. 9.5% of papers has no cited references. These are mostly meeting abstracts.

SLIDE 13.4

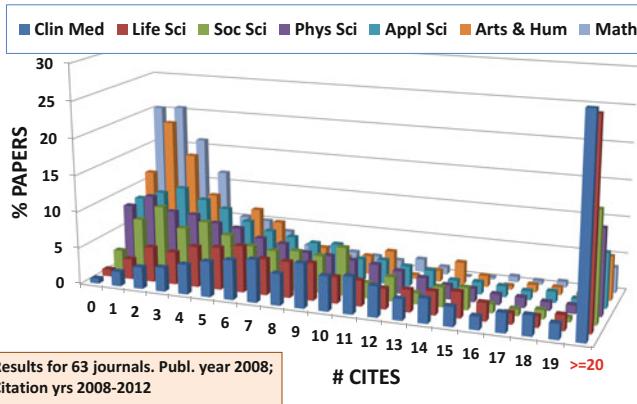
Fig. 2: Percentages (relative to total number of cited papers) of papers cited various numbers (n) of times, for a single year 1961



Slide 13.4 presents Figure 2 from Price's article². It shows the distribution of the number of received *citations* across all cited documents in the 1961 SCI. Uncited docs are not considered. The vertical axis has a logarithmic scale. The distribution is highly skewed. 75% of (cited) papers are cited only once.

SLIDE 13.5

Differences in citation distributions between subject fields



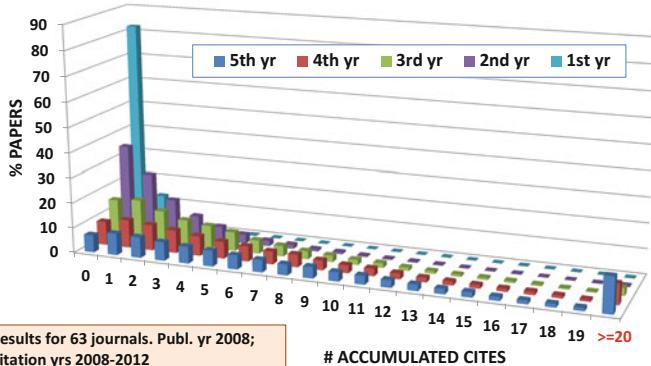
¹Reprinted with permission from Price (1965, p. 511).

²Reprinted with permission from Price (1965, p. 511).

Slides 13.5 and 13.6 are not taken from Price's work, but were created by the current author from a dataset of 63 journals listed in Slide 13.11 below. Slide 13.5 shows that the shape of the citation distribution varies substantially between subject fields. In Clinical Medicine and Life Sciences papers tend to be cited more often than in Arts & Humanities and Mathematics. Data is extracted from Scopus.

SLIDE 13.6

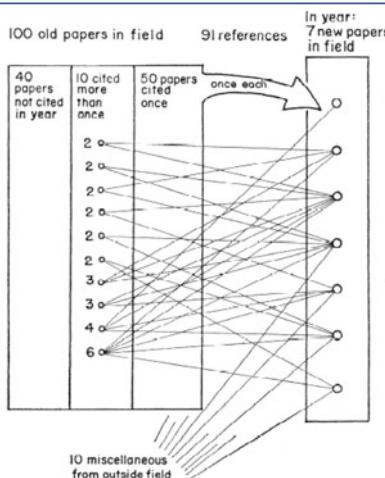
Effect of length of citation time window upon citation distributions



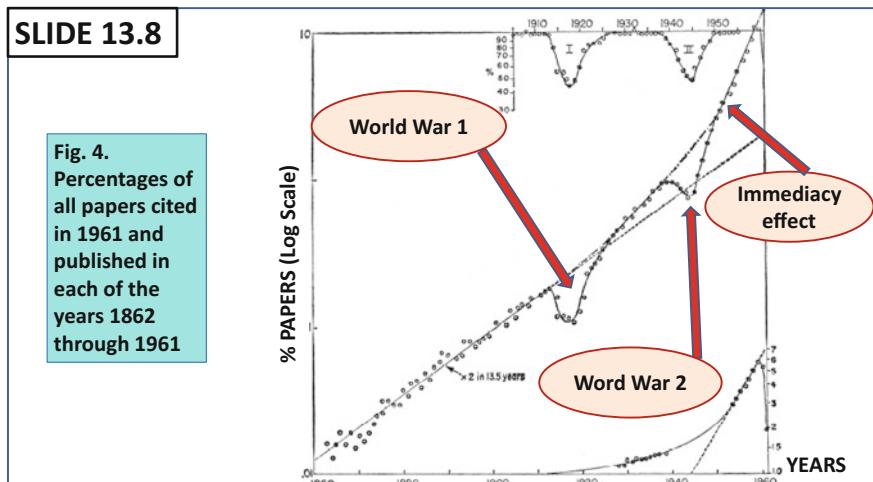
Slide 13.6 illustrates that the shape of the citation distribution not only depends upon the discipline, but also upon the citation 'window' that is applied. The distribution at the inner side of the graph is based on citation counts obtained in the year in which the articles were published, and that at the front, on counts during the first *five* years. The former is much more skewed than the latter.

SLIDE 13.7

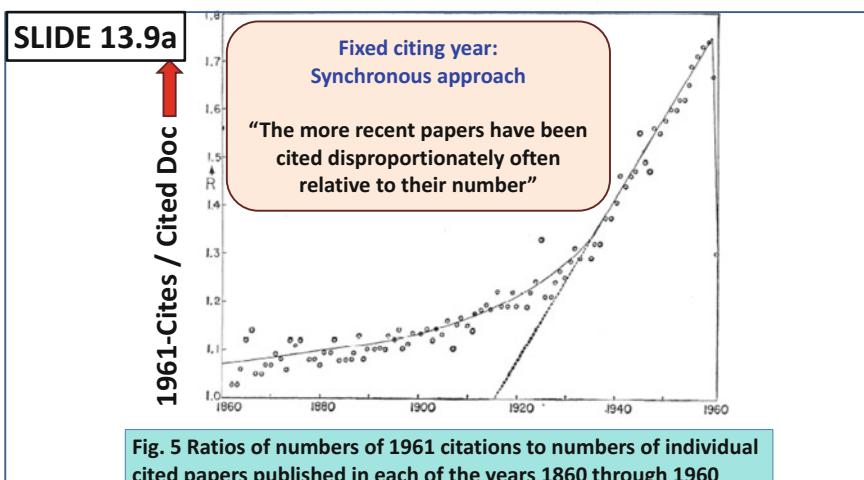
Fig. 3. Idealized representation of the balance of papers and citations for a given "almost closed" field in a single year.



Slide 13.7. In this figure Price illustrates how papers from various years are connected³ He writes: "Since only a small part of the earlier literature is knitted together by the new year's crop of papers, we may look upon this small part as a sort of growing tip or epidermal layer, an active research front" (Price, 1965, p. 512).



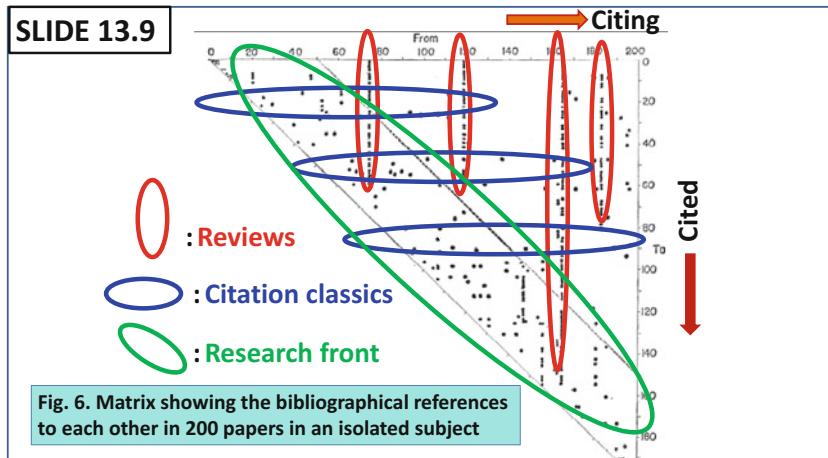
Slide 13.8. This figure is among the most famous graphs ever produced in bibliometrics⁴. It shows the age distribution of references cited in source articles in the SCI of 1961. There are clear traces of the two world wars.



³Reprinted with permission from Price, (1965, p. 512).

⁴Reprinted with permission from Price, (1965, p. 513).

Slide 13.9a. While Slide 13.8 displays on the vertical axes the percentage of cited papers (on a log scale), Slide 13.9a gives the average number of citations *per cited document*⁵. As indicated in the quote, the more recent papers are cited disproportionately relatively to their number. This phenomenon is termed as the *immediacy effect*, and it reflects the *research front*.



Slide 13.9. Price's Figure 6 is also a famous graph⁶. It presents citation relations between papers in a small, closed research field. Each dot represents a citation in a citing document of a particular year to a cited document. Review articles containing a large number of cited references are clearly visible. This is also true for the often cited citation 'classics', and for the research front.

SLIDE 13.10

Two different literature practices and information needs [Price, 1965]

Type of subject	Citation pattern	Information need
Research front subjects	The research front builds on recent work, and the network becomes very tight	Alerting service that will keep him posted, probably by citation indexing, on the work of his peers
Taxonomic subjects	The random scattering corresponds to a drawing upon the totality of previous work	Systematize the added knowledge from time to time in book form, topic by topic, or make use of a system of classification

⁵Reprinted with permission from Price, (1965, p. 514).

⁶Reprinted with permission from Price, (1965, p. 514).

Slide 13.10. Price distinguishes two different types of research fields with different literature practices and information needs. *Research front subjects* reveal a concentration on recent work and dense networks, and need citation indexing. *Taxonomic subjects* can be found especially in humanities. They show more scattering, publish books and need a subject classification system.

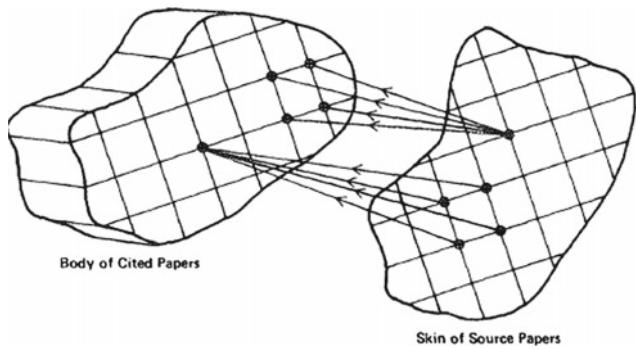
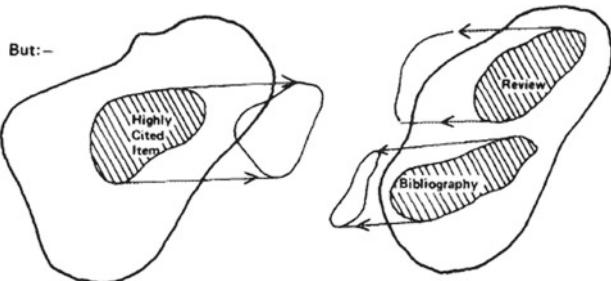
SLIDE 13.11	63 Journal Set
<ul style="list-style-type: none"> • Annals of Pure and Applied Logic • Applied Clay Science • Applied Surface Science • Biochimica et Biophysica Acta - Bioenergetics • Biochimica et Biophysica Acta - Biomembranes • Biochimica et Biophysica Acta - Gene Regulatory Mechanisms • Biochimica et Biophysica Acta - General Subjects • Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids • Biochimica et Biophysica Acta - Molecular Basis of Disease • Biochimica et Biophysica Acta - Molecular Cell Research • Biochimica et Biophysica Acta - Proteins and Proteomics • Biochimica et Biophysica Acta - Reviews on Cancer • Bioinorganic Chemistry • Cancer Letters • Differential Geometry and its Application • Earth and Planetary Sciences Letters • European Journal of Cancer • Fuzzy Sets and Systems • Journal of Applied Geophysics • Journal of Cultural Heritage • Journal of Dentistry • Journal of Econometrics • Journal of Economics and Business • Journal of Hydrology • Journal of Informetrics • Journal of International Economics • Journal of Logic and Algebraic Programming • Journal of Medieval History • Journal of Wind Engineering and Industrial Aerodynamics • Limnologica • Lingua • Materials Science & Engineering A: Structural Materials: Properties, Microstructure and Processing • Materials Science & Engineering B: Solid-State Materials for Advanced Technology • Materials Science and Engineering C 	<ul style="list-style-type: none"> • Molecular Oncology • Ophthalmology • Performance Evaluation • Physica A: Statistical Mechanics and its Applications • Physica B: Condensed Matter • Physica C: Superconductivity and its Applications • Physica D: Nonlinear Phenomena • Physica E: Low-Dimensional Systems and Nanostructures • Phytochemistry • Phytochemistry Letters • Plant Physiology and Biochemistry • Plant Science • Poetics • Powder Technology • Stem Cell Research • Surface Science • Tectonophysics • Tetrahedron Letters • Thin Solid Films • Topology and its Applications • Trends in Plant Science • Water Research • Journal of Science and Medicine in Sport • Applied Ergonomics • design studies • Journal of Historical Geography • Journal of Phonetics • Child Abuse and Neglect • Behavior Therapy

Slide 13.11 lists the set of journals analyzed in Slides 13.5 and 13.6 above. They are all published by Elsevier. The citation data on these journals were extracted from Scopus.

13.2 Modelling the Relational Structure of Subject Space

Slide 13.12 visualizes the citation process. The term source paper indicates a document that is indexed for the Science Citation Index (SCI) by the Institute for Scientific Information (ISI). All cited references in a source paper are included in the citation index (Price, 1980, p. 630)⁷.

⁷Slides 13.12–13.14 are copied from a reprint of Price's paper published in Garfield, E., Essays of an Information Scientist, vol. 4, pp. 621-633, 1979/1980. The page numbers relate to this version.

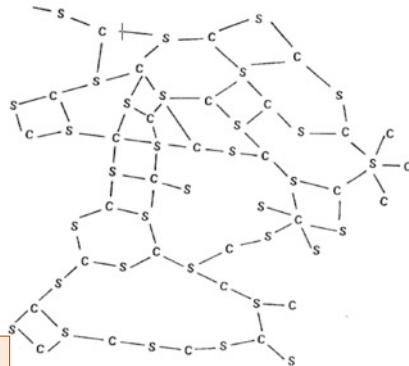
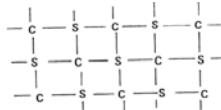
SLIDE 13.12**From: Price, The Citation Cycle****SLIDE 13.13****From: Price, The Citation Cycle**

Slide 13.13 gives a reminder of a phenomenon already described in Slides 13.3, 13.4 and 13.9, namely that some documents (reviews) contain many cited references, and some (highly cited items, citation classics) are cited often (Price, 1980, p. 630).

SLIDE 13.14

If four is only a statistical mean, the corresponding lattice with various numbers of links would look rather like a very torn and deformable fishing net

if there were exactly four links per item, the result would be a perfect lattice

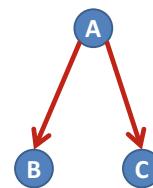


From: Price, The Citation Cycle

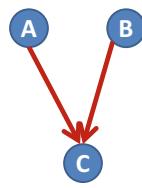
In Slide 13.14 Price visualizes the citation relationship between documents. The structure caused by these relationships is represented as a chemical structure. S means source document, and C cited document. The next slides present a series of maps unravelling the structure of science (Price, 1980, p.631).

SLIDE 13.15**Three citation relationships**

A cites B
B is a cited reference in A



B and C are co-cited by A

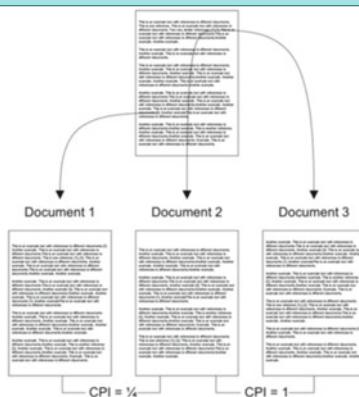


A and B are bibliographically coupled via C

But first, Slide 13.15 illustrates three citation relationships: citation, co-citation and bibliographic coupling. The co-citation technique was introduced by Irina Marshakova and Henry Small, independently from each other, in 1973. Bibliographic coupling was invented by Kessler in 1963.

SLIDE 13.16

Co-citation proximity Index (CPI): co-citation strength depends upon placement of citations relative to each other



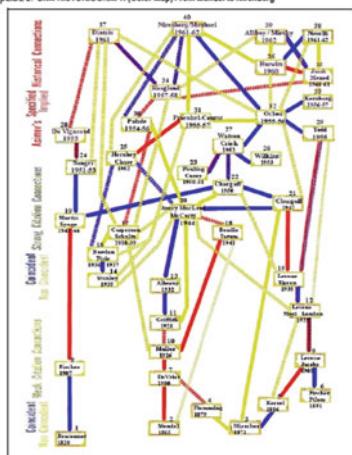
Source: Wikipedia

Slide 13.16⁸ illustrates a recent refinement of co-citation analysis that expresses the strength of a co-citation relationship between two co-cited documents on the basis of the ‘distance’ between the two citations in the citing text. The link between documents 2 and 3 is stronger than that between documents 1 and 2, or between 1 and 3, because they were co-cited in the same *segment* of the full text (perhaps even in the same sentence). The idea to define co-citation strength between documents in this way was introduced by Gipp & Beel (2009) and further explored by Boyack, Small & Klavans (2013).

SLIDE 13.17

SLIDE 2: DNA HISTORIOGRAPH (Color Map) From Mendel to Nirenberg

40 key milestone events in the history of DNA from Mendel to Nirenberg, 1962



Source: Garfield, Algorithmic Historio-bibliography, 2001.

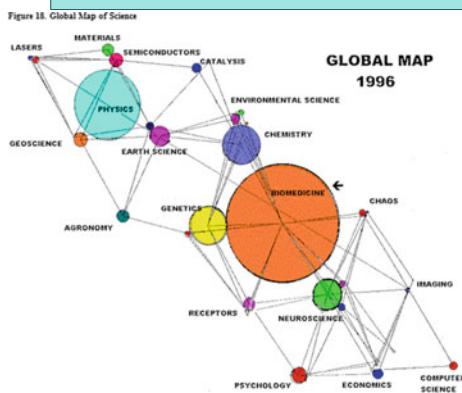
<http://garfield.library.upenn.edu/papers/drexelbelvergrifffith92001.pdf>

⁸Reprinted with permission from Gipp & Beel (2009).

Slide 13.17 presents one of the very first science maps, created manually by Eugene Garfield and co-workers, based on a mini citation index of 65 key papers in DNA research. The figure is copied from a lecture presented by Eugene Garfield (Garfield, 2001). Each box represents one or more key papers. The dark blue lines indicate that there is a direct citation link between the nodes involved. The dotted lines represent implicit links. The red lines indicate relationships mentioned in a popular text book on the emergence of DNA research.

SLIDE 13.18

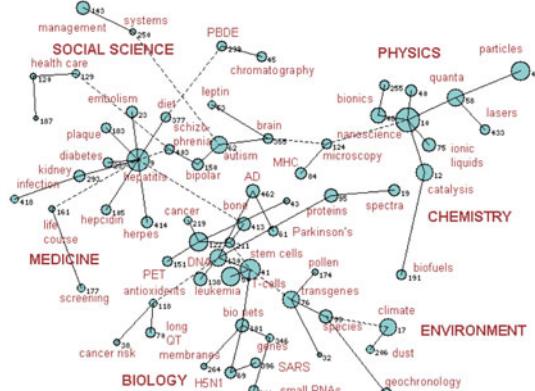
Global Map of Science, 1996.
Co-citation clustering using ISI data (Henry Small).
Source: <http://www.garfield.library.upenn.edu/papers/mapsciworld.html>



Slide 13.18 presents one of the first comprehensive, global maps of science, based on co-citation analysis. The figure is copied from a lecture presented by Eugene Garfield (Garfield, 1998). The diameter of a discipline's circle indicates the number of papers in the cluster, and the lines between clusters the strength of the relationship between disciplines.

SLIDE 13.19

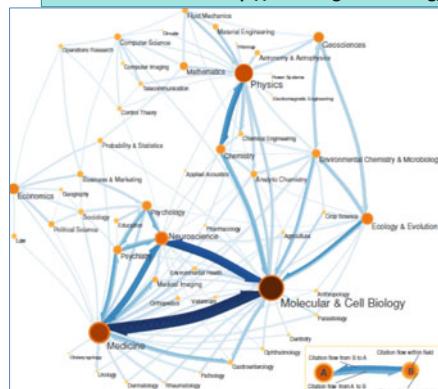
Co-citation based research fronts (2004-2009)
Based on Web of Science; Source: ScienceWatch, 2010



Slide 13.19 presents a more recent global map of science based on *co-citation analysis*⁹. A co-citation cluster named *brain* establishes a (weak) link between physics on the one hand, and social science and medicine on the other.

SLIDE 13.20

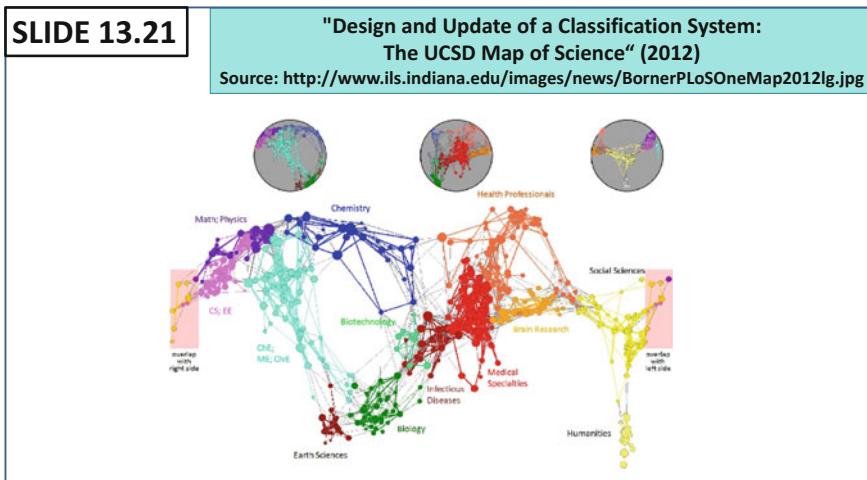
Using Journal Citation Reports (JCR) data, 6,128 journals connected by 6,434,916 citations were partitioned into 88 modules (2004)
Source: <http://www.eigenfactor.org/map/Sci2004.pdf>



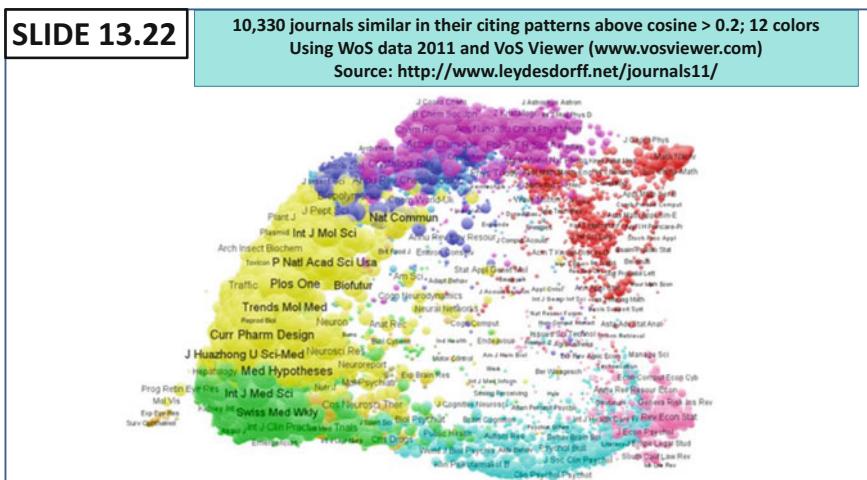
The map presented in Slide 13.20 is based on an *information flow* method for mapping large networks. Reprinted with permission from Rosscall, M. & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. PNAS, 105, 1118–1123. Copyright (2008) National Academy of Sciences, U.S.A. Orange circles represent subject fields, with larger, darker circles

⁹Source: Web of Science, Essential Science Indicators. Used by permission of Clarivate Analytics.

indicating larger field size. Blue arrows represent citation flow between fields. An arrow from field A to field B indicates citation traffic from A to B, with larger, darker arrows indicating higher citation volume.

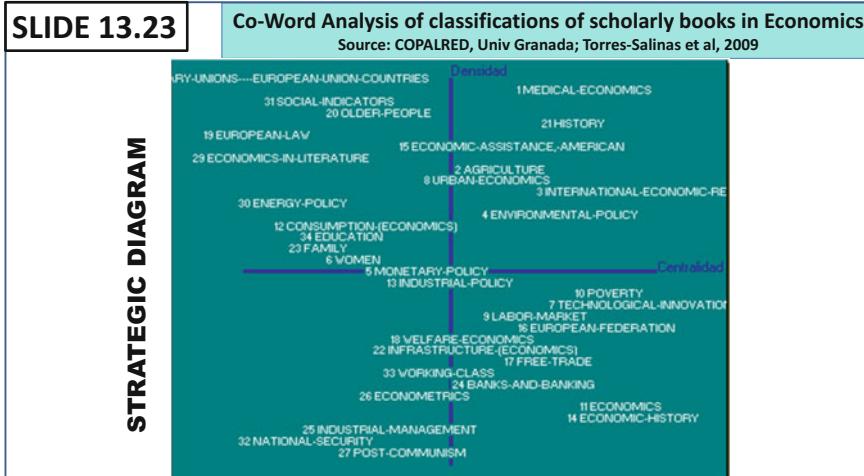


Slide 13.21¹⁰ presents a map on a three-dimensional surface that is constructed as a world map. The circles are *co-citation* clusters. It suggests that, compared to the situation in Slide 13.19, the role of brain research has become stronger in recent years. There is a link between social science and physics, but this map does not give the name of the cluster that is responsible for it.



¹⁰Reprinted with permission from Borner, Klavans, Patek, Zoss, Biberstine, et al. (2012).

Slide 13.22 is based on an analysis of the way in which *journals cite one another*¹¹. A *similarity* measure between two journals is based on the extent to which they cite the same journals: the more cited journals they have in common, the more similar they are. The resulting similarity matrix is analyzed and visualized using a special software named VosViewer (see Slide 13.24 below).



The map presented in slide 13.23 is *not* based on citation analysis. It is a so called *co-word map*¹². Co-word analysis identifies significant terms from a set of documents, for instance, cognitive terms extracted from their full texts, or indexing terms such as keywords or content classification codes. See the seminal paper by Callon, Courtial, Turner, & Bauin (1983), and Callon, Law, & Rip (1986) for technical details and a theoretical background. Similarity between two documents is based on the number of terms they have in common. Clusters are formed, and categorized in terms of their density (the strength between terms within a cluster) and centrality (based on the strength of the links between clusters). The map in Slide 13.23 analyzes classification codes assigned by libraries to books in the field of economics (Torres-Salinas & Moed, 2009). For a combination of co-citation and co-word analysis the reader is referred to Braam, Moed & van Raan (1991).

¹¹Reprinted with permission from Leydesdorff, Rafols & Chen (2013). It can also be web-started at:

http://www.vosviewer.com/vosviewer.php?map=http://www.leydesdorff.net/journals11/citing_all.txt.

¹²Reprinted with permission from Torres-Salinas & Moed (2009).

13.3 Mapping Software

SLIDE 13.24**Four science mapping softwares**

Name	Website
Vosviewer – Visualising scientific landscapes	www.vosviewer.com
SCIMAT – Science mapping analysis tool	http://sci2s.ugr.es/scimat/
SCI2 – A tool for science of science research and practice	https://sci2.cns.iu.edu/user/index.php
Pajek – Program for large network analysis	http://pajek.imfm.si/doku.php?id=pajek

Slide 13.24 gives information on four often used mapping software packages. Obviously, there are many more software packages, but these four can be found relatively often in the informetric literature. All four were developed in an academic environment, namely in Leiden University (the Netherlands), University of Granada (Spain), Indiana University (USA), and University of Ljubljana (Slovenia), respectively. All tools model the relational structure of subject space, inspired by Derek de Solla Price's visionary papers.

Chapter 14

From Eugene Garfield's Citation Index to Scopus and Google Scholar

Abstract This presents a comparative analysis of three large, multi-disciplinary citation indexes. It starts with a presentation of the basic principles of the Science Citation Index (SCI, later Thomson Reuters' Web of Science), a multi-disciplinary citation index created by Eugene Garfield in the early 1960s. Next, it presents a study conducted in 2009 comparing the Web of Science with Scopus, a comprehensive citation index launched by Elsevier in 2004, and a recent study comparing Scopus with an even more comprehensive citation index, Google Scholar, also launched in 2004.

Keywords ArXiv.org · Bradford's Law · Content advisory board · Coverage analysis · Garfield's Law · Indexing speed · Literature retrieval · Source coverage · Ulrichsweb

14.1 Science Citation Index and Web of Science

SLIDE 14.1

Eugene Garfield's base idea

"If the **literature of science** reflects the activities of science, a comprehensive, multidisciplinary **citation index** can provide an interesting view of these activities.

This view can shed some useful light on both the **structure** of science and the **process** of scientific development" (Garfield, 1979, p. 62).

This chapter is based on a lecture presented by the author in a doctoral course given in February 2015 at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome.

Slide 14.1 reveals the agreement as well as the complementarity between Eugene Garfield and Derek de Solla Price's ideas. They both understood that one can study the structure and the process of science by analyzing the scientific literature. Eugene Garfield was instrumental to this notion as the builder of one of the key tools to conduct such analysis: a comprehensive citation index.

SLIDE 14.2

How many scientific/scholarly journals are there?

- There is no agreement on what constitutes a journal (Garfield)
- Derek de Solla Price (1980): 40,000
- Garfield (1979): 10,000
- Scopus: Covers about 19,000 journals in 2014
- Web of Science covers 11,000 journals in 2014

Slides 14.2 and 14.3 reveal that there has always been a debate as to how many scientific-scholarly journals are being published. Of course, there are changes over time. But, as Garfield has pointed out, perhaps the most critical issue is how one defines the concept of journal.

SLIDE 14.3

Journals in Ulrichsweb (6 Febr 2015)

Collection	Number	Percentage
Active, academic/scholarly journals	111,770	100 %
Peer-reviewed/refereed	66,734	60 %
Available online	47,826	43 %
Open access	15,025	13 %
Included in Thomson-Reuters Journal Citation Reports (JCR)	10,916	10 %

Slide 14.3 gives figures extracted from Ulrichs' journal database Ulrichsweb on 6 February 2015. Apart from the issue as to how to define a journal, the definition of the concepts *academic* or *scholarly* is highly relevant as well. The term 'active' means that journals do publish papers in the base year.

SLIDE 14.4**Bradford's Law of Dispersion**

"Articles of interest to a specialist must not only occur in periodicals specializing in his subject [core] but also in other periodicals, which grow in number as the relation of their fields to that of the subject lessens, and the number of articles on his subject in each periodical diminishes [tail]"

[Bradford, S.C., Documentation, 2nd ed. (London: Lockwood, 1953)]

Slide 14.4 gives a quote of an important law in library science: Bradford's Law of Dispersion or Scattering, describing how the relevant documents in a subject are statistically distributed among publishing sources (journals).

SLIDE 14.5**Bradford's Law of Dispersion:
Rules of thumb**

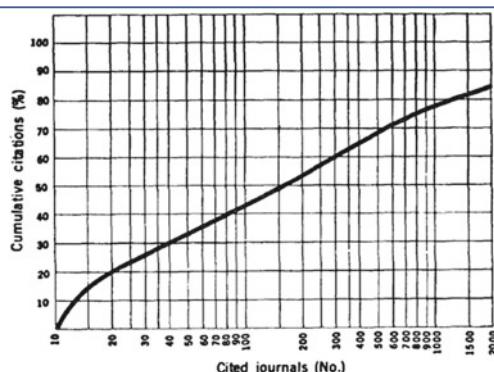
- A **core** of 20 % of journals contain 80 % of relevant documents
- 500-1000 journals cover 95 % of significant literature in a field

Slide 14.5 presents a rule of thumb that librarians have developed based on Bradford's Law. One might argue that, if 500–1000 journals cover the significant literature in a field, and if one assumes that there are, say, 100 fields, the total number of journals needed to cover all science and scholarship might be half a million or more. But the next slide shows that this assertion is incorrect.

SLIDE 14.6**Garfield's Law of Concentration**

- “The tail of the literature in one discipline consists, in a large part, of the cores of the literature of other disciplines” (Garfield, 1979)
- “The core literature for all scientific disciplines involves a group of no more than 1,000 journals, and may involve as few as 500” (Garfield, 1979)

Slide 14.6 explains why this assertion is incorrect. It does not take into account *Garfield's Law of Concentration*, stating that the tail of the literature in one field consists of core journals from other fields. Garfield also claimed that the number of core journals in the domain of all science and scholarship is between 500 and 1000.

SLIDE 14.7

“Fig. 5. Distribution of citations among cited journals. The curve shows that a relatively small core of 152 journals accounts for about half of all citations and that only 2000 or so journals account for 84 percent of all citations” (Garfield, 1972, p 535)

Slide 14.7 is copied from Garfield (1972), with permission from AAAS. This graph has played an important role in later discussions on the adequacy of coverage of the Science Citation Index. It indicates that about 2000 journals account for 84% of citations made in sources indexed for the SCI. Two key issues have been addressed: how does this percentage vary with expanding source coverage? After all, the 84% estimate is not based on an analysis of citation in all journals published in the world, but only in those indexed for the SCI. Secondly: How does it vary between disciplines?

SLIDE 14.8**Basic coverage criterion**

- “Because the problem of coverage is one of practical economics, the criterion for what is covered is **cost effectiveness**” (Garfield, 1979)
- “A cost-effective index must restrict its coverage, as nearly as possible, to only those items that people are likely to find **useful**” (Garfield, 1979)

The emphasis on cost effectiveness in Slide 14.8 is partially due to the relatively large costs for electronic data storage and for manual meta-data entry from indexed sources when Garfield started creating the SCI. Nowadays, these two activities are cheaper.

SLIDE 14.9**Selection of source journals**

The real problem is to “make the coverage as complete as possible by expanding it **beyond the core** of journals whose importance to a given field is obvious” (Garfield 1979)

Slide 14.9. Garfield argued that in each field of scholarship its practitioners can easily identify the most important journals, publishing the highest quality materials, but that for a citation indexer the real challenge is to go a step further and select a next group of important journals.

SLIDE 14.10**How is the core expanded?**

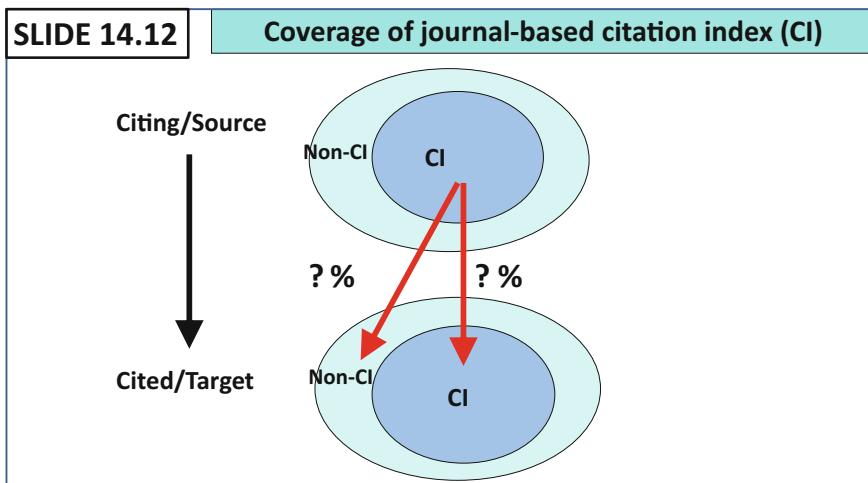
- **Garfield's criterion:** The frequency at which journals are cited in the source journals that are already included in the index
- **Assumption:** The number of times a journal's items are cited is an expression of its utility as communication medium
- **Indicator:** Journal impact factor

Slide 14.10 reveals how the source coverage is expanded. The method that Garfield developed is the key element in the creation of a comprehensive citation index. Journals are added that are most frequently cited in the set of source journals already included in the index. The journal impact factor was developed in this context. This indicator was used in combination with expert knowledge from a content advisory board of active researchers.

SLIDE 14.11**Coverage analysis of references**

- Source (citing) year: 2002
- References to items published <1980 were removed
- Total references: about 20 million
- Cited journal items: rough estimate

Slide 14.11. The graph presented above in Slide 14.7 revealed that 2000 journals accounted for 84% of all citations indexed in the SCI. Slides 14.12 and 14.3 extend this analysis and assess the coverage of the ISI Citation Indexes in the year 2002 by discipline. Slide 14.11 present some of the technical details. The set of ISI Indexes contained the Science Citation Index (SCI), Social Science Citation Index (SSCI) and Arts & Humanities Citation Index (AHCI).



Slide 14.12 shows how the coverage of the three ISI indexes was assessed. The method takes into account all about 20 million cited references published as from 1980 onwards and included in source journals indexed in 2002. The question is: to which extent are these cited references themselves published in journals processed for the index? It was found that this coverage percentage amounted to 75% for the three indexes combined, and 81% for the SCI.

SLIDE 14.13

CI coverage by field

Journals		Books, proceedings	
EXCELLENT (>80%)	GOOD (60-80%)	FAIR (40-60%)	MODERATE (<40%)
Biochem & Mol Biol	Appl Phys & Chem	Mathematics	Other Soc Sci
Biol Sci — Humans	Biol Sci — Anim & Plants	Economics	Humanities & Arts
Chemistry	Psychol & Psychiat	Engineering	
Clin Medicine	Geosciences		
Phys & Astron	Soc Sci ~ Medicine		

Slide 14.13 shows that large differences in ISI coverage exist between scholarly disciplines. It is above 80% in biochemistry & molecular biology, biological sciences related to humans, chemistry, clinical medicine and physics & astronomy, but below 40% in sociology, political science, educational science and other social sciences, and in the humanities and arts. More results are presented in Moed (2005a, Chap. 7).

SLIDE 14.14**A partial view**

- Analysis relates to ISI citation indexes on CD-ROM in 2002!
Updating and expanding is necessary
- Analysis is based on references in (WoS covered) journal articles only
- Authors publishing in a particular journal tend to cite that journal more frequently than they cite it in their papers published in other journals
- Analyses of references in non-covered journals, books and other non-journal sources would provide a more complete picture
- This type of coverage data is not available for Scopus and Google Scholar

The coverage analysis provides a partial view, and Slide 14.14 shows some of its limitations. It must be underlined that the data relate to 2002; an updating would be interesting, so that the trend in coverage over the years can be analyzed as well. Also, it would be of great interest to conduct and publish this type of analysis also for Scopus and Google Scholar.

14.2 Scopus Versus Web of Science

SLIDE 14.15**Scopus is a genuine alternative to WoS**
[CWTS, Study for HEFCE, 2007]

- Scopus tends to include all science journals covered by the WoS (papers>=1996)
- And Scopus contains some 40 % more papers
- Scopus is larger and broader in terms of subject and geographical coverage
- Web of Science is more selective in terms of citation impact

A study conducted in 2007 compared the source coverage of Thomson Reuters' Web of Science (WoS, the continuation of the ISI citation indexes) on a paper-by-paper basis, and found that in science fields almost all WoS source journals were indexed in Scopus. But Scopus indexed many journals *not* covered in WoS. All in all, Scopus contained in 2007 some 40% more source articles than WoS.

SLIDE 14.16**Scopus journals not covered in WoS tend to show
(SCImago-CWTS case study on oncology):**

- Lower citation rates
- More dispersion among publishers countries
- More in non-English languages
- More recently founded between 1996-2006
- More often freely available online
- More nationally oriented

Slide 14.16 summarizes the characteristics of the Scopus journals that are *not* covered by WoS, based on a case study conducted in 2008 in the field of oncology (Lopez-Illescas et al., 2008).

SLIDE 14.17**What is the overlap in coverage of oncological journals
between WoS and Scopus? [Lopez-Illescas et al., 2008]**

- All 126 WoS oncology journals are in Scopus (112 in Scopus Cancer categories)
- About 50 % of Scopus oncology journals is not in WoS (106 out of 231)

Slide 4.17 compares Web of Science and Scopus in one particular subject field: Oncology (or Cancer Research). The conclusions obtained in this study are similar to those obtained by Gorraiz & Schloegl (2008) in their analysis of another discipline: pharmacology and pharmacy.

SLIDE 14.18

Scopus cancer journals not in WoS have lower impact factors than WoS covered journals do [Lopez-Illescas et al., 2008]

Quartile (based on Scopus impact factors)	No. (%) journals in Scopus (n=206)	Impact factor range (2006)	No. (%) of Scopus journals in WoS (n=112)
4 (bottom)	51 (25%)	0-0.5	3 (6%)
3	52 (25%)	0.5-1.8	25 (48%)
2	52 (25%)	1.8-3.3	36 (69%)
1 (top)	51 (25%)	3.3-63.0	48 (94%)

Slide 14.18 analyzes the citation impact of the various oncological journals in the study. It shows that among the 25% of journals with the *lowest* journal impact factor in Scopus, only 6% is indexed in WoS. But among the 25% of journals with the *highest* citation impact in Scopus, 94% were covered in WoS.

SLIDE 14.19

Comments on comparison Scopus-Web of Science

- Both databases have an Advisory Content Selection Board
- Scopus and Web of Science coverage change significantly over time
- Both databases added more conference proceedings
- Both databases added a Book Citation Index
- Garfield's citation-based coverage criterion is not the only coverage criterion

Slide 14.19 gives more information about Web of Science and Scopus. It must be underlined that both indexes made significant changes in their source coverage during the past decade.

14.3 Google Scholar Versus Scopus

SLIDE 14.20

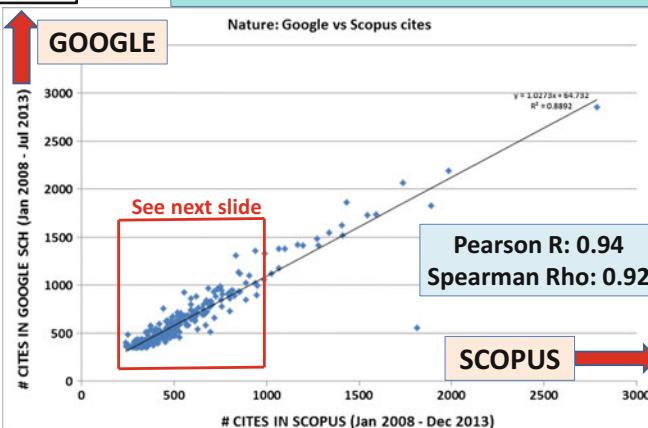
Scopus vs. Google Scholar: Citation counts to individual articles in selected journals

- Google Scholar publishes for a large number of journals the **H index** (for the time period 2009-2013, label: **H5**)
- It also publishes citation counts of individual articles (only “top” articles with **score > H5**)
- Research Question:
- **How do the article citation counts in Google Scholar compare to those in Scopus?**

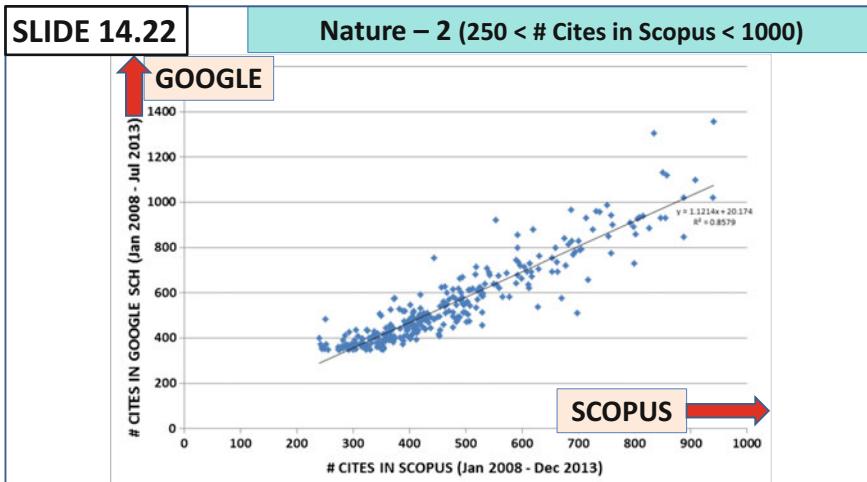
Slide 14.20 formulates one of the research questions in a comparative analysis of Scopus and Google Scholar published by Moed, Bar-Ilan and Halevi (2016).

SLIDE 14.21

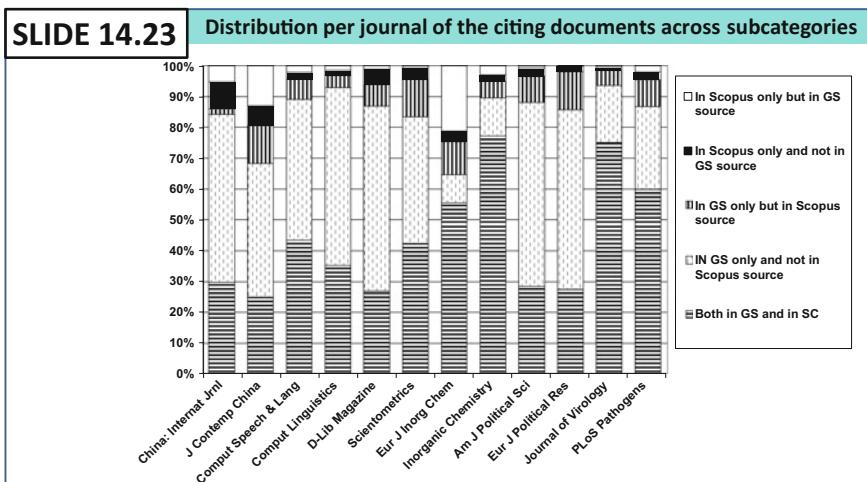
Google Scholar vs. Scopus: Nature -1



A scatter plot presented in Slide 14.21 compares the citation count in Scopus with that generated in Google Scholar for a set of articles published in the journal Nature in 2008. The linear and rank correlation coefficients between the two counts are both above 0.9.



Slide 14.22 zooms in on a particular subset of papers presented in Slide 14.21, namely the papers with a citation count in Scopus below 1000. The Pearson correlation coefficient between Scopus and Google Scholar citation counts in this data segment is 0.86.



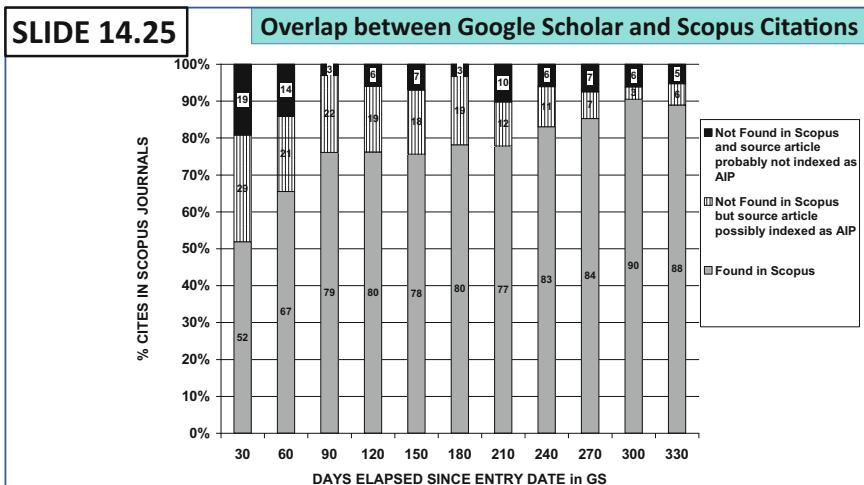
The study mentioned above in Slide 14.20 conducted a detailed analysis of 12 journals from 6 subject fields.¹ Their titles are listed below the horizontal axes in Slide 14.23. This slide analyses per journal the overlap between the *citing documents* retrieved from Scopus and those extracted from Google Scholar. It shows for instance that this overlap is much larger in the virology and chemistry journals than it is in political science and the other social science or humanities journals.

SLIDE 14.24**Google Scholar vs. Scopus (Moed, Bar-Ilan & Halevi, 2016):
Conclusions**

- Exploratory, hypothesis-generating study comparing Google Scholar and Scopus for 12 journals in 6 subject fields
- The ratio of GS over Scopus citation counts varies across fields between 1.0 and 4.0
- OA journals in the sample show higher GS/Scopus ratios than non-OA outlets
- The linear correlation between GS and Scopus citation counts at the article level is high: Pearson's R is in the range of 0.8-0.9.
- A median Scopus indexing delay of two months compared to GS is largely though not exclusively due to missing cited references in articles in press in Scopus.
- The effect of double citation counts in GS due to multiple citations with identical or substantially similar meta-data occurs in less than 2 per cent of cases.
- Several data quality issues were found in GS: inconsistencies between GS segments; erroneous citation links to some individual documents
- While Scopus is article-based, GS is more like a 'concept-based' citation index

The main conclusions of the study are presented in Slide 14.24. Two particular outcomes are presented in Slides 14.25 and 14.26 below.

¹Figure reprinted with permission from Moed, Bar-Ilan & Halevi (2016).



Not only the sheer number of documents indexed in a database is of interest, also the indexing *speed* is an important characteristic. Slide 14.25 compares the indexing speed of Google Scholar with that of Scopus.² A complicating factor is that Scopus indexes articles in press (denoted as AIP in the slide) *without* their cited references. Slide 14.25 analyzes cited references contained in documents published in journals indexed in Scopus. It marks the day in which they are entered into Google Scholar as a starting point, and counts after how many days it appears in Scopus. For instance, between 30 and 60 days after entry in Google Scholar, only 52% of cited references in Scopus-covered journals is actually found in Scopus; after 90–120 days this percentage amounts to 79. The overall outcome is that the Google Scholar indexes documents somewhat faster than Scopus does.

²Figure reprinted with permission from Moed, Bar-Ilan & Halevi (2016).

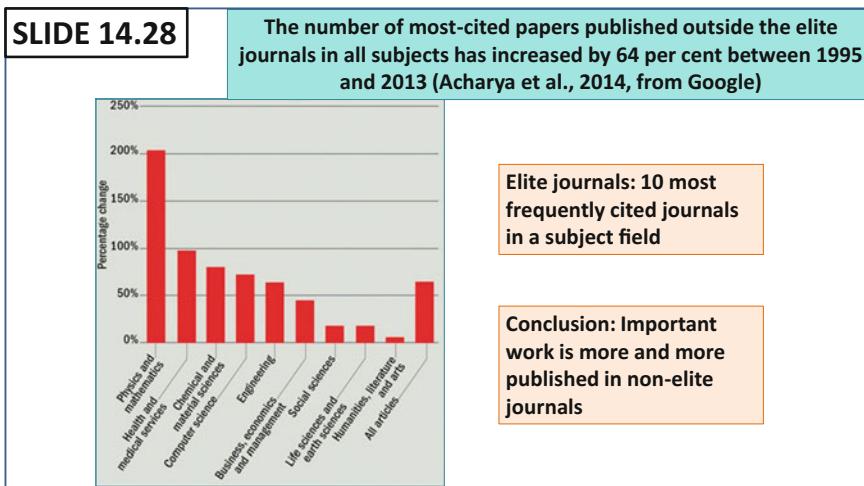
SLIDE 14.26		The 10 most frequently occurring <i>web links</i> of citing documents in sources in GS <i>not</i> indexed in Scopus
# Cites	Web link info	Comments
156	books.google.com	Google's Book Index
140	Springer	Monographs, book chapters, proceedings papers published by Springer
93	papers.ssrn.com	Documents posted in the Social Sciences Research Network
86	researchgate.net	Social networking site for scientists and researchers to share papers
63	dl.acm.org	ACM Digital Library containing full texts of all articles published by ACM
54	arxiv.org	Repository of freely available e-prints of scientific papers in physics a.o.
53	aclweb.org	Website of the Association for Computational Linguistics
39	anthology.aclweb.org	Digital Archive of research papers in Computational Linguistics
38	Wiley Online Library	Monographs, book chapters, proceedings papers published by Wiley
36	ieeexplore.ieee.org	Provides abstracts and full-text articles published by IEEE and IET

Slide 14.26 gives, for the sample of 12 journals listed in Slide 14.23, an overview of the most important sources in Google Scholar that are not indexed in Scopus. It must be noted that these outcomes very much depend upon the study sample. If the sample would have contained physics journals, ArXiv.org would probably rank much higher.

14.4 Concluding Remarks

SLIDE 14.27		Overall conclusion
<ul style="list-style-type: none"> • Google Scholar is a powerful tool to search relevant literature • It is also a fantastic tool to track one's own citation impact • It is up-to-date, and has a broad coverage • But there are many data quality issues • Its online metrics features are poor • Use in research evaluation requires data verification by assessed researchers themselves 		

Major conclusions related to Google Scholar are summarized in Slide 14.27. A distinction is made between its use as a literature retrieval tool, and as a research assessment tool.



The source coverage of Google Scholar, and, to a lesser extent, that of Scopus, are not based on Eugene Garfield's principle of selecting journals with a relatively large citation impact. While it can be maintained that Web of Science tends to contain in all fields the journals with the highest citation impact, this is *not* true for Google Scholar and Scopus. While Scopus has an active content advisory board responsible for quality control, Google Scholar tends to cover all sources that are available via the Web and revealing 'scholarly' characteristics, such as a reference list. Interestingly, an article by Acharya et al. (2014) found that important work is more and more published in non-elite journals. A selective citation index processing only journals with a higher citation impact could therefore miss important work. The figure in Slide 14.28 is based on Table 1 in Acharya et al. (2014).

Chapter 15

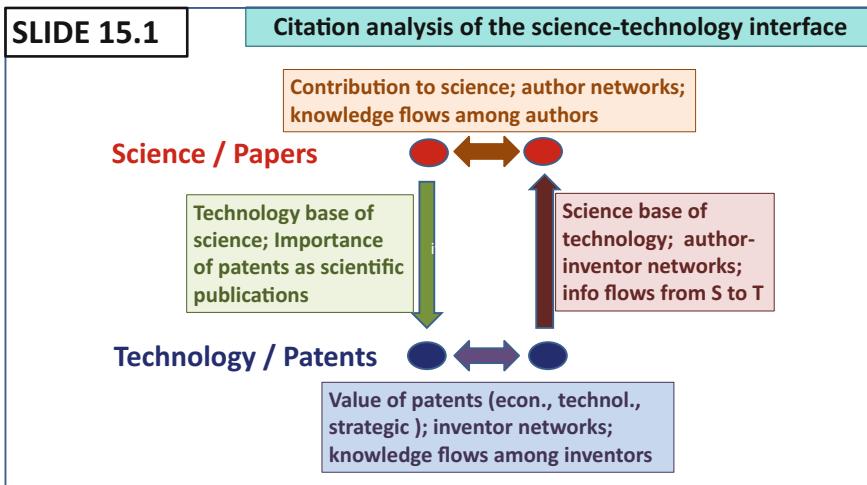
From Francis Narin's Science-Technology Linkages to Double Boom Cycles in Technology

Abstract This chapter presents a series of studies on the relationship between science and technology. It starts with showing the pioneering work by Francis Narin and co-workings on the citation analysis of the linkage between science and technology. Next, it presents several theoretical models of the relationship between science and technology. As an illustration of an analysis of the development of a technological field, it presents key findings from a study on industrial robots. The chapter ends with studies that illustrate the limitations of citation analysis of the scientific literature for the measurement of technological performance.

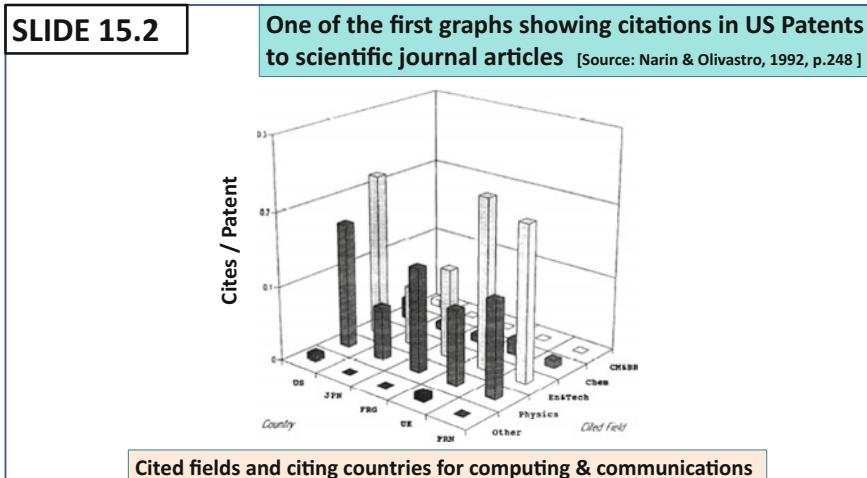
Keywords Customer satisfaction · Disciplinary specialization · Double-boom cycles · European patent office · Industrial robots · Investment in R&D · Knowledge production · Market pull · Mode 2 · Patent-to-paper citations · Patent-to-patent citations · Science-technology linkages · Technology push · Triple helix overlay model · US patent office

This chapter is based on a lecture presented by the author in a doctoral course given in February 2015 at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome.

15.1 Citation Analysis of the Science-Technology Interface



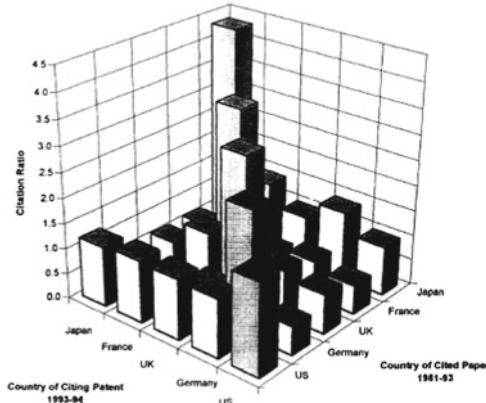
This chapter focuses on the science-technology interface, especially on the study of the science base of modern technology, as reflected in citations in patents to the scientific literature. Slide 15.1 gives an overview of the various types of citation analysis involving scientific publications in journals, denoted as papers, and patents.



Slide 15.2 presents one of the first analyses of citations in patents to scientific papers, made by Narin and Olivastro.¹ Francis Narin was the founder and, for many years, director of the information company CHI Research, specializing in analyses of scientific and technological performance. The strongest dependence of technology on science is found for all countries in physics and in engineering and technology.

SLIDE 15.3

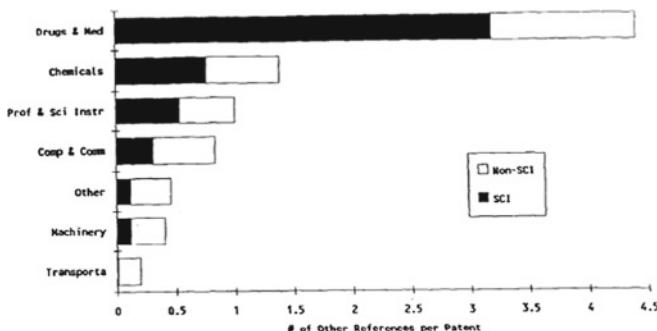
The linkage between technology and science has a strong national component [Source: Narin, Hamilton & Olivastro, 1997, p. 322]



Another analysis of citations in patents to papers reveals a national component and is presented in Slide 15.3.² Patents from a country tend to cite papers from the same country. But the extent to which countries do so varies from one country to another. This tendency is especially strong in Japan. A validation study of the use of patent citations for the study of the science base of technology, and the technological dependence of a nation to the technology from other countries, is presented in Carpenter and Narin (1983).

¹Figure reprinted with permission from Narin and Olivastro (1992).

²Figure reprinted with permission from Narin Hamilton and Olivastro (1997).

SLIDE 15.4**Average number of citations in US patents to non-patents (“other references”)** [Source: Narin, Hamilton & Olivastro, 1997, p. 245]

Linkage to science by product fields, and type of reference

If the average number of citations in patents to non-patents (mainly scientific papers) reflects the science intensity of a technological field, Slide 15.4 shows that large differences exist in science intensity among fields.³

SLIDE 15.5**Science-Technology Linkages: Indicators**

Citations in patents to patents or papers	Author-inventor networks; knowledge flows; science base; impact [next slide]
Authors as inventors	Author-inventor networks & collaboration
Co-publications academia - industry	Collaboration; author-inventor networks
Authors moving from acad to industry v.v.	Knowledge flows; networks
Funding acknowledgements in papers to industry	Funding base and industrial relevance of research
Multiple appointments	Networks; industrial relevance of research

Slide 15.5 gives an overview of indicators that are used to study science-technology linkages. Patent-to-paper citations and vice versa constitute one type of such linkages. Other types of linkages relate to authors and inventors,

³Figure reprinted with permission from Narin Hamilton and Olivastro (1997).

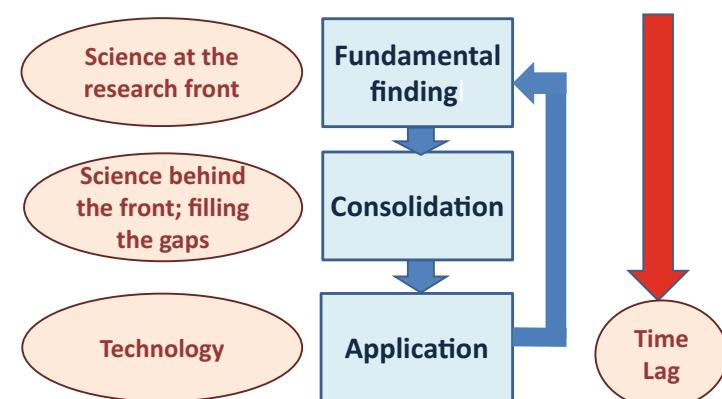
co-publications, scientific migration, funding acknowledgements and multiple professional appointments.

SLIDE 15.6**Citation gap (LePair, 1988)**

- The influence of technology upon scientific development (e.g., electron microscope)....
- and technological performance in general (e.g., half-open Eastern Scheldt dam)....
- tend to be **poorly** reflected in **cited references** in science papers
- **Full text analysis of scientific papers may provide insight into the influence of technology upon science**

It was Cees le Pair, former director of the Technology Foundation in the Netherlands, and founding father of the field of scientometrics in the Netherlands, who underlined that technological performance is not necessarily well reflected in cited references given in the scientific literature. Slide 15.6 shows that this especially applies to the impact of scientific equipment upon the development of a scientific field.

15.2 Theoretical Models of the Relationship Between Science and Technology

SLIDE 15.7**The Science-Technology Helix (Casimir, 1970s)**

This section starts with a model proposed by Hendrik Casimir, a Dutch physicist and former director of the Physics laboratory of the multinational company Philips. The model describes the relationship between scientific and technological development as a helix or a spiral.

SLIDE 15.8

Casimir's Science-Technology Helix - 1

- Technical application of a fundamental finding requires a time period of consolidation
- Consolidation: Activity behind the research front; solving details; clarification of unsolved issues
- Time lag is around 30 years. It did not decline during 1900-1980.
- Industrial/political pressures can to some extent shorten the time lag.
- Science-technology helix is an autonomous process
- Technology essential for development of science

Casimir conceived the science-technology helix as an *autonomous* process. As shown in Slide 15.8, he pointed towards the *time lag* that exists between the date a fundamental finding has been made and the date that a technical application based on this finding is launched. In between there is a phase of consolidation, which, using a notion introduced by Price, takes place *behind* the research front.

SLIDE 15.9

Casimir's Science-Technology Helix - 2

	Fundamental finding	Consolidation	Technology	
Laser	1917: Einstein's paper on quantum theory of radiation ("stimulated emission")	1940s-1950s: physicists needed and conceived a device producing light with short wavelengths	1960: concept patented; first laser built at Hughes Aircraft Co.	
Atomic bomb	1913: Bohr's Atom Model provides principle for studying spectral lines	1910s-1930s: Extensive research on spectral lines and behavior of atoms	1945: production of an atomic bomb in Manhattan project	
Industrial radiography	1895: Rontgen radiation can be used to study crystal structures	1890s-1900s: immediate application in medicine ; Numerous studies on cristal structures of materials	1912: Invention of high-vaccum X-Ray tubes; Industrial radiography . 1948: Discovery transistor	

Slide 15.9 gives three examples of developments from fundamental findings to technical application. It is taken from Sarlemijn (ed., 1984).

SLIDE 15.10

Casimir's Science-Technology Helix - 3

- Attempts to stimulate academic institutions towards consolidating fundamental knowledge often failed
- Hence, consolidation is a primary task for industrial laboratories
- An industry lab is a matching device between fundamental science and technological innovation

Slide 15.10. Casimir developed his views in the 1960s and 1970s. During the past decades, there is a tendency that industries close their laboratories or substantially reduce the size of their in-house research activities, and outsource a part of their R&D to academic institutes. If it is a regular task of universities to carry out R&D in what Casimir termed as the consolidation of fundamental knowledge, this has important consequences for the criteria that universities or external evaluation agencies should apply in the assessment of academic performance. As consolidation takes place behind the research front, informetric indicators measuring the visibility at the research front such as citation-based indicators are not the most appropriate measures.

SLIDE 15.11

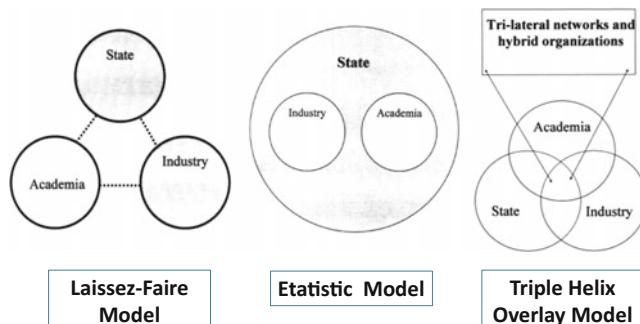
Types of knowledge production [Gibbons et al. 2000]

Mode 1	Mode 2	Comments/examples
Academic, fundamental	Context-driven	Mode 2 is the original format of science; mode 1 is construct to justify autonomy (E & L, 2000)
Investigator-initiated	Problem-focused	Fundamental research can be problem-focused ("strategic")
Discipline-oriented	Inter-disciplinary	Problem-solving may require both types of research
Traditional organization	Organizational diversity	Research group vs. practitioner group (e.g. lawyer's office)
Traditional quality contr.	New forms of quality control	Peer review vs. customer satisfaction surveys

Slide 15.11 presents a distinction between two types of knowledge production, proposed by a group of experts (Gibbons et al. 2000) in a study that has generated a lot of impact in the first decade of the 21th century. It aims to capture a mode of knowledge production (mode 2) that is context-driven, problem focused and inter-disciplinary, and is supposedly different from research normally conducted in academic institutions. Typical examples of institutions producing mode 2 knowledge are a lawyer's office or technical bureau. Note that customer satisfaction surveys are main tools in the performance assessment of this type of knowledge production.

SLIDE 15.12

Three models of the relationship between academia, industry and government [Etzkowitz & Leydesdorff, 2000]



Source: Etzkowitz and Leydesdorff, 2010, Research Policy 29, 109–123

Slide 15.12⁴ presents three models of the relationship between academic institutions, industry and government: a laissez-faire model (also denoted as double helix model) in which the three domains have their own dynamics and have mostly bilateral interactions; an estatistic model in which the state controls the other two domains, as for instance in many Eastern European countries before the fall of the Wall; and a Triple Helix Overlay Model with strong interaction between the domains, tri-lateral networks and hybrid organizations. The model is developed by Leydesdorff and Etzkowitz (2000).

⁴Figure reprinted with permission from Leydesdorff and Etzkowitz (2000).

SLIDE 15.13**Triple Helix Model** [Etzkowitz & Leydesdorff, 2000]

- “Triple Helix III is generating a knowledge infrastructure in terms of overlapping institutional spheres....
- with each taking the role of the other and with hybrid organizations emerging at the interfaces”(p. 111).
- “In contrast to a double helix (or a coevolution between two dynamics), a Triple Helix is not expected to be stable”(p. 112)
- “The Triple Helix hypothesis is that systems can be expected to remain in transition” (p. 113)
- “Most countries and regions are presently trying to attain some form of Triple Helix III” (p. 113)

Some main characteristics of the Triple Helix Overlay Model, developed by Etzkowitz and Leydesdorff (2000) are presented in Slide 15.13 and 15.14. These authors argue that most countries and geographic regions are seeking to implement (components of) such a model.

SLIDE 15.14**The common objective is to realize an innovative environment consisting of** [Etzkowitz & Leydesdorff, 2000]

- University spin-off firms
- Tri-lateral initiatives for knowledge based economic development
- Strategic alliances among firms, government laboratories, and academic research groups.

Governments encourage but do not control these arrangements as follows:

- Direct or indirect financial assistance
- through the Bayh–Dole Act in the USA
- Creation of new foundations to promote innovation

Slide 15.14 presents main characteristics of Triple Helix Overlay Model.

SLIDE 15.15

Innovation can be defined at different levels and from different perspectives [Etzkowitz & Leydesdorff, 2000]

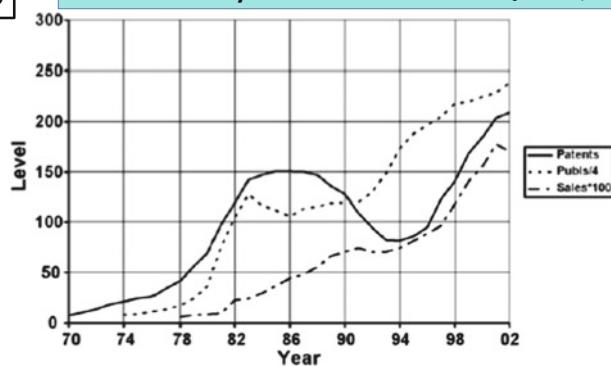
Perspective	View
Evolutionary economy	Considers firms as the units of analysis, since they carry the innovations and they have to compete in markets
Policy	Defines national systems of innovation as a relevant frame of reference for government interventions
Formal, network	Has networks as more abstract units of analysis: the semi-autonomous dynamics of the networks may exhibit lock-ins, segmentation

Slide 15.15 illustrates how according to Etzkowitz and Leydesdorff innovation can be defined and stimulated at different levels and from distinct perspectives: economic, policy and network-based.

15.3 Double Boom Cycles in Product Development

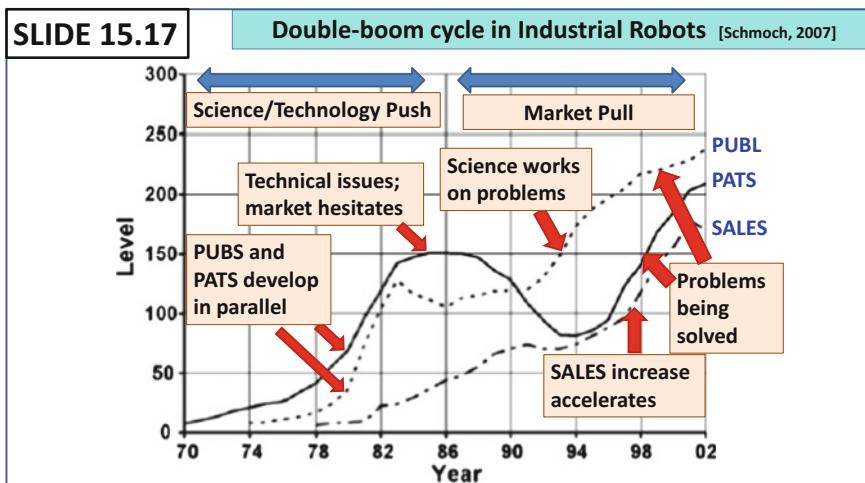
SLIDE 15.16

Double-boom cycle in Industrial Robots [Schmoch, 2007]



Number of EPO applications (by priority years)*, publications (by submission years)* and sales of industrial robots in Germany (in bn Euro) (*3-year moving average).

The figure on Slide 15.16 and on Slide 15.17 is taken from a paper by Ulrich Schmoch (2007) on the development of the technological field of *industrial robots*, and on the relationship between science and technology during its development. It plots per year the number of patents, scientific papers and German sales figures related to these robots, over a time period of more than 30 years.⁵



Slide 15.17 gives an interpretation of the trends measured in Slide 15.16. It clearly reveals a double boom in all three variables: publication counts, patent counts and sales. First there is a sharp increase, followed by a decline phase, and a subsequent period of increase. About halfway the time period, the development of industrial robots encountered serious problems, which had to be solved first in further scientific research. In terms of the Casimir model presented in Slide 15.7 this activity can perhaps be characterized as consolidation.

⁵Figure reprinted with permission from Schmoch (2007).

SLIDE 15.18**Double-boom cycles: observations [Schmoch, 2007]**

- Long time delays between fundamental finding and application (first “lab” robot created in 1950s)
- Innovation process has non-linear features.
- “First technology push, next market pull” is too simplistic
- Double boom cycles found in many science-based areas
- Additional (public) funding needed during stagnation period, by the end of the first boom.

Schmoch argues that in both phases technology push and market pull occurs, but the former dominates in the first period and the latter in the second. He underlines the long time delay between a first fundamental finding (a lab robot) and its application on a large scale.

SLIDE 15.19**Schmoch (2007): Science ← → Technology**

- “The results of pure basic, curiosity-oriented research referring to the fields analysed often appear several decades before the first patent boom, but are not taken up either by technology or by science” (p. 1009)
- “The basic problem is obviously to realise the basic theoretical concepts even on a laboratory scale. If a stage is reached where the transfer into technology seems to be realistic, the scientific and technical activities immediately begin in parallel” (p. 1009)

In Slide 15.19 Schmoch further explains his ideas about the interplay between science and technology. Perhaps the linkages between science and technology tend to have become stronger during the time period analyzed by Schmoch than they had been when Casimir formulated his model, which took place several decades earlier.

15.4 Selected Case Studies

SLIDE 15.20

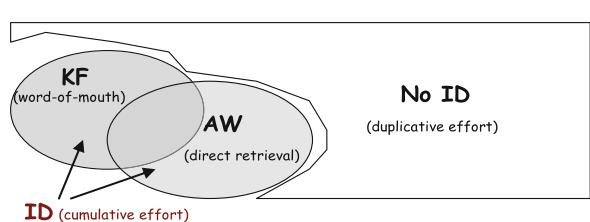
Differences between USPTO and EPO [Breschi & Lissoni, 2004]

Aspect	USPTO	EPO
Patent examiner's search report	Examiner's references are listed on front page	Separate document attached to patent once published
Other references, e.g., describing invention	Available but in less easily available format	Available but in less easily available format
Duties of applicants	Applicants must disclose all prior art they are aware of;	No requirement for applicant;
Role of examiner	Filtered by examiner; source indicated as from 2001	All references are from the examiner
Examiner's degree of selectivity	Provides a broader document search	Only prior art that threatens the application's patentability
Number of citations given by examiner	On average, around 10-16, depending upon study; For international patent applications similar to EPO	On average, around 3-5 cites, depending upon study; For international patent applications similar to USPTO

In studies based on patent citations many different patent databases are available. Slide 15.20 summarizes the main differences between two important ones: the database of the US Patent Office (USPTO), and that of the European Patent Office (EPO). (from Breschi & Lissoni 2004).

SLIDE 15.21

The meaning of patent citations [Breschi & Lissoni, 2004]



ID = INTELLECTUAL DEBT between cited-citing patent (cumulative effort); there is none if the citation is the result a mere duplicative effort

KF = KNOWLEDGE FLOW; the citation is the final outcome of a word-of-mouth diffusion process; an intellectual debt between cited-citing patent exists for sure, but not necessarily the citing inventor is aware of it

AW = AWARENESS; the citing inventor knows about the cited patent early on during his inventive effort; an intellectual debt between cited-citing patent exists, but it may be either the result of a direct retrieval of the cited patent from a database (AW still belongs to ID, no KF supports it) or a word-of-mouth diffusion process which includes news of the patent's existence (AW overlaps KF)

Slide 15.21.⁶ There is an ongoing debate about the meaning of patent citations. For instance, Breschi and Lissoni formulated a hypothesis related to patent-to-patent citations that distinguishes between citations that reflect a cumulative effort and intellectual debt between a citing and a cited patent, and those that are merely a result of a duplicate effort.

SLIDE 15.22**The Technological Impact of Library Science Research: A Patent Analysis**

(G. Halevi & H.F. Moed, Research Trends, 2012).

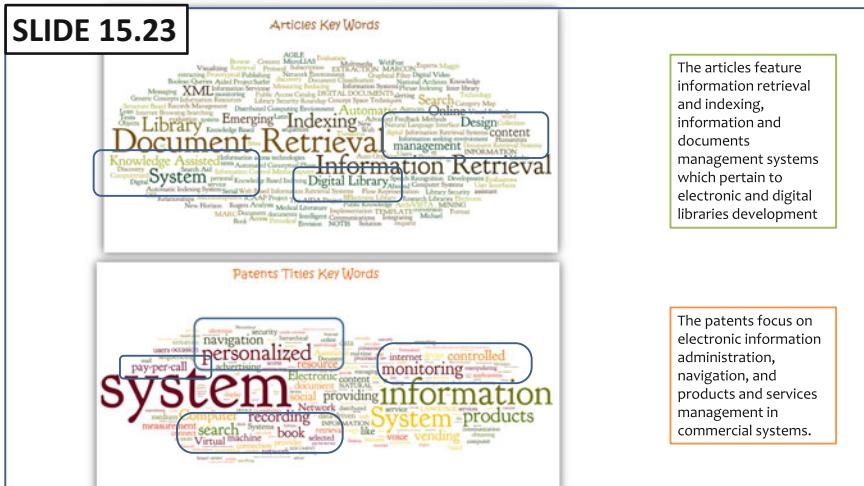
- The study examines the characteristics of 42 research articles published in **Library Science journals** and the manner by which they are cited in **patents** between 1991 and 2011

A study by Halevi et al. (2012) examined the technological impact of library science, by analyzing the patents that cited papers published in a set of library science journals. Slides 15.23–15.25 present more details on the study.⁷

⁶Reprinted with permission from Breschi and Lissoni (2004).

⁷Figures in Slides 15.23–15.25 are reprinted with permission from Halevi et al. (2012).

SLIDE 15.23

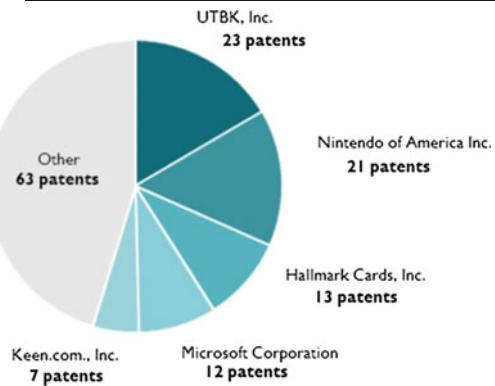


The upper text cloud in Slide 15.23 is extracted from the titles of the papers that were *cited in patents*, and that in lower part from the titles of the *citing patents*. The former shows terms from information retrieval and digital libraries, and the latter words from electronic information systems and products.

SLIDE 15.24

Patents citing Library Science journals, by industry

[Source: Halevi et al, 2012]



Slide 15.24 shows a breakdown of citing patents by assignee.

SLIDE 15.25

Patents citing Library Science journals, by main subject category [Source: Halevi et al, 2012]



Slide 15.25 shows that the key subject category is data processing, and that the most important sub-category relates to financial, business practices, management or cost aspects.

SLIDE 15.26

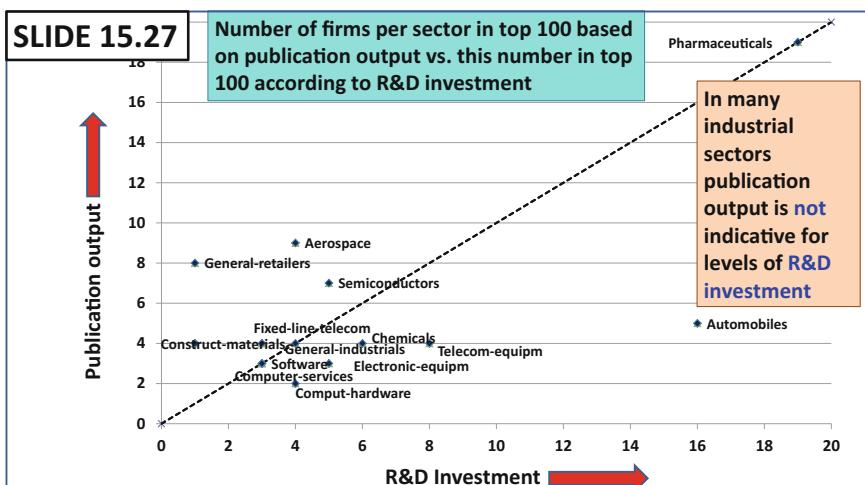
Position of firms in bibliometric rankings

Source: Moya-Anegon, Lopez-Illescas & Moed, 2014

- Among the institutions with the largest publication output, firms are underrepresented
- In many industrial sectors publication output is not indicative for levels of R&D investment.
- Among the institutions with the largest citation impact, firms are overrepresented.
- The firms with the largest citation impact tend to collaborate with the best public institutions
- Outcomes illustrate the crucial importance of public research institutions for the advancement of science and technology

The main conclusions of a study by Moya et al. (2014) on the position of firms in bibliometric rankings of institutions are presented in Slide 15.26. The study analyzed some 3600 institutions showed that the number of articles published by a company is not a good indicator of its investment in R&D. Interestingly, in a

ranking of all types of institutions according to their citation impact, firms appear to be over-represented. Slides 15.27 and 15.28 shed more light upon these outcomes.



Slide 15.27 clearly shows a lack of statistical correlation between a firm's publication output and the level of its investment in R&D. For instance, automobile firms invest large amounts of money in R&D, at a level similar to that of pharmaceutical companies, but they publish relatively few papers.

SLIDE 15.28

Among the institutions with the largest citation impact, firms are overrepresented

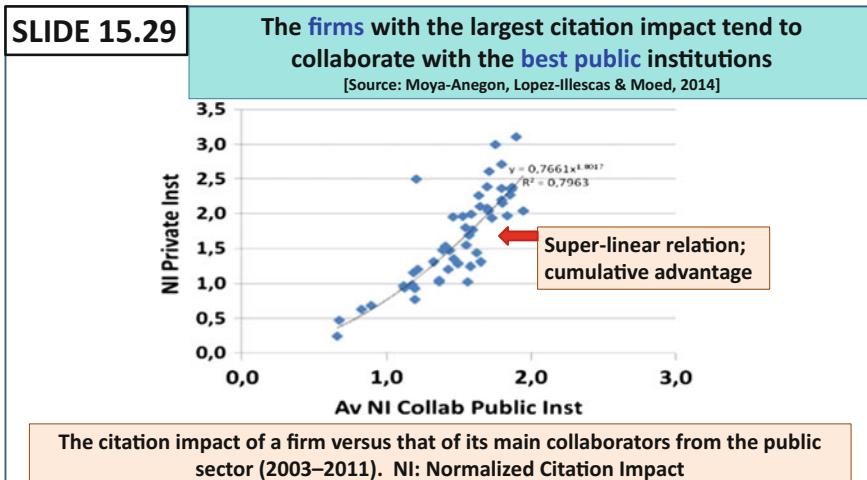
[Source: Moya-Anegon, Lopez-Illescas & Moed, 2014]

Table 7 The position of firms in the top quartile based on citation impact during 2003–2011(3,500 set)

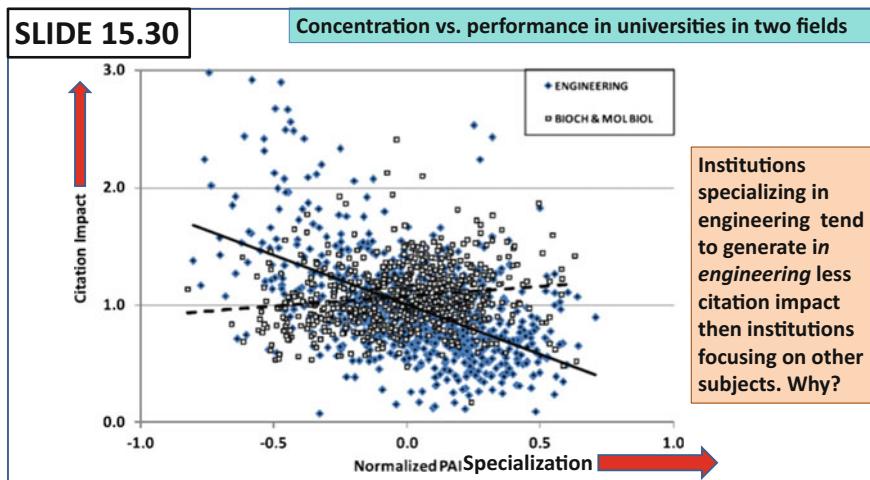
Sector	Total institutions		Institutions in top NI quartile			Institutions with NI > 1.75		
	(a) N	(b) %	(c) N	(d) %	(e) = (d)/(b)	(f) N	(g) %	(h) = (g)/(b)
HE	2,240	65.77	312	36.66	0.56	99	22.86	0.35
Governm	398	11.69	141	16.57	1.42	77	17.78	1.52
Health	630	18.50	316	37.13	2.01	191	44.11	2.38
Companies	115	3.38	67	7.87	2.33	55	12.70	3.76
Other	23	0.68	15	1.76	2.61	11	2.54	3.76
All sectors	3,406	100	851	100		433	100	

The position of firms in the top quartile based on citation impact during 2003–2011 (3,500 set). NI: Normalized Citation Impact.
Data from: SCIMAGOIR.com.

Slide 15.28⁸ shows that 2.33% of the total set of 3406 institutions analyzed by Scimago are companies. But in the subset of institutions with a normalized citation impact above 1.75, 12.7% are companies, which is 3.76 times the value in the total set.



According to Slide 15.29, when companies appear in the top of a ranking based on citation impact, they tend to have strong collaborations with the best public institutions (in terms of normalized citation impact denoted as NI).⁹



⁸Table reprinted with permission from de Moya-Anegon, Lopez-Illescas & Moed (2013).

⁹Figure reprinted with permission from de Moya-Anegon, Lopez-Illescas & Moed (2013).

A study presented in Slide 15.30 analyzes the statistical relationship between disciplinary specialization, i.e., the degree to which institutions specialize in a particular subject field, and the citation impact of their papers published in that field.¹⁰ Disciplinary specialization is normalized and expressed in an index that ranges between -1 and $+1$. In the field of engineering there appears to be a negative linear correlation between these two variables. A possible explanation is that performance in engineering is not well reflected in citation traffic between papers in scientific journals. This may be due to the citation gap outlined in Slide 15.6.

SLIDE 15.31	Conclusions
<ul style="list-style-type: none">• Relationship between science and technology is complex, non-linear• Time delays of one or more decades between fundamental invention and practical application• Citations in patents to papers constitute an interesting tool in the study of the S-T interface• There is a citation gap: technological performance may not be well reflected in citations in scientific articles• Double-boom phenomena in technological developments• In many fields scientific articles are not indicative of the level of R&D investment	

Some of the conclusions of the slides presented in this chapter are listed in Slide 15.31.

¹⁰Figure reprinted with permission from Moed de Moya Anagon Lopez Illescas and Visser (2011).

Chapter 16

From Journal Impact Factor to SJR, Eigenfactor, SNIP, CiteScore and Usage Factor

Abstract This chapter presents a series of journal metrics. It starts with a discussion of the journal impact factor, probably the most well-known bibliometric measure. It shows some of its technical limitations and dedicates in an analysis of editorial self-citations special attention to its sensitivity to manipulation. Next, a series of alternative journal citation measures is presented: SJR, Eigenfactor, SNIP, CiteScore, and indicators based on usage.

Keywords Citation potential · Citable documents · Free citations · Harmonic mean · Journal citation reports · Source normalization

16.1 Journal Impact Factors

SLIDE 16.1 **ISI/JCR Journal Impact Factor of journal J for year T**

Citations in year T to items published in J in years
T-1 and T-2

÷

Number of “citable” items published in J in years T-1
and T-2

This chapter is based on a lecture presented by the author in a doctoral course given in February 2015 at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome.

Slide 16.1 gives a definition of the journal impact factor, published by Thomson Reuters (formerly Institute for Scientific Information, currently Clarivate Analytics) in the *Journal Citation Reports*. The term ‘citable’ is explained in Slide 16.7 below.

SLIDE 16.2		ISI Journal impact factor citation window									
		← C I T I N G (Impact Factor) Y E A R S →									
C	I	06	07	08	09	10	11	12	13	14	15
I	06				↑↓						
T	07	X			↓↑						
E	08	X	X		↓↑						
D	09	X	X	X		↓↑					
Y	10	X	X	X	X		↓↑				
E	11	X	X	X	X	X			↑		
A	12	X	X	X	X	X	X		↑		
R	13	X	X	X	X	X	X	X	↑		
S	14	X	X	X	X	X	X	X	X	↑	
	15	X	X	X	X	X	X	X	X	X	

The citation window applied in the calculation of the journal impact factor (JIF) is explained in Slide 16.2. For instance, the JIF for the year 2015 is based on citations given in the (citing) year 2015 to documents published in the journal in 2013 and 2014.

SLIDE 16.3		InCites™ Journal Citation Reports®																							
NATURE ISSN: 0028-0836 NATURE PUBLISHING GROUP MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XH, ENGLAND ENGLAND		Titles ISO: Nature JCR: Abbrv: NATURE Categories MULTIDISCIPLINARY SCIENCES SCIE Languages ENGLISH 51 Issues/Year																							
Go to Journal Table of Contents Go to Ulrich's		<div style="border: 1px solid #ccc; padding: 5px;"> <h4>Journal Impact Factor</h4> <p>Cites in 2013 to items published in 2012 = 3478 Number of items published in 2012 = 669 2011 = 3762 Sum: 7240 2011 = 841 Sum: 1710</p> <p>Calculations: $\frac{\text{Cites to recent items}}{\text{Number of recent items}} = \frac{7240}{1710} = 42.351$</p> </div>																							
Key Indicators Year: 2013 Total Cites: 590		Article Influence Score (AIS) <table border="1"> <tr><td>105</td><td>22.184</td></tr> <tr><td>139</td><td>20.801</td></tr> <tr><td>124</td><td>20.373</td></tr> <tr><td>29</td><td>19.306</td></tr> <tr><td>105</td><td>18.062</td></tr> <tr><td>145</td><td>17.279</td></tr> <tr><td>80</td><td>16.996</td></tr> </table>										105	22.184	139	20.801	124	20.373	29	19.306	105	18.062	145	17.279	80	16.996
105	22.184																								
139	20.801																								
124	20.373																								
29	19.306																								
105	18.062																								
145	17.279																								
80	16.996																								
2013: 590 2012: 554 2011: 526 2010: 511 2009: 483 2008: 443 2007: 417.228 2006: 28.751 2005: 29.263 2004: 30.616 2003: 7.385 2002: 841 2001: 8.0 2000: 4.8 1999: 1.83870 1998: 16.996																									

Slide 16.3 shows how the JIF of one particular journal (Nature) is calculated from the raw citation and publication counts.

SLIDE 16.4

InCites™ Journal Citation Reports®

THOMSON REUTERS

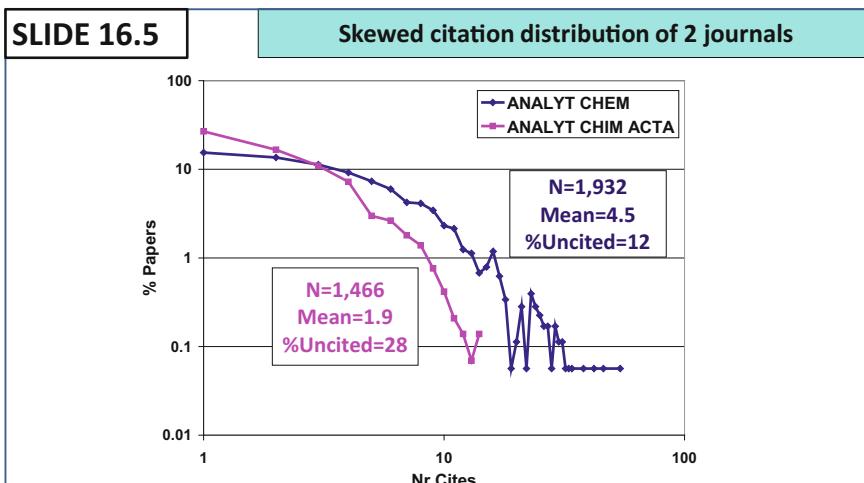
NATURE
ISSN: 0028-0836
NATURE PUBLISHING GROUP
MACMILLAN BUILDING, 4 CRINAN ST, LONDON N1 9XX, ENGLAND
ENGLAND

Go to Journal Table of Contents Go to Ulrich's

Key Indicators

Year	Total Citations (Growth)	Journal Impact Factor (Growth)	Impact Factor Without Journal Self-Cites	5 Year Impact Factor (Growth)	Immediacy Index (Growth)	Citable Items (Growth)	Cited Half-Life (Growth)	Citing Half-Life (Growth)	Eigenfactor Score (Growth)	Article Influence Score (Growth)
2013	590,324	42.351	41.650	40.783	8.457	857	9.8	5.4	1.60305	22.184
2012	554,745	38.597	37.956	36.159	9.243	869	9.6	5.2	1.56539	20.801
2011	526,505	38.280	35.707	36.235	9.690	841	9.4	5.1	1.65524	20.373
2010	511,248	36.104	35.527	35.248	8.792	862	9.1	5.2	1.73520	19.306
2009	483,039	34.480	33.855	32.906	8.209	866	8.9	5.1	1.74605	18.062
2008	443,967	31.434	30.864	31.210	8.194	899	8.5	4.9	1.76345	17.279
2007	417,228	28.751	29.263	30.616	7.385	841	8.0	4.8	1.83870	16.995

The JIF is the most frequently used indicator of journal impact, but the Journal Citation Reports present a series of citation-based indicators as well. Their names are indicated in Slide 16.4.



The JIF represents a mean value of a citation distribution that tends to be skewed. Both the horizontal and the vertical axis have a logarithmic scale. Slide 16.5 shows such distributions for two chemical journals. Data relate to the (citing)

year 2002.¹ The journal with the lowest JIF value of the two journals (1.46 for *Analytica Chimica Acta*) has a highest percentage of uncited papers or papers cited only once or twice, but the lowest percentage of papers with 4 or more citations.

SLIDE 16.6**Thomson/JCR Journal Impact Factor****Citations to all docs****# Citable docs**

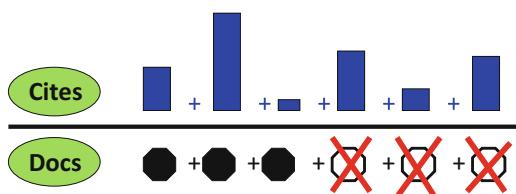
The JIF is a ratio of the number of received citations and the number of published documents. Scientific scholarly journals may publish many different types of documents. A problem with the JIF is that the sets of document types taken into account in the calculation of the numerator does not coincide with the documents included in the denominator. This is further explained in Slides 16.7, 16.8 and 16.9.

SLIDE 16.7**Citable vs. non-citable docs**

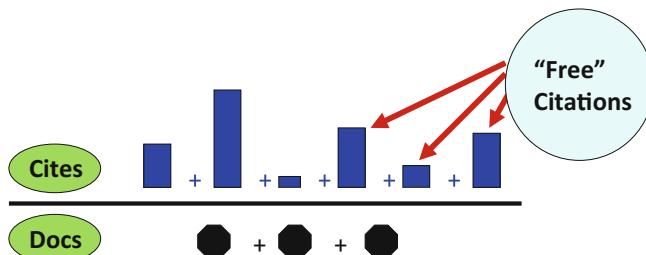
Citable documents	“non-citable” documents
Articles	Letters
Reviews	Editorials
	Discussion papers

¹Figure reprinted with permission from Moed (2005a).

Slide 16.7 indicates some document types. In the calculation of the JIF, articles and reviews are defined as ‘citable documents’. Other types, such as letters, editorials, are considered ‘non-citable’. But ‘non-citable’ documents can be cited, some even quite frequently.

SLIDE 16.8**The problem of “free” citations - 1**

In Slide 16.8 the horizontal line indicates the ratio’s division line. Six documents are plotted below this line. Three are ‘citable’, three others are ‘not citable’. The bars above the division line indicate the citation counts of each of these six documents. It is assumed that they are all cited, which for many journals is a realistic assumption. In the calculation of the JIF, the numerator includes citations to all six documents published in the journal, regardless of whether the cited documents are ‘citable’ or not, whereas in the denominator only the three citable documents are counted.

SLIDE 16.9**The problem of “free” citations - 2**

Slide 16.9 explains the notion of free citations. In this way, journals publishing for instance short letters, or discussion papers provoking a debate within the journal, may receive so called free citations: these citations are included in the citation count of the JIF' numerator, but the documents they cite are *not* counted in the denominator. In this sense, these citations are 'free'.

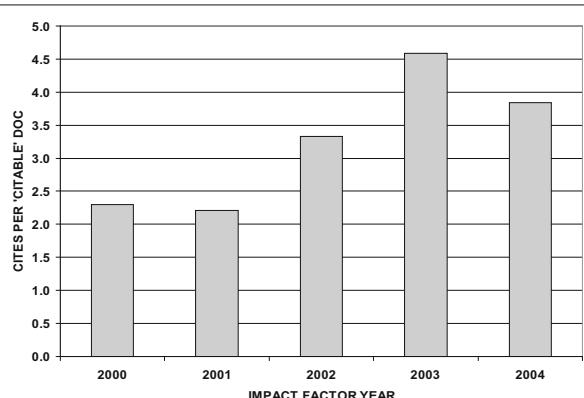
16.2 Effect of Editorial Self-citations

SLIDE 16.10

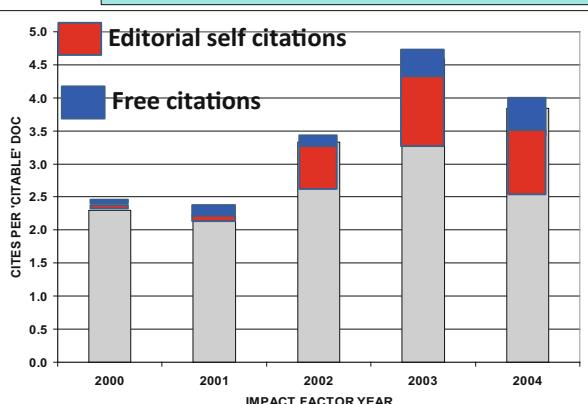
Effects of editorial self-citations upon journal impact factors [Reedijk & Moed, J. Doc., 2008]

- Editorial self-citations: A journal editor cites in his editorials papers published in his own journal
- Focus on 'consequences' rather than 'motives'

The next slides show how in principle JIF values can be manipulated. It presents one particular case. The analysis focus on *consequences* of the behavior of the journal editor, not on his *motives*. The current author does not want to suggest that the editor in this case manipulated the JIF deliberately. He may even not have been aware of the effects of his behavior.

SLIDE 16.11**Case: ISI/JCR Impact Factor of a Gerontology Journal
(published in the journal itself)**

Slide 16.11 shows the development over time of the value of the journal's JIF. The JIF increases over the years. The editor published this graph in one of his editorials, and used it as an illustration of the success of his editorial policy.

SLIDE 16.12**Decomposition of the IF of a Gerontology journal**

A decomposition of the JIF calculated the portion of the JIF value that is caused by *editorial self-citations*, and a second portion caused by the *free citations* outlined in Slide 16.9. These two portions are indicated in Slide 16.12. Deleting these two types of citations, the JIF hardly increases. In other words, the observed increase in Slide 16.11 is caused by specific editorial behavior.

SLIDE 16.13

One can identify and correct for the following types of strategic editorial behavior

- Publish ‘non-citable’ items
- Publish more reviews
- Publish ‘top’ papers in January
- Publish ‘topical’ papers (with high short term impact)
- Cite your journal in your own editorials
- Excessive journal self-citing
- Advertise via social media inc. blogs

Slide 16.13 shows editorial strategies that could increase the value of a journal’s JIF (Reedijk & Moed, 2009). The effects of the use of indicators upon author and editorial practices is also discussed in Sect. 9.4. It must be noted that there are ways to control at least partly for most types of strategic behavior. For instance, one can take into account only citations in peer-reviewed documents to other peer review documents, thus eliminating citations in or to editorials and other non-peer reviewed material. This is done in the calculation of the SNIP (see Slide 16.22).

16.3 Alternative Journal Metrics

SLIDE 16.14**Five Alternative journal citation-based indicators**

Indicator	Producer	Source	Availability
SJR	SCImago	Scopus	www.SCImago.com www.journalmetrics.scopus.com www.Scopus.com
Eigenfactor	Eigenfactor.org	Web of Science	www.webofscience.com www.eigenfactor.org
SNIP	CWTS, Leiden Univ	Scopus	www.scopus.com www.journalmetrics.scopus.com; www.journalindicators.com
H, H5	Google	Google Scholar	www.scholar.google.com
CiteScore	Elsevier	Scopus	https://journalmetrics.scopus.com/

Slide 16.14 gives a list of the indicators that will be discussed below.

SLIDE 16.15

www.scimagojr.com

The screenshot shows the Scimago Journal & Country Rank website. At the top left is the SJR logo. To its right is the text "SCImago Journal & Country Rank". Below this is a navigation menu with links to Home, Journal Rankings, Journal Search, Country Rankings, Country Search, Compare, Map Generator, Help, and About Us. To the right of the menu is a section titled "The Shape of Science" featuring a colorful network visualization of scientific publications. Below the visualization is a text box stating: "The Shape of Science is a new graphical interface designed to access the bibliometric indicators database of the SCImago Journal & Country Rank portal (based on 2012 data)." At the bottom right of this section is a blue button labeled "Open The Shape of Science".

Slide 16.5 presents the front page of a website created by Scimago.com, named Scimago Journal Rank. This database contains freely available indicators of citation impact for almost all journals indexed in Scopus, based on a citation analysis of an in-house version of Scopus. The acronym of this indicator is SJR (González-Pereira, Guerrero-Bote & Moya-Anegón, 2010). This indicator is also available via the Elsevier website www.journalmetrics.scopus.com.

SLIDE 16.16

SJR weights citations according to 'prestige' of citing source

The basic feature of the SJR is that it weights citations according to the ‘prestige’ of the citing journal. For instance, a citation in a prestigious journal as Nature has a higher weight than a citation from a more peripheral source. Slide 16.7 gives more information.

SLIDE 16.17**How SJR is calculated**

- The idea of **recursion** is essential
- Step by step, SJR **weights citations** in one step according to the SJR of the citing journal in the previous step
- Under certain conditions this process **converges**
- in the end a citation from a source with **high SJR** is **worth more than a citation from a source with low SJR**

The Eigenfactor metrics (Eigenfactor Score and Article Influence) available at <http://www.eigenfactor.org/projects/journalRank/journalsearch.php> are based on the same principle. The approach was invented by Pinski & Narin (1976), and constitutes also the basis of the Page Rank algorithm applied in Google.

SLIDE 16.18www.journalindicators.com

The screenshot shows the homepage of the CWTS Journal Indicators website. At the top, there is a navigation bar with links for Leiden University, CWTS, CWTS B.V., and Other CWTS sites. Below the navigation bar is a large banner featuring a photograph of bookshelves filled with books. The banner contains the text "Welcome to CWTS Journal Indicators". Below the banner, a paragraph of text provides information about the service: "CWTS Journal Indicators provides free access to bibliometric indicators on scientific journals. The indicators have been calculated by Leiden University's Centre for Science and Technology Studies (CWTS) based on the Scopus bibliographic database produced by Elsevier. Indicators are available for over 20,000 journals indexed in the Scopus database."

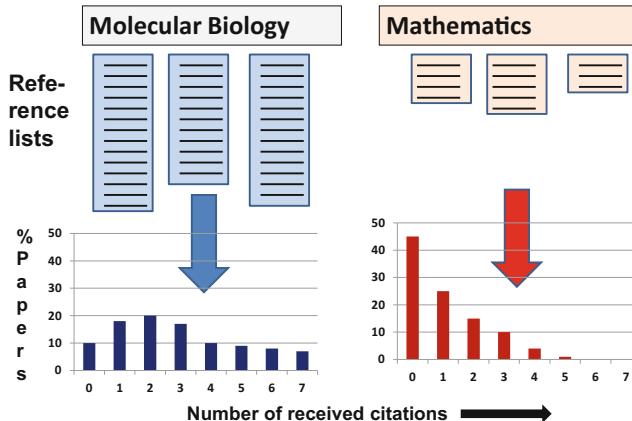
Slide 16.18 shows the front page of the website journalindicators.com, created by the Centre for Science and Technology Studies (CWTS) at Leiden University. It contains a series of freely available journal indicators calculated for a long range of years. One of the indicators is the Source Normalized Impact per Paper (SNIP).

SLIDE 16.19

SNIP corrects for disparities in citation potential among fields

SNIP = Source Normalized Impact per Paper

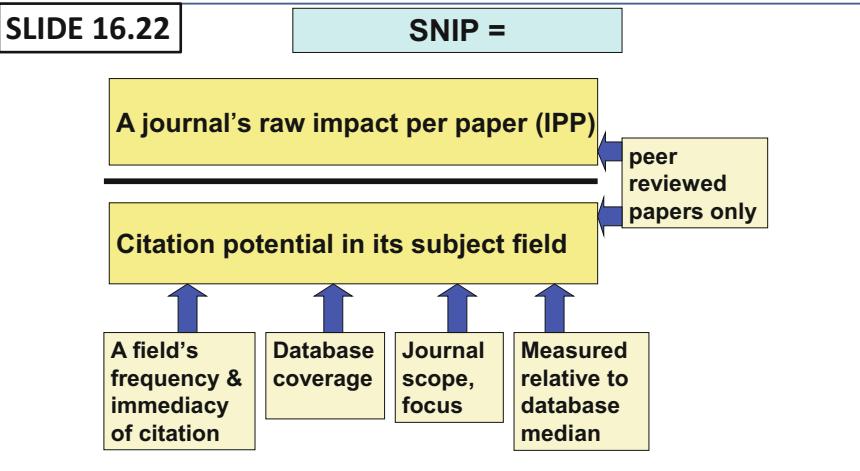
As indicated in Slide 16.19, SNIP aims to correct for differences in citation potential among subject fields. Its concept is partially based on an idea launched by Eugene Garfield (1979), and on an approach proposed by Zitt & Small (2008) and Zitt (2011). It was launched in Moed (2010). A modified version was proposed in 2013 (Waltman et al., 2013). The new version is also available via the Elsevier website www.journalmetrics.scopus.com.

SLIDE 16.20**Differences in citation potential between fields**

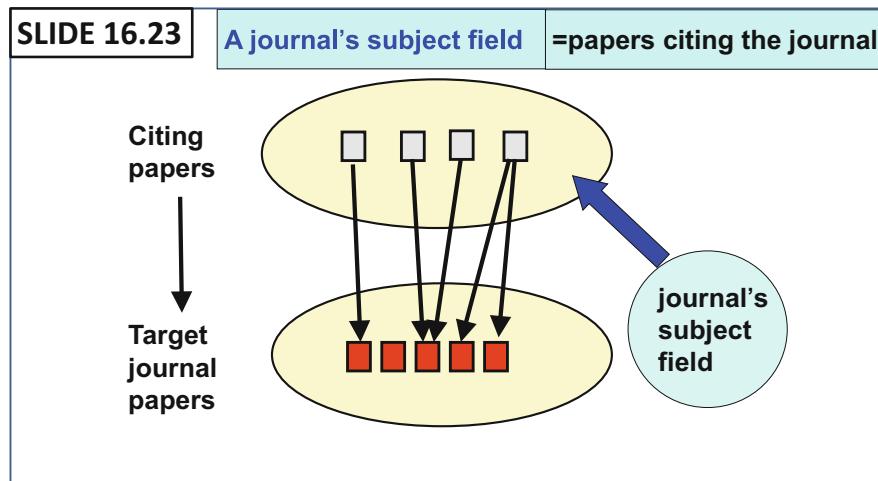
A key factor of citation potential is the length of the cited reference lists in articles published in a subject field. The more often authors cite other papers in their articles, the higher are the citation rates that papers in the authors' subject field receive on average. In technical terms, the field normalization of SNIP is based on the source or 'citing side' of the citation, and is therefore denoted as source normalization.

SLIDE 16.21		SNIP and SJR citation window										
		CITING (SNIP, SJR) YEARS										
C	I	06	07	08	09	10	11	12	13	14	15	
I	06											
I	07	X										
E	08	X	X									
E	09	X	X	X								
Y	10	X	X	X	X							
E	11	X	X	X	X	X						
A	12	X	X	X	X	X	X					
R	13	X	X	X	X	X	X	X				
S	14	X	X	X	X	X	X	X	X			
	15	X	X	X	X	X	X	X	X	X		

The citation window applied in the SNIP is indicated in Slide 16.21. It is one year longer than the window applied in the JIF; it relates to citations in a particular (citing) year to documents published during the *three* previous years. A window of three years is more appropriate than a 2-year window especially in fields in which the citation impact is maturing slowly.



Slide 16.22 gives the basic structure of the SNIP. It is a ratio of two indicators, each of which is a ratio itself. SNIP's numerator is the citation per article ratio over all articles and reviews published in a journal, termed as raw impact per paper or IPP. This is similar to the JIF, but there are important differences as well. (a) SNIP's citation window is one year longer; (b) SNIP's calculation is based on articles and reviews only, both on the citing (source) and on the cited (target) side, so free citations and *editorial self-citations* are *not* counted. SNIP's denominator is the citation potential in the subject field, defined as the average number of 1–3 year-old references per (source) paper. It is normalized so that the *median* journal obtains a value of one. As a consequence, 50% of journals has a (normalized) citation potential above 1.0, and another 50% below 1.0.



Slide 16.23 shows how a journal's subject field is defined, namely as the total collection of articles citing a particular journal.² SNIP does not use an a priori classification of journals into journal categories. In the SNIP method, a journal's subject field is directly determined by its citation links.

²Figure reprinted with permission from Moed (2010).

SLIDE 16.24**Example : Molec Biol vs. Mathematics**

<i>Journal</i>	<i>RIP(IPP)</i> <i>Raw impact / paper</i>	<i>Cit Pot</i> <i>Citation Potential</i>	<u><i>SNIP</i></u> <u>(= RIP/ Cit Pot)</u>
INVENT MATH	1.5	0.4	<u>3.8</u>
MOLEC CELL	13.0	3.2	<u>4.0</u>

Slide 16.24 shows that the journal Inventiones Mathematicae has a low IPP but also a low citation potential (far below 1.0). Molecular Cell has high values of these two parameters. But since the SNIP is defined as the ratio between the two parameters, their SNIP values are statistically most similar. Four main differences between the original SNIP and the new version (Waltman et al., 2013) are: (a) in the latter version the citation potential is expressed as a harmonic mean rather than an arithmetic mean; (b) certain sets of citing sources with few cited references are excluded; (c) the definition of a journal's subject field has changed; d) the new SNIP value is normalized so that an average journal has a value of 1.0. For a systematic comparison, see Waltman et al., 2013, and Moed (2016b).

SLIDE 16.25**H Index for journals (Google Scholar)**

- H index for a set of articles is the largest number H so that H articles have at least H citations
- H5-index is the H-index for articles published in the last 5 complete years:
- It is the largest number H such that H articles published in 2009-2013 have at least H citations each

The H index for authors is defined in Table 3.5 in Chap. 3. It can be easily extended to journals. A h-index of 20 means that a journal has published 20 articles with at least 20 citations (but not 21 articles with at least 21 citations). Google Scholar publishes h-index values for many journals. H5 is based on a 5-year time period.

SLIDE 16.26**Elsevier's CiteScore**

- *The three-year citation window.* Research has found that in slower-moving fields, two years' worth of data is too short; yet five years is too long to consider in faster-moving fields. Three years is the best compromise for a broad-scope database.
- CiteScore's numerator and denominator *both include all document types*. This means not only articles and reviews but also letters, notes, editorials, conference papers and other documents indexed by Scopus are included. As a result, the numerator and the denominator used in the CiteScore calculation are consistent.

SOURCE: www.journalmetrics.scopus.com

Recently Elsevier has launched a journal metric termed as CiteScore. The time window is identical to that applied in IPP and SNIP. But while SNIP and SJR focus on document types that tend to be subjected to peer review, thus creating a sub-universe of citation relations between peer-reviewed documents as the basis of the calculation, CiteScore includes *all* document types, *both* at the citing *and* at the cited side.

SLIDE 16.27 Usage- or downloads-based journal indicators

Indicator	Definition
Journal Usage Factor (JUF) [Counter, n.d.]	Ratio of the total usage over period X of items published in a journal during period Y, and the total number of items published in that journal during period Y (COUNTER, n.d.)
Download Immediacy Index (DII) [Wan et al., 2010]	Number of downloads in a particular year to a journal's items published in the same year

Slide 16.27 presents two journal measures based on usage, measured by the number of downloads or html-views of the full text of documents published in a journal. The first, Journal Usage Factor (JUF) is not yet fully specified; the description indicates the general scheme but no time windows. The second, Download Immediacy Index specifies the time window. It is identical to that of another journal citation indicator generated by Thomson Reuters in the Journal Citation Reports, namely the Immediacy Index.

Chapter 17

From Relative Citation Rates to Altmetrics

Abstract This chapter presents definitions and properties of a series of informetric indicators that are discussed in earlier chapters of this book: relative citation rates, h-index, Integrated Impact Indicator, usage-based indicators, social media mentions, and research efficiency measures. It highlights their potential and limits, and gives typical examples of their application in research assessment.

Keywords Concentration measure · Economies of agglomeration · Economies of scale · Gini index · Productivity indicators

17.1 Citation-Based Indicators

SLIDE 17.1

Relative Citation Rate (RCR) or Crown Indicator (CI)

The average citation rate of a unit's papers
÷
world citation average in the subfields in
which the unit is active

Corrects for differences in
citation practices among fields,
publication years and type of article

This chapter is based on a lecture presented by the author in a doctoral course given in February 2015 at the Department of Computer, Control and Management Engineering in the Sapienza University of Rome.

One of the earliest types of citation-based indicators is a field-normalized indicator of citation impact, also termed as relative citation rate or by some as crown indicator. There are various versions of this type of indicator. Slide 17.1 shows one that does not only correct for differences in citation rates between subject fields, but also for differences in type of document—as review articles tend to be more often cited than normal research articles—and in publication year or age, since older papers tend to be cited up-to-date more often than recently published articles.

SLIDE 17.2

Normalized citation rate

- PRO

- Corrects for differences in citation practices among subject field
- May also take into account document type (e.g., review, full length article), and age of cited article
- Takes into account the full citation distribution

- CON

- Should be used with caution when comparing entities with very different publication volumes or active in highly specialized subjects
- Field delimitation must be sound

The main pros and cons of relative citation rates are summarized in Slide 17.2, and also in Table 3.5 in Sect. 3.6, and in Sect. 4.3 in Part 2 of this book. One limitation is that, since the measure is expressed as a (normalized) citation-per-paper ratio, one may easily lose a sense of the underlying absolute numbers.

SLIDE 17.3

Two ways to calculate Crown Indicator (CI)

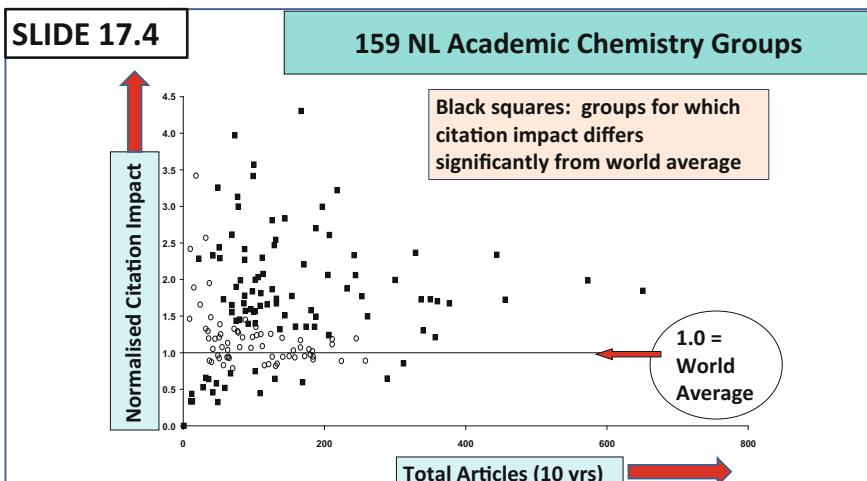
Group Publication Table

Pa per	Field	Type	Year	Cites
A	Math	Rev	2001	11
B	Astro	Art	2005	7
		$\text{CI-1} = \frac{11 + 7}{8.9 + 6.7} = 1.15$		
		$\text{CI-2} = \frac{11 - 7}{8.9 - 6.7} = 1.14$		2

Field averages Table

Field	Type	Year	Cites/ Paper
Math	Art	2001	5.4
Math	Rev	2001	8.9
.....			
Math	Art	2002	5.1
Math	Rev	2002	7.9
.....			
Astro	Art	2005	6.7
Astro	Rev	2005	13.5

Slide 17.3 compares two ways for calculating relative citation rates. The slide shows how the various operations are carried out. The first calculates a ratio of averages, and the second an average of ratios. In a series of tests on real data, Lariviere and Gingras (2011) found a strong correlation between the two measures, but in some cases also statistically significant differences between the two distributions generated by the two methods.



As an illustration of relative citation rates, Slide 17.4¹ plots for 159 research groups in chemistry the relative citation rate of their papers against the number of papers they published. This type of graphs was used in the Netherlands in the assessment of academic performance in science fields (van Raan, 2004a). See also Sect. 6.5 in Part 2 about assessment models and strategies.

¹Figure reprinted with permission from Moed (2005a).

SLIDE 17.5**Definition of Integrated Impact Indicator (I3)**

[Leydesdorff, Bornmann, Mutz & Ophof (2011); Leydesdorff & Bornmann (2011)]

1. For each paper to be evaluated, a reference set of papers published in the same year, of the same type and belonging to the same WoS category is determined.
2. These are rank ordered and split into percentile rank (PR) classes, for example the top 1 percent (99th percentile), 5 percent, 10 percent, 25 percent, 50 percent and below 50 percent. For each PR, the minimum number of citations necessary to get into the class is noted. Based on its citations, the paper is then assigned to one of the classes. This particular classification is known as PR.
3. The procedure is repeated for all the target papers and the results are then summated, giving the overall percentage of papers in each of the PR classes. The resulting distributions can be statistically tested against both the field reference values and against other competitor journals or departments.

SOURCE: Leydesdorff & Mingers, 2015

Slide 17.5 gives a definition of the Integrated Impact Indicator (I3). For a further discussion the reader is referred to Sect. 7.3 in Part 3 on how evaluative assumptions shape indicators.

SLIDE 17.6**Integrated Impact Indicator (I3)****• PRO**

- Combines an assessment of both quantity (published papers) and impact (citations)
- Corrects for differences in citation practices among subject field
- Takes into account all publications, rewarding entities with a large article production

• CON

- Its value is biased in favor of senior researchers compared to juniors;
- Maps all actual values onto a 0-100 scale;
- One may lose the sense of underlying absolute differences, and undervalue extraordinary papers

The potential and limits of the indicator (in the view of the current author) are listed in Slide 17.6. A strong feature is that this indicator does take into account differences in citation practices between disciplines, but is *not* constructed as a field-normalized citation-per-article *ratio* of the type presented in Slide 17.1.

SLIDE 17.7**Definition h-index**

- A scientist has index h if h of his/her N_p papers have at least h citations each, and the other ($N_p - h$) papers have no more than h citations each [Hirsch, 2005]
- H-index does not take into account academic age, but another parameter introduced by Hirsch, m : if n denotes the length of the time window taken into account, “quite generally, the slope of h versus n , the parameter m , should provide a useful yardstick to compare scientists of different seniority”.

The observation that h-index is biased towards senior researchers is valid, but it must be noted that Hirsch was very well aware of this bias and presented in his paper that launched his h-index also a parameter m that can be used as yardstick to compare researchers of *different seniority*. But the publishers of h-indices, Web of Science, Scopus and Google Scholar, and many others, do not present any information on the latter parameter.

SLIDE 17.8**All three publication lists have a Hirsch Index of 5**

	Author 1	Author 2	Author 3
1	30 P1	30 P1	100 P1
2	10 P2	10 P2	70 P2
3	8 P3	8 P3	8 P3
4	6 P4	6 P4	6 P4
5	5 P5	5 P5	5 P5
6	1 P6	4 P6	1 P6
7	0 P7	4 P7	0 P7
8		4 P8	
9		4 P9	

H= 5 **H= 5** **H= 5**

Slide 17.8 shows the lists of publications ranked by citation count of three authors. The lists are rather different. Compared to author 1, author 2 has published more papers, while author 3 has published two highly cited publications. But the

value of the h-index is in all three cases the same. This slide illustrates which type of differences between citation distributions tends to disappear in the calculation of the h-index.

SLIDE 17.9

H-Index

- PRO
- Combines an assessment of both quantity (published papers) and impact (citations).
- Tends to be insensitive to highly cited outliers and to poorly cited papers
- CON
- Its value is biased in favor of senior researchers compared to juniors;
- Impact of the most cited papers hardly affects its value
- Does not correct for differences between subject fields

Slide 17.9 describes the potential and limits of the h-index. For a further discussion the reader is referred to Sect. 4.3 in Part 2 and Sect. 7.3 in Part 3 of this book.

17.2 Usage-Based Indicator and Altmetrics

SLIDE 17.10

Downloads-citations: Analogy Model

<u>Formal use (citations)</u>	<u>Informal use (downloads)</u>
(Collections of) publishing authors	(Collections of) users
Citing a document	Downloading the full text of a document
Article	User session
Author's institutional affiliation	User's account name
Number of times cited	Number of times downloaded as full text

Slide 17.10 marks analogies between citations and full text downloads. It shows that downloads can to some extent be quantitatively modelled and analyzed in the same way as citations. Many concepts that are used in citation analysis make sense also in usage analysis. But Slide 17.11 shows that there are large differences as well.

SLIDE 17.11**10 important factors differentiating between downloads and citations**
[Moed & Halevi, JASIST, 2014]

- 1 **Usage leak:** Not all downloads may be recorded.
- 2 **Citation leak:** Not all citations may be recorded.
- 3 **Downloading** the full text of a document does **not** mean that it is **read**.
- 4 The **user (reader)** and the **author (citer)** population may not coincide.
- 5 Distribution # downloads **less skewed** than that of # cites, and depends upon the **type of document** differently
- 6 Downloads and citations show different **obsolescence functions**.
- 7 Downloads and citations measure **distinct concepts**.
- 8 Downloads and citations may **influence one another** in multiple ways.
- 9 Download counts are more sensitive to **manipulation**.
- 10 Citations are **public**, usage is **private**.

Main differences between downloads and citations are listed in Slide 17.11. Chap. 19 presents a paper presenting a detailed comparison between the statistical properties of these counts.

SLIDE 17.12**Full text article download counts****• PRO**

- Are in principle available immediately after publication
- Enable researchers to assess the effectiveness of their communication strategies
- May reveal attention of scholarly audiences from other research domains or of non-scholarly audiences

• CON

- Downloaded articles may be selected according to their face value
- Incomplete data availability across providers
- Affected by differences in reading behavior between disciplines and institutions
- Counts can be manipulated

Slide 17.12 lists the potential and limits of usage-based indicators. They are further discussed in Sect. 4.3. A major concern is that while citations are the results of at least some kind of reflection upon their relevance of the cited work for the research described in a paper, usage counts tend to reflect a publication's face value. They are less appropriate as indicators of research performance, but are apt to assess the effectiveness of a communication strategy.

SLIDE 17.13

Mentions in social media

- | | |
|---|---|
| <ul style="list-style-type: none"> • PRO • Are immediately available after publication • May reveal impact upon non-scholarly audiences • Measure attention rather than influence • May provide tools to link expertise to societal needs | <ul style="list-style-type: none"> • CON • Scientific-scholarly impact and societal impact are distinct • Do not measure scientific-scholarly impact • Interdependence of the various social media may boost numbers • Altmetrics are easily be manipulated |
|---|---|

Slide 17.13 focuses on the use of mentions in social media. The potential and limits of this indicator, and of altmetrics in general, is extensively discussed in Chap. 11 in Part 4.

SLIDE 17.14

Issues in efficiency measurement of research [Bonacorsi & Daraio, 2007]

Problem	Specification
Definition of inputs and outputs	All factors can be conceived both as input and as output
Identification of unit of analysis	Research team or lab is the best level but often no data is available
Methodologies for data collection	Lack of standardisation ; input and output data collection developed independently
Comparative issues	Comparisons difficult to make due to lack of standardisation
Endogeneity	Output is partly an effect of input
Assumptions made	Production function approach may be invalid
Dynamic relations	Outputs follow inputs with an unknown, variable time lag structure

17.3 Efficiency Indicators

Efficiency indicators or productivity indicators measure the ratio of the output of an activity to its inputs. While these two terms are often used interchangeably, some authors distinguish efficiency from productivity by defining it as a distance between the quantity of input and output rather than a ratio. Major issues in the measurement of efficiency are summarized in Slide 17.14. Important articles on this topic are published by Bonacorsi & Daraio (2003a, 2003b, 2004, 2005, 2007), Daraio & Simar (2007), and Daraio, Bonacorsi & Simar (2015).

SLIDE 17.15

Efficiency analysis of research: Inputs

Type	Examples
Human resources	Number of researchers differentiated by age, level of qualification or seniority; technical staff, administrative staff
Financial resources	Governmental research funds, funds raised from the market (block grant – competitive funding)
Physical resources	Physical capital, equipment, laboratories, libraries
Cumulated stock of knowledge	Number and quality of publications in the past

The various types of inputs to be considered in an efficiency analysis are listed in Slide 17.15.

SLIDE 17.16

Two topics in the economics of science

	Topic	Question
1	Economies of scale	Is there a positive effect of concentration of resources in big institutions upon research productivity?
2	Economies of agglomeration	Is there a positive effect of territorial concentration resources?

This section dedicates attention to the two topics indicated in Slide 17.16: Economies of scale and economies of agglomeration. These topics are illustrated below with a number of empirical case studies.

SLIDE 17.17

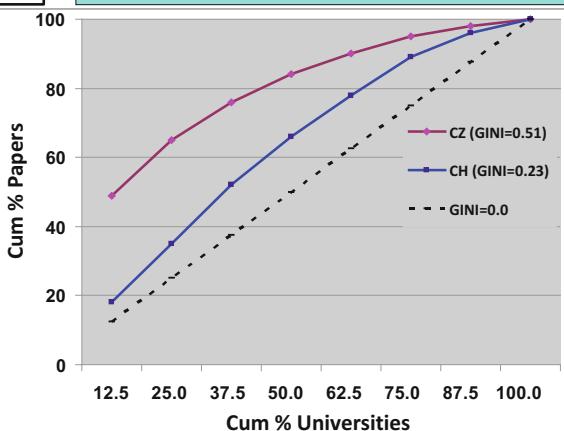
1. Is there an economy of scale in scientific research?

- Evidence for economy of scale found in manufacturing industry
- Research differs from manufacturing
- Existence of a minimum efficient scale for administrative costs and physical infrastructures is plausible
- In research, economies of scale may be important up to a threshold level, then become irrelevant.
- Research group size in universities is limited by senior-junior relationship and varies across disciplines
- Peer-peer collaborations are often outside the institution

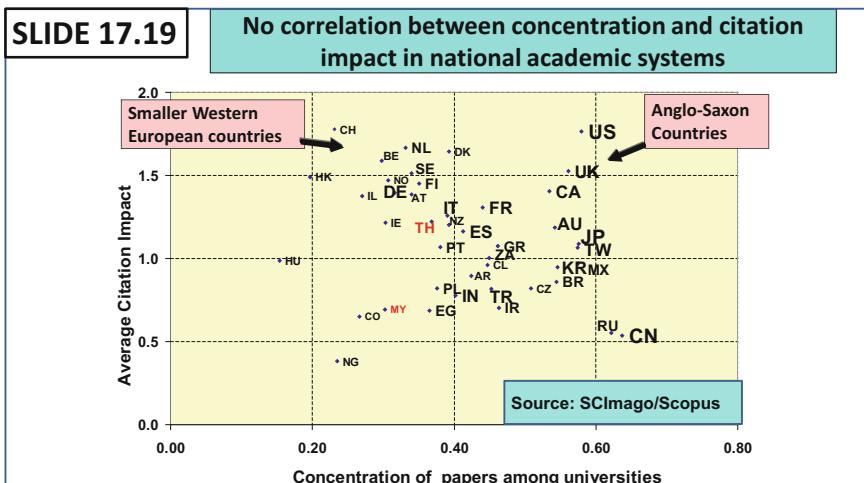
Bonacorsi and Daraio (2003a; 2003b; 2004; 2005; 2007) made a series of observations on the issue of economy of scale. The main ones are listed in Slide 17.17. The third observation suggests that input size may have a positive effect on efficiency up to a certain point. When production units increase their size above a certain threshold, their efficiency tends to decline. Also, the slide underlines the differences in research structures between the various research disciplines.

SLIDE 17.18

GINI Index of Concentration (for countries)



Slide 17.18 visualizes the value of the Gini index, a measure of concentration often used in economics. As an example, it shows data on the distribution of publication output among universities in two countries: Czech Republic (CZ) and Switzerland (CH). Per country, universities are ranked by their publication count. The figure plots the cumulative percentage of published papers against the cumulative percentage of universities publishing these papers. The surface between the curve and the diagonal (dashed line) reflects the value of the Gini concentration measure. It ranges between 0 (no concentration) and 1 (all publication output concentrated in one university only).



Slide 17.19 plots for 40 major countries the average citation impact of papers published by the universities in a country, and the degree of concentration (Gini index) of papers among universities in that country.² The plot shows in the upper part (with an impact above 1.5) both a group of Anglo-Saxon countries with large concentration (US, UK, on the right hand side of the graph), as well as a set of smaller European countries (Switzerland, Netherlands, Belgium, Denmark, Sweden) with a smaller degree of concentration in their national academic systems.

²Figure reprinted with permission from Moed, de Moya Anagon, Lopez Illescas & Visser (2011).

SLIDE 17.20**Does more concentration lead to better academic research?** [Moed, de Moya-Anegón, López-Illescas, Visser (2011)]

- In a group of 40 major countries:
- No significant linear or rank correlation is found between national research performance and the degree of concentration of research among their universities

The conclusions from Slide 17.19 are drawn in Slide 17.20. In the set of 40 countries no significant linear or rank correlation is found between citation impact and concentration within a national academic system. It must be underlined that the analysis takes into account only 2 variables. National academic systems are much more complex. But at least it questions a claim not seldom heard in policy circles, namely that more concentration of research leads to a better performance. Interestingly, the study also analyzed the relationship between performance and disciplinary concentration *within* universities (Moed et al., 2011).

SLIDE 17.21**Bonaccorsi & Daraio in a study of 169 CNR institutes**

- Scientific productivity seems not favored by the concentration of resources into larger institutes geographically agglomerated.
- Interestingly enough, size negatively affects the performance of all CNR institutes.
- Study applied advanced, robust, non-parametric efficiency analysis

Slide 17.21. In a study by Bonacorsi and Daraio (2003a) on 169 institutes of CNR, the Italian Research Organization, size is found to be *negatively* correlated with performance.

SLIDE 17.22**2. Agglomeration economies: base hypothesis**

- The **diffusion of knowledge** may take place via codification and distant transmission;
- but in most cases requires also **personal acquaintance and face to face interaction**.
- This is made easier and cheaper by **physical proximity**.

Slide 17.23 formulates the basic hypothesis of the economies of agglomeration (Bonacorsi & Daraio, 2005).

SLIDE 17.23**2. Agglomeration economies: evidence**

- **Industrial districts** provide specialized suppliers, highly trained workforce, shared residential neighborhoods
- **Costs of production** are therefore **lower** in an agglomerated area than outside it.
- There is evidence of knowledge spillover effects (exchange of ideas among individuals)

Slide 17.23 summarizes some of the empirical evidence in favor of the economies of agglomeration hypothesis.

SLIDE 17.24**Fractional Scientific Strength (FSS) of an individual**

[Abramo & D'Angelo, 2014; 2016]

Relative Citation Impact of one single paper: $\frac{\text{Number of citations}}{\text{Citation average of benchmark group} * \text{number of co-authors}}$

÷

Fractional scientific strength of a researcher: $\frac{\text{Sum of Relative Citation Impact over all papers by the researcher}}{\text{Average yearly salary} * \text{number of years}}$

÷

Slide 17.24 describes the efficiency indicator proposed by Abramo & D'Angelo, and discussed in Sect. 7.3 of this book. It can be calculated for an individual researcher, but also for a group or institution. On the output side, it takes into account only citation impact. The measure of citation impact corrects for differences in citation practices among subject fields, and also for the number of co-authors of the cited publications, under the assumption that all co-authors of a paper contribute equal portions to it. On the input side, the only parameter is the average yearly salary of the participating researchers. Thus, the proposed efficiency measure expresses the field-normalized, ‘co-author normalized’, citation count generated per Euro (or dollar) spent on salaries.

Part VI

Papers

Chapter 18

A Comparative Study of Five World University Rankings

Abstract To provide users insight into the value and limits of world university rankings, a comparative analysis is conducted of 5 ranking systems: *ARWU*, *Leiden*, *THE*, *QS* and *U-Multirank*. It links these systems with one another at the level of individual institutions, and analyses the overlap in institutional coverage, geographical coverage, how indicators are calculated from raw data, the skewness of indicator distributions, and statistical correlations between indicators. Four secondary analyses are presented investigating national academic systems and selected pairs of indicators.

Keywords Academic reputation · Alumni · Awards · Citations per faculty · European higher education area · Faculty-student ratio · Geographical distributions · Highly cited researchers · Industry income · Institutional overlap · International faculty · Per capita performance · Quality of faculty · Ranking methodology · Teaching performance

18.1 Introduction

In most OECD countries, there is an increasing emphasis on the effectiveness and efficiency of government-supported research. Governments need systematic evaluations for optimizing their research allocations, re-orienting their research support, rationalizing research organizations, restructuring research in particular fields, or augmenting research productivity. In view of this, they have stimulated or imposed evaluation activities of their academic institutions. Universities have become more diverse in structure and are more oriented towards economic and industrial needs.

In March 2000, the European Council agreed a new strategic goal to make Europe “the most competitive and dynamic knowledge-based economy in the world, capable of sustainable economic growth with more and better jobs and greater social cohesion”. Because of the importance of research and development to

This chapter re-uses with permission the text of Moed (2017b).

“generating economic growth, employment and social cohesion”, the Lisbon Strategy says that European universities “must be able to compete with the best in the world through the completion of the European Higher Education Area” (EU Council, 2000). In its resolution ‘Modernizing Universities for Europe’s Competitiveness in a Global Knowledge Economy’, the European Council expressed the view that the “challenges posed by globalization require that the European Higher Education Area and the European Research Area be fully open to the world and that Europe’s universities aim to become worldwide competitive players” (EU Council, 2007, p. 3).

An Expert Group on the assessment of university-based research noted in 2009 that university rankings have become an increasing influence on the higher education landscape since US News and World Report began providing consumer-type information about US universities in 1983. They “enjoy a high level of acceptance among stakeholders and the wider public because of their simplicity and consumer type information” (AUBR Expert Group, 2009, p. 9).

University ranking systems have been intensely debated, for instance by van Raan (2005), Calero-Medina et al. (2008), Salmi (2009), Hazelkorn (2011), Rauhvargers (2011; n.d.) and Shin, Toutkoushian and Teichler (eds.) (2011). A report from the European University Association concluded that despite their shortcomings, evident biases and flaws, rankings are here to stay. “For this reason it is important that universities are aware of the degree to which they are transparent, from a user’s perspective, of the relationship between what it is stated is being measured and what is in fact being measured, how the scores are calculated and what they mean” (Rauhvargers, 2011, p. 7).

A base notion underlying the current article is that a critical, *comparative* analysis of a *series* of university ranking systems can provide useful knowledge that helps a wide range of interested users to better understand the information provided in these systems, and to interpret and use it in an informed, responsible manner. The current article aims to contribute to such an analysis by presenting a study of the following five ranking systems: **ARWU** World University Rankings 2015, CWTS **Leiden** Ranking 2016, **QS** World University Rankings 2015–2016, **THE** World University Rankings 2015–2016, and **U-Multirank** 2016 Edition. An overview of the indicators included in the various systems is given in Table 18.10.

ARWU, the Academic Ranking of World Universities, also indicated as ‘Shanghai Ranking’ is the oldest ranking system. Initially created by the Center for World-Class Universities (CWCU) at Shanghai Jiao Tong University, since 2009 it has been published and copyrighted by ShanghaiRanking Consultancy. It combines bibliometric data from Thomson Reuters with data on prizes and awards of current and former academic staff or students. The **ARWU** 2015 Ranking of World Universities, available online and analyzed in the current article, covers 500 institutions. The **Leiden** Ranking is not a ranking in the strict sense but rather a bibliometric information system, containing for about 850 universities bibliometric data extracted from Web of Science related to publication output, citation impact and scientific collaboration. This article uses the 2016 version of the database.

U-Multirank is prepared with seed funding from the European Union by a Consortium lead in 2016 by the Center for Higher Education Policy Studies (CHEPS), The Netherlands; Centre for Higher Education (CHE) in Germany; and the Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands. This article is based on the 2016 version. A key feature of the ***U-Multirank*** system is the inclusion of teaching and learning-related indicators. While some of these relate to a university as a whole, the core part is concerned with 13 specific scientific-scholarly disciplines, and based on a survey among students.

Between 2004 and 2009, **Times Higher Education (THE)** and **Quacquarelli Symonds (QS)** jointly published the **THES-QS** World University Rankings. After they had ended their collaboration, the methodology for these rankings continued to be used by **QS** as the owner of its intellectual property. Since 2010 these rankings are known as the **QS** World University Rankings. At the same time, **THE** started publishing another ranking, applying a methodology developed in partnership with Thomson Reuters in 2010, known as the Times Higher Education or **THE** World University Rankings and related rankings. At present, both organizations have a collaboration with Elsevier, and use bibliometric data from Scopus.

A series of interesting studies analyzed statistical properties and validity *within* particular university ranking systems (e.g., Soh, 2013; Paruolo, Saisana and Saltelli, 2013; Soh, 2015a; Soh, 2015b), mostly focusing on the so called *Overall* indicator which is calculated as a weighted sum of the various indicators. For instance, a factor analysis per ranking system conducted by Soh (2015a) found that the factors identified in **ARWU**, **THE** or **QS** systems are negatively correlated or not correlated at all, providing evidence that the indicators covered by each system are not “mutually supporting and additive”. Rather than dealing with the internal consistency and validity *within* a particular system, the current chapter makes comparisons *among* systems.

All five systems listed above claim to provide valid and useful information for determining academic excellence, and have their own set of indicators for measuring excellence. Three systems, **ARWU**, **THE** and **QS**, present an overall indicator, by calculating a weighted sum of scores of a set of key indicators. The **Leiden** Ranking and ***U-Multirank*** do not have this type of composite measure. The current chapter examines the consistency among the systems. As all systems claim to measure essentially academic excellence, one would expect to find a substantial degree of consistency among them. The overarching issue addressed in the current chapter is the assessment of this consistency-between-systems. To the extent that a lack of consistency is found,—and the next chapters will show that it exists,—what are the main causes of the observed discrepancies? What are the systems’ profiles? How can one explain to potential users the ways in which the systems differ one from another? What are the implications of the observed differences for the interpretation and use of a particular system as a ‘stand-alone’ source of information?

The article consists of two parts. In the *first* part, a series of statistical properties of the 5 ranking systems are analyzed. The following research questions are addressed.

- *Overlap in institutional coverage* (Sect. 18.3). How many institutions do the rankings have pairwise in common? And what is the overlap between the top 100 lists in the various rankings? If this overlap is small, one would have to conclude that the systems have different ways to define academic excellence, and that it is inappropriate to speak of “*the*” 100 global top institutions.
- *Differences in geographical coverage* (Sect. 18.4). How are the institutions distributed among countries and world regions in which they are located? Are there differences in this distribution between ranking systems? All five systems claim to adopt a global viewpoint; ***ARWU***, ***THE*** and ***QS*** explicitly speak of *world* universities. But do they analyse the world in the same manner? Are differences between global geographical regions mainly due to differences in excellence in those regions, or do regional indicator normalizations play a significant role as well?
- *Indicator distributions and their skewness* (Sect. 18.5). Firstly, to which extent do the systems present for each institution they cover scores for *all* indicators? When assessing the information content of a system, it is important to have an estimate of the frequency of occurrence of missing values. Secondly, which methods do the systems apply to calculate scores from the raw data? Such methods determine how differences in indicator scores should be interpreted in terms of differences in underlying data. For instance, ***ARWU***, ***THE*** and ***QS*** express an indicator score as a number ranging from 0 to 100, while U-Multirank uses five so called performance classes (A to E). How precisely are these scores defined, and, especially, which differences exist between systems? Finally, how does the skewness of indicator distributions vary between indicators and between ranking systems? To what extent is skewness as measured by the various systems a base characteristic of the global academic system, or is it determined by the way in which the systems calculate their indicators?
- *Statistical correlations between indicators* (Sect. 18.6). The least one would expect to find when comparing ranking systems is that (semi-) identical indicators from different systems, such as the number of academic staff per student, show a very strong, positive correlation. Is this actually the case? Next, how do indicators from different systems measuring the same broad aspect (e.g., citation impact or academic reputation) correlate? If the correlation is low, what are the explanations? To what extent are indicators complementary?

In the *second* part of the chapter (Sect. 18.7) four analyses show how a more detailed analysis of indicators included in a system, and, especially, how the *combination* of indicators from *different* systems can generate useful, new insights and a more comprehensive view on what indicators measure. The following analyses are presented.

- *Characteristics of national academic systems.* What is the degree of correlation between citation- and reputation-based indicators in major countries? This analysis is based on indicators from the ***THE*** ranking. It aims to illustrate how simple data representations, showing for instance in scatterplots how pairs of

key indicators for a given set of institutions are statistically related, can provide users insight into the structure of underlying data, raise critical questions, and help interpreting the indicators.

- ***QS versus Leiden citation-based indicators.*** What are the main differences between these two indicators? How strongly do the correlate? Are they interchangeable? The main purpose of this analysis is to show how indicator normalization can influence the rank position of given universities, and also to underline the need to systematically investigate the data quality of ‘input-like’ data such as number of students or academic staff obtained via institutional self-reporting or from national statistical offices.
- ***THE Research Performance versus QS Academic Reputation.*** What are the main differences between the **THE** and **QS** reputation-based indicators? How strongly do they correlate? Which institutions show the largest discrepancies between **THE** and **QS** score? This analysis provides a second illustration of how indicator normalization influences university rankings.
- ***ARWU Highly Cited Researchers versus Leiden Top Publications indicator.*** Gingras (2014) found severe biases in the Thomson Reuters List of Highly Cited Researchers, especially with respect to Saudi Arabian institutions. Do these biases affect the **ARWU** indicator that uses this list as data source? This fourth study shows how a systematic comparison of indicators of the same broad aspect from different systems can help interpreting the indicators, and evaluating their data quality and validity.

Finally, Sect. 18.8 presents a discussion of the outcomes and makes concluding remarks

18.2 Analysis of Institutional Overlap

In a first step, data on the names and country of location of all institutions, and their values and rank positions for all indicators in as far as available were extracted from the websites of the 5 systems, indicated in Table 18.10 at the end of Sect. 18.5. Next, names of institutions were standardized, by unifying major organizational and disciplinary terms (e.g., ‘university’, ‘scientific’) and city names (e.g., ‘Roma’ vs. ‘Rome’), and an initial version of a thesaurus of institutions was created, based on their appearance in the first ranking system. Next, this thesaurus was stepwise expanded, by matching it against the institutional names from a next ranking system, manually inspecting the results, and updating it, adding either new variant names of institutions already included, or names of new institutions not yet covered. As a final check, names of institutions appearing in the top 100 of one system but not found in the other systems, were checked manually. In the end, 1715 unique institutions were identified, and 3248 variant names. 377 universities (22%) appear in all 5 ranking systems, and 182 (11%) in 4 systems.

A major problem concerning university systems in the USA was caused by the fact that it was not always clear which components or campuses were covered. For instance, University of Arkansas System has 6 main campuses. **ARWU** has two entries, ‘U Arkansas at Fayetteville’ and ‘U Arkansas at Little Rock’. **Leiden** includes ‘U Arkansas, Fayetteville’ and ‘U Arkansas for Medical Sciences, Little Rock’. **QS**, **THE**, and **U-Multirank** have one entry only, ‘U Arkansas’. Similar problems occur for instance with ‘Univ Colorado’, ‘Univ Massachusetts’, ‘Purdue Univ’ and ‘Univ Minnesota’. If it was unclear whether two institutions from different ranking systems covered the same components or campuses, they were considered as different, even if there is a substantial overlap between the two.

Table 18.1 presents the institutional overlap between each pair of ranking systems. The numbers in the diagonal give the total number of institutions covered by a particular system. Table 18.2 gives key results for the overlap in the top 100 lists of all 5 systems. It shows that the total number of unique institutions in the top 100 lists of the five systems amounts to 194. Of these, 35 appear in all lists.

Table 18.3 shows the institutional overlap between *the top 100 lists* of the various systems. For **ARWU**, **QS** and **THE** the ‘overall’, weighted ranking was used. **Leiden** and **U-Multirank** do not include such an overall ranking. For **Leiden**, two top 100 lists were created, one size-dependent, based on the number of

Table 18.1 Institutional overlap between the 5 ranking systems

Ranking	ARWU	LEIDEN	QS	THE	U-MULTIRANK
ARWU	500	468	444	416	465
LEIDEN		840	585	589	748
QS			917	635	638
THE				800	627
U-MULTIRANK					1293

Table 18.2 Key results overlap analysis of top 100 lists in all 5 ranking systems

Indicator	N
Total number of different institutions	194
Number of institutions appearing in the top 100 lists of all 5 systems	35

Table 18.3 Institutional overlap between the top 100 lists of 4 ranking systems

Ranking	LEIDEN-CIT	LEIDEN-PUB	QS	THE
ARWU	60	67	60	66
LEIDEN-CIT		49	51	56
LEIDEN-PUB			64	68
QS				75

publications (labelled as LEIDEN-PUB in Table 18.3), and a second size-independent (LEIDEN-CIT), based on the Mean Normalized Citation Score (MNCS), a size-normalized impact measure correcting for differences in citation frequencies between subject fields, the age of cited publications, and their publication type (see Leiden Indicators, n.d.). Since there is no obvious preferred ranking in ***U-Multirank***, this system was not included in Table 18.3. The number of overlapping institutions per pair of systems ranges between 49 for the overlap between the two ***Leiden*** top lists, and 75 for that between ***QS*** and ***THE***.

It should be noted that the overwhelming part of the top institutions in one ranking but missing in the *top 100* of another ranking were found at *lower* positions of this other ranking. In fact, the number of cases in which a top institution in a system is not linked to any university in another system ranges between 0 and 6, and most of these relate to institutions in university systems located in the USA.

Several cases were detected of institutions that could not be found in a system, while one would expect them to be included on the basis of their scores in other systems. For instance, *Rockefeller University*, occupying the 33th position in the overall ***ARWU*** ranking, and first in the ***Leiden*** ranking based on relative citation rate, is missing in the ***THE*** ranking. *Freie Univ Berlin* and *Humboldt Univ Berlin*—both in the top 100 of the overall ***THE*** ranking and in the top 150 of the ***QS*** ranking—could not be found in the ***ARWU*** system, while *Technical Univ Berlin*, ranking 178th in the ***QS*** system, was not found in the ***THE*** system. In the ***THE World Ranking*** the Italian institutions *Scuola Normale Superiore di Pisa* and *Scuola Superiore Santa Anna* are in the range 101–200. In fact, the first has the largest score on the ***THE*** Research Performance indicator. But institutions with these two names do *not* appear in the ***QS World University Ranking***; it is unclear whether the entity ‘*University of Pisa*’, appearing in the overall ***QS*** ranking on position 367, includes these two schools.

18.3 Geographical Distributions

The preference of ranking system R for a particular country C is expressed as the ratio of the actual and the expected number of institutions from C appearing in R, where the expected number is based on the total number of institutions across countries and across systems, under the assumption of independence of these two variables. A value of 1.0 indicates that the number of institutions from C in R is ‘as expected’. See the legend to Table 18.4 for an exact definition. Table 18.4 gives for each ranking system the five most ‘preferred’ countries. It reveals differences in geographical coverage among ranking systems. It shows the orientation of ***U-Multirank*** towards Europe, ***ARWU*** towards North America and Western Europe, ***LEIDEN*** towards emerging Asian countries and North America, and ***QS*** and ***THE*** towards Anglo-Saxon countries, as Great Britain, Canada and Australia appear on both.

Table 18.4 Five most 'preferred' countries per ranking system

System	Country	Nr. Univs	Preference	System	Country	Nr. Univs	Preference
ARWU	Canada	20	2.1	THE	Taiwan	24	2.0
	USA	146	2.1		Great Britain	78	1.9
	Netherlands	12	2.1		Australia	31	1.8
	Great Britain	20	1.9		Canada	25	1.7
	Germany	39	1.5		Japan	41	1.4
LEIDEN	China	108	1.9	U-MULTI-RANK	Netherlands	20	1.3
	Korea	33	1.8		Spain	67	1.3
	Canada	28	1.8		Poland	45	1.3
	Taiwan	19	1.5		Germany	84	1.3
	USA	173	1.5		Portugal	27	1.3
QS	Australia	33	1.7				
	Great Britain	75	1.6				
	Brazil	22	1.6				
	Canada	26	1.5				
	Korea	27	1.4				

Table 18.5 Country of location of unique institutions in top 100 lists

Ranking system	Nr unique univs	Country of location with ≥ 2 univs
ARWU	11	USA (4), Israel (2)
THE	8	Germany (3), USA (2), Netherlands (2)
QS	14	Great Britain (3), Hong Kong (2) Korea (2)
LEIDEN-PUB	11	China (6), Italy (2)
LEIDEN-CIT	26	USA (9), Great Britain (6), Switzerland (2) France (2)

Legend to Table 18.4. The preference P of ranking system R for a particular country C is defined as follows. If $n[i,j]$ indicates the number of institutions from country i in system j , $\sum_i n[i,j]$ the sum of $n[i,j]$ over all i (countries), and $\sum_j n[i,j]$ the sum of $n[i,j]$ over all j (systems), $P = (n[i,j]/ \sum_i n[i,j]) / (\sum_j n[i,j]/ \sum_i \sum_j n[i,j])$.

A second way to analyze differences in geographic orientation among ranking systems focuses on the *top 100 lists* in the *ARWU*, *QS* and *THE* rankings based on their overall score and on the two *Leiden top lists*, rather than on the total set of covered institutions analyzed in Table 18.4, and identifies for each system the country of location of ‘unique’ institutions, i.e., universities that appear in a system’s the top list but that are not included in the top list of any other system. The results presented in Table 18.5 are not fully consistent with those in Table 18.4, due to differences among countries in the frequency at which their institutions appear in top 100 lists, but there is a considerable agreement between the two tables. Table 18.5 reveals that in the *ARWU* and the *Leiden CIT* top list most unique institutions are from the USA, and in the *QS* top from Great Britain and two Asian entities: Korea and Hong Kong (formally a part of China). Unique institutions in the *Leiden PUB* top list are especially located in China, and, to a lesser extent, in Italy, and those in the *THE* top list in Germany, USA and The Netherlands.

18.4 Indicator Scores and Their Distributions

18.4.1 Missing Values

In the *ARWU*, *THE* and *QS* rankings the *overall* indicators are presented only for the first 100, 200 and 400 universities, respectively. In addition, *QS* presents on its website for *all* its indicators only values for the first 400 institutions. Occasionally, values are missing. This is true, for instance, in the *QS* system for the values of Rockefeller University on the indicators Academic Reputation, Employer Reputation and Overall Score. As regards *U-Multirank*, not all universities have participated in the surveys per subject field, and those who did were not necessarily involved in each subject field. Of the about 1300 institutions retrieved from the *U-Multirank* website, 28% has a score for the indicator quality of teaching in at least one subject field, and 12% in at least three fields.

18.4.2 From Data to Indicators

Both **ARWU** and **QS** apply the method of *normalizing by the maximum*: for each indicator, the highest scoring institution is assigned a score of 100, and other institutions are calculated as a percentage of the top score. Standard statistical techniques are used to adjust the indicator if necessary. The **QS** documentation adds that for some indicators a cut-off is applied so that multiple institutions have score 100. In fact, for the indicators citations per faculty, academic reputation and employer reputation the number of institutions with score 100 is 10, 12 and 11, respectively.

The **THE** system applies a *percentile rank-based approach*: For all indicators except the Academic Reputation Survey, a cumulative probability function is calculated, and it is evaluated where a particular institution's indicator sits within that function, using a version of Z-scoring. For the Academic survey, an exponential component is added. This is illustrated in Fig. 18.1. It plots the scores in the **THE** Ranking 2016 against percentile rank scores calculated by the author of this article. For the citations all observations are plotted on the diagonal. This illustrates that **THE** citation scores are in fact percentile rank scores. Figure 18.1 reveals how radically the **THE** research and teaching performance scores deviate from percentile rank scores, and how strong the exponential component is. 90% of institutions has a Research or Teaching Performance Score below 55 or 50, respectively.

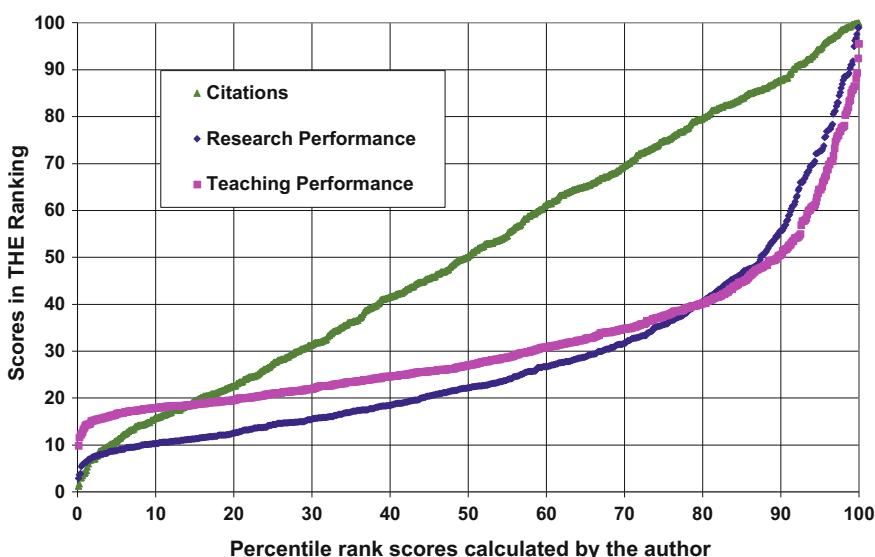


Fig. 18.1 Scores of three key indicators in the **THE** ranking plotted against their percentile rank calculated by the author of the current article

U-Multirank applies a ‘*distance to the median*’ approach. Per indicator, universities are assigned to 5 performance groups ranging from excellent (= A) to weak (= E), based on the distance of the score of an individual institution to the median performance of all institutions that **U-Multirank** has data for. It should be noted that the distribution of indicator values (A-E) may substantially vary from one indicator to another, and deviates strongly from a distribution based on quintiles. For instance, as regards the absolute number of publications the percentage of institutions with score A, B, C, D and E is 2.6, 47.3, 25.5, 20.7 and 0.0, respectively (for 3.9% no value is available). For the number of publications cited in patents these percentages are 30.6, 7.4, 11.6, 30.3 and 8.8 (for 11.2% no value is available), and for the number of post doc positions 15.3, 4.0, 3.9, 15.3 and 5.0 (for 56.5% data is unavailable).

18.4.3 Skewness of Indicator Distributions

Table 18.6 presents for a group of 17 indicators the skewness of the indicator distributions related to all institutions for which data are available. Figure 18.2 visualizes the distribution of 7 key indicators by plotting the institutions’ scores as a function of their rank. Table 18.5 shows that the **Leiden** absolute number of ‘top’ publications,—i.e., the number of publications among the 10% most frequently

Table 18.6 Skewness of 17 indicator distributions

	All Universities	
	Nr.Univs	Skew-ness
LEIDEN Nr. Top Publications (Top 10%)	840	4.03
ARWU Awards	500	3.03
LEIDEN Publications	840	2.56
ARWU Alumni	500	2.55
ARWU Publ in Nature, Science	498	2.30
ARWU World Rank	100	2.08
ARWU Highly Cited Researchers	500	1.81
THE Teaching	799	1.63
THE Research	799	1.49
THE Overall	199	1.01
QS Overall	400	0.65
LEIDEN % Top Publications (Top 10%)	840	0.54
LEIDEN Mean Normalized Citation Score (MNCS)	840	0.46
QS Academic Reputation	400	0.43
QS Employer Reputation	400	0.36
QS Citations per Faculty	399	0.26
THE Citations	799	0.07

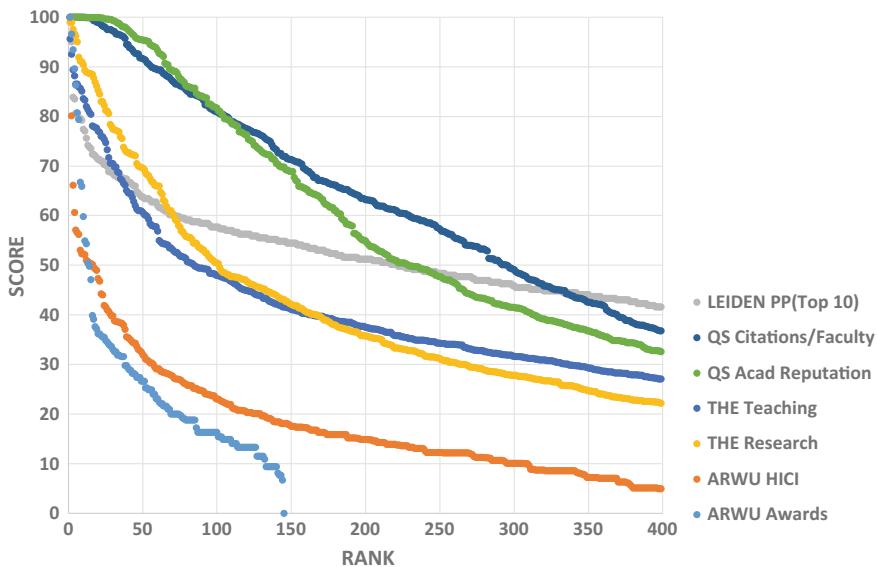


Fig. 18.2 Institutions' scores as a function of their ranks in 7 key indicator distributions. Legend: LEIDEN PP (Top 10): The percentage of publications among the top 10% most frequently cited articles published worldwide. THE Teaching: THE Teaching Performance; THE Research: THE Research Performance. ARWU HICI: ARWU Highly Cited Researchers

cited articles published worldwide—has the highest skewness, and the **THE** citations indicator the lowest. The latter result is not surprising, as Fig. 18.1 revealed already that the values obtained by this indicator are percentile ranks, for which the skewness is mathematically zero. Disregarding **Leiden** Number of Top Publications and **THE** Citations, the 5 **ARWU** indicators have the highest skewness, followed by 3 **THE** indicators, and 4 **QS** jointly with the two **Leiden** relative impact indicators the lowest.

18.5 Statistical Correlations

Tables 18.7–18.9 presents the Spearman coefficients (denoted as Rho) of the rank correlation between pairs of selected indicators, arranged into 3 groups: a group with pairs of seemingly identical indicators related to staff, student and funding data; citation-based indicators; and a group combining reputation- and recognition-based indicators with key indicators from the group of the citation-based measures. The correlations between two indicators are calculated for those institutions that have non-missing values for both measures. Row N gives the number of institutions involved in a calculation. Unless indicated differently, all correlations in Tables 18.7–18.9 are statistically significant at the $p = 0.001$ level.

Table 18.7 Spearman rank correlations between specific pairs of identical/very similar variables from different sources

Variable 1	Variable 2	Statistic	Score
QS Internat. students	THE % internat. students	Rho	0.87
		N	311
QS Faculty-Student Ratio	THE Student-Staff Ratio	Rho	-0.47
		N	289
QS Internat. Faculty	UMULTI Internat. Acad. Staff	Rho	0.13
		N	107
THE Industry Income	UMULTI Income from private sources	Rho	0.48
		N	201

Table 18.8 Spearman rank correlations between citation-based indicators

		LEIDEN MNCS	LEIDEN % Publ. in Top 10%	QS Citation per Faculty	THE Citations	UMULTI Top Cited Publ
ARWU Highly Cited Researchers	Rho	0.69	0.70	0.38	0.70	0.61
	N	468	468	308	416	461
LEIDEN MNCS (Mean Relative citation rate)	Rho		0.98	0.32	0.92	0.86
	N		840	344	589	742
LEIDEN % Publ. in Top 10% Most Cited Articles	Rho			0.34	0.92	0.89
	N			344	589	742
QS Citations per Faculty	Rho				0.38	0.26
	N				348	343
THE Citations	Rho					0.81
	N					620

Rank correlations above 0.8 are printed in bold, and those below 0.4 in italic. If one qualifies correlations with absolute values in the range 0.0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8 and 0.8–1.0 as ‘very weak’, ‘weak’, ‘moderate’, ‘strong’ and ‘very strong’, respectively, it can be said that correlations printed in bold-but-not-italic are very strong; correlations in bold and italic are weak or very weak, while all other are moderate or strong.

Unsurprisingly, a very strong correlation is found between an institution’s number of publications in the *ARWU* ranking and that in the *Leiden* Ranking ($\text{Rho} = 0.96$, $n = 468$), as both numbers are extracted from the Web of Science. On the other hand, the *ARWU* number of publications in Nature and Science correlates 0.73 with the *Leiden* (absolute) number of ‘top’ publications, suggesting that top publications are not merely published in these two journals.

Table 18.9 Spearman correlations between citation, reputation and teaching-related indicators

	ARWU Highly Cited Res	LEIDEN MNCS	THE Research	THE Teaching	QS Acad Reput	QS Citations/faculty	UMULTI quality Teaching
ARWU Awards	Rho N	0.43 500	0.50 468	0.46 416	0.47 416	0.45 314	0.30 308
ARWU Highly Cited Res	Rho N	0.69 468	0.64 416	0.60 416	0.53 314	0.38 308	0.22* 60
LEIDEN Mean Norm.	Rho N	0.60 589	0.54 589	0.41 349	0.32 349	0.36 344	82
Citation Score (MNCS)	Rho N		0.81 799	0.76 356	0.52 348	0.42 94	
THE Research	Rho N			0.76 356	0.50 348	0.43 94	
THE Teaching	Rho N				0.34 264	0.33* 53	
QS Academic Reputation	Rho N					0.47 29	
QS Citations per Faculty	Rho N						

*Not significant at $p = 0.05$

The most striking outcome in Table 18.7 is that the *QS* Faculty-Student Ratio correlates only moderately with the *THE* student-staff ratio ($\rho = -0.47$). From the data descriptions in the two systems it does not become clear why there are such large differences between the two. This is also true for the very weak correlation between *QS* International Faculty and *U-Multirank*'s International Academic Staff.

Noteworthy in Table 18.8 is first of all the very high correlation between the two *Leiden* citation impact measures ($\rho = 0.98$). Apparently, at the level of institutions it does not make a difference whether one focuses on the mean (MNCS) or the top of the citation distribution. Interestingly, also the *THE* Citation indicator shows a strong correlation with the *Leiden* impact measures. The description of this measure on the *THE* Ranking Methodology page (*THE* Ranking Methodology, n.d.) suggests that it is most similar if not identical to the *Leiden* MNCS, but a key difference is that it is based on Scopus, while the *Leiden* indicators are derived from the Web of Science. The *U-Multirank* indicator of top cited publications is provided by the Leiden Centre for Science and Technology Studies using the same methodology as that applied in the *Leiden* Ranking. The most remarkable outcome in Table 18.7 is perhaps that the indicator *QS* Citation per Faculty shows only a weak correlation with the other citation-based indicators. This result is further analyzed in Sect. 18.7 below.

Table 18.9 presents pairwise correlation coefficients between seven citation-, reputation- or teaching-related indicators. The only very strong rank correlation is that between *THE* Research and *THE* Teaching. Both measures are composite indicators in which the outcomes of a reputation survey constitute the major component. On the *THE* Ranking Methodology page it is unclear whether the reputation components in the two indicators are different. The very strong correlation between the two indicators suggests that these components are very similar if not identical.

The weak correlation between *QS* Citations per Faculty and other citation-based indicators has already been mentioned above. Table 18.9 shows that there is also a weak rank correlation inside the *QS* system between the citation and the academic reputation measure ($\rho = 0.34$). The major part of the pairs shows moderate or strong, positive Spearman correlation coefficients.

The *U-Multirank* Quality of Teaching score in Table 18.9 is calculated by the author of the current chapter, based on the outcomes of the survey among students, conducted by the *U-Multirank* team in 13 selected subject fields, and mentioned in Sect. 18.2. For institutions participating in at least two surveys, the performance classes (A-E) were quantified (A = 5, B = 4, etc.), and an average score was calculated over the subject fields. The number of cases involved in the calculation of the rank correlation coefficients between this indicator and other measures is relatively low, and the major part of the coefficients are not statistically significant at $p = 0.05$. Table 18.10 gives an overview of the 5 ranking systems.

Table 18.10 Overview of five information systems on the performance of higher education institutions

Aspect	ARWU World University Rankings 2015	CWTS Leiden Ranking 2016	QS World University Rankings 2015–2016	THE World University Rankings 2015–2016	U-Multirank 2016 Edition
Website	http://www.shanghairanking.com/ARWU2015.html	http://www.leidenranking.com/	http://www.opinuniversities.com/university-rankings	https://www.timeshighereducation.com/world-university-rankings	http://www.umultirank.org
Universities included	Every university that has any Nobel Laureates, Fields Medallists, Highly Cited Researchers, or papers published in Nature or Science, or significant amount of papers indexed by SCIE/SSCI. The best 500 are published on the web.	All 842 universities worldwide with more than 1000 fractionally counted Web of Science indexed core publications in the period 2011–2014 are included in the ranking.	918 universities are included	800 universities with at least 200 articles per year published in journals indexed in Scopus, and teaching at least undergraduates in each year during 2010–2014	In principle all higher education institutions can register for participation. The current version includes about 1300 institutions
Indicators/ dimensions and their weights	<ul style="list-style-type: none"> • Quality of Education Alumni (10%) Awards (20%) • Quality of Faculty Highly cited researchers (20%) Publ. in Nature, Science (20%) • Research output Publications (20%) • Per Capita Performance (10%) 	<ul style="list-style-type: none"> • Publication counts Articles in English, authored in core journals • Citation Impact Nr., % Top 1.10, 50% publications Mean Normalized Citation Rate • Collaboration Nr., % publ from different institutions Nr., % publ with geographical collab distance <100 or >5000 km 	<ul style="list-style-type: none"> • Academic Reputation (40%), based on QS survey • Employer Reputation, based on QS survey (10%) • Faculty Student Ratio (20%) • Citations per Faculty (20%) • International Students (10%) • International Faculty (10%) 	<ul style="list-style-type: none"> Performance indicators: • Teaching (30%), mainly based on reputation survey • International Outlook (7.5%) • Research (30%), mainly based on reputation survey • Citations (30%) • Industry Income (2.5%) 	Over 30 indicators covering the following main dimensions: <ul style="list-style-type: none"> • teaching and learning • research • knowledge transfer • international orientation • regional engagement Typical examples of indicators: Quality of teaching (based on survey); citation rate; income from regional sources; spin-offs
Data sources used	Databases on Nobel prizes and field medals; Thomson Reuters Web of Knowledge and Highly Cited researchers; data on academic staff from national agencies	All bibliometric data are extracted from Thomson Reuters' Web of Science	QS Academic Reputation Survey; self-reported data from universities; data from government and other agencies; bibliometric data from Elsevier's Scopus	THE Reputation Surveys; self-reported data from universities; bibliometric data from Elsevier's Scopus	U-Multirank student surveys; self-reported data from universities; bibliometric data of Science and PATSTAT database on patents

18.6 Secondary Analyses

18.6.1 Characteristics of National Academic Systems

A secondary analysis based on **THE** data examined for the 19 major countries with more than 10 institutions the rank correlation between **THE** Citations and **THE** Research Performance. According to the **THE** Ranking Methodology Page, the citation-based (research influence) indicator is defined as the number of times a university's published work is cited by scholars globally, compared with the number of citations a publication of similar type and subject is expected to have. **THE** Research Performance is a composite indicators based on three components: Outcomes of a Reputation Survey (weight(W) = 0.6); Research income (W = 0.2); and Research productivity (W = 0.2).

The results are presented in Fig. 18.3. Countries can be categorized into three groups. A first group with rho scores up or above 0.7 consists of four Anglo-Saxon countries, India and Switzerland. A second group, with scores between 0.4 and 0.6 contains four Asian countries and Spain. Finally, the group with scores below 0.4 includes four Western-European countries, Turkey and Russia, and also Brazil. As an illustration, Figs. 18.4 and 18.5 present a scatterplot representing the scores of the institutions in Italy and The Netherlands, respectively. In Italy, but also in Brazil and Russia, a large subset of universities has statistically similar Research Performance scores, but assumes a wide range of citation scores; at the same time, a few universities with high Research Performance scores have median or low citation scores. The Netherlands and Germany show a different, partly opposite pattern: a relatively large set of universities has similar, high citations scores, but reveals a

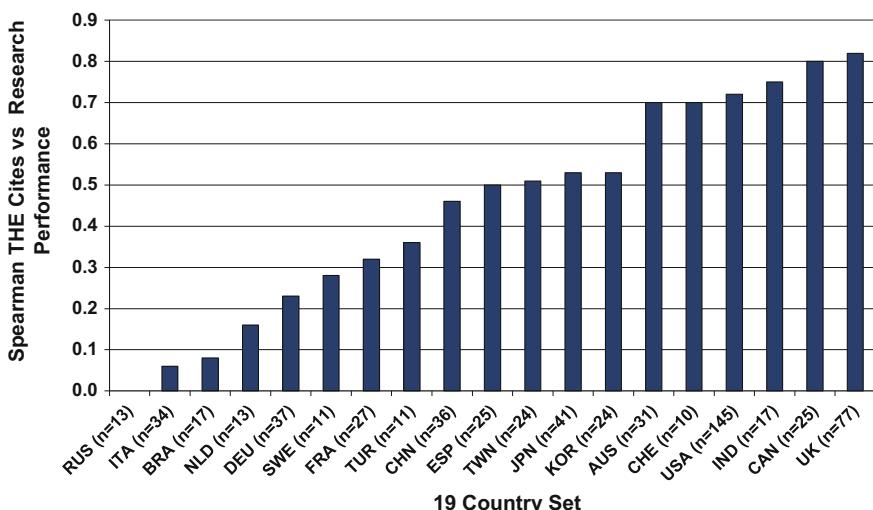


Fig. 18.3 Spearman rank correlation coefficient between Citations and Research Performance per country (*Data source THE Ranking 2016*)

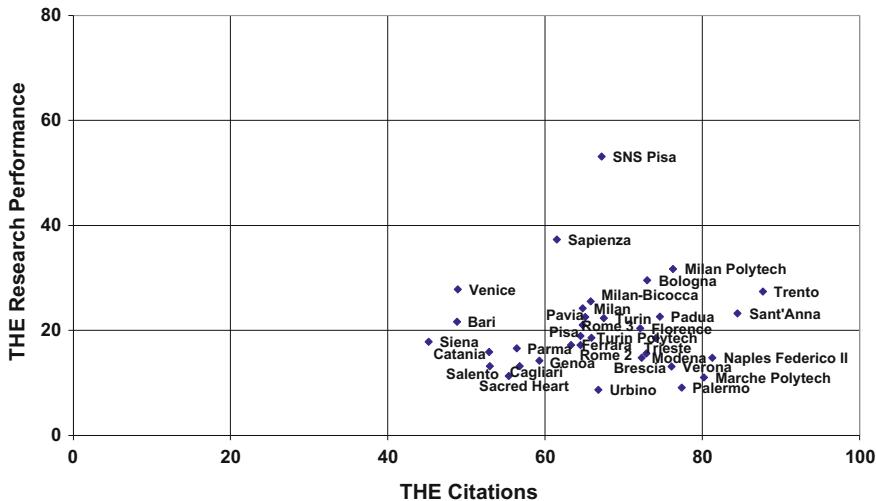


Fig. 18.4 Scatterplot of THE Research Performance versus THE Citations for Italy

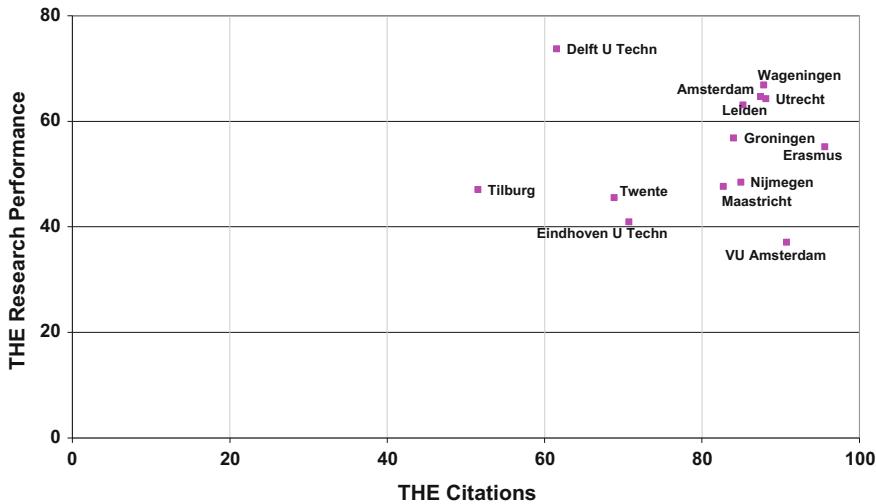


Fig. 18.5 Scatterplot of THE Research Performance versus THE Citations for The Netherlands

wide range of Research Performance scores. Both patterns result in low rank correlation coefficients.

The interpretation of the observed patterns is unclear. The figure suggests that there are differences among global geographical regions. A low correlation may reflect a certain degree of conservatism in the national academic system in the sense that academic reputation is based on performances from a distant past, and does not keep pace well enough with recent performances as reflected in citations.

18.6.2 QS Versus Leiden Citation-Based Indicators

Figure 18.6 plots for institutions in 6 countries the scores on the *QS* Citations Per Faculty indicator against the *Leiden* percentage of publications among the top 10% most frequently cited documents published worldwide. Both scores were expressed as percentile ranks by the current author. For details on the *QS* measure the reader is referred to *QS Normalization* (n.d.) and *QS Methodology* (n.d.) and on the *Leiden* indicators to *Leiden Indicators* (n.d.).

Five countries in Fig. 18.6 have institutions among the top 20% worldwide in the *QS* ranking, seemingly regardless of their citation scores on the *Leiden* indicator: Taiwan, Germany and The Netherlands have three institutions, China (including Hong Kong) six, and Canada two. This outcome raises the question whether the *QS* measure applies ‘regional weightings’ to correct for differences in citation counts between world regions, analogously to the application of regional weightings to counter discrepancies in response rates in the *QS Academic Reputation survey*. It must be noted that the current author could not find an explicit reference to such weightings in the *QS* document on normalization (*QS Normalization*, n.d.), although this document does indicate the use of weightings by scientific-scholarly discipline.

A second normalization of the *QS* measure calculates the ratio of citations and number of faculty. Interestingly, this leads to a negative correlation with the *Leiden* measure for Italy, The Netherlands, and, especially, for Germany, two institutions in which—*Humboldt University Berlin* and *University of Heidelberg*—have a *Leiden* percentile rank above 60 but a *QS* Citation per Faculty percentile rank below 20.

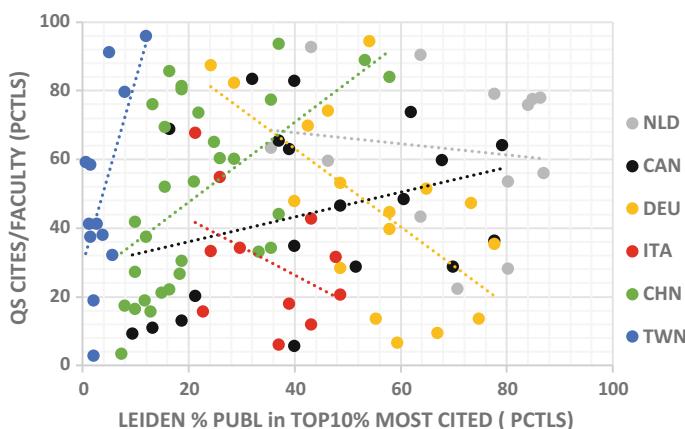


Fig. 18.6 QS and Leiden citation impact indicators for institutions in 6 selected countries

18.6.3 THE Research Performance Versus QS Academic Reputation

Figure 18.7. presents a scatterplot of the reputation-based **THE** Research Performance against **QS** Academic Reputation. As in the previous secondary analysis in this section, both measures were expressed as percentile ranks by the current author. The figure displays the names of the top 20 institutions with the largest, and the bottom 20 with the smallest difference between the **THE** and the **QS** measure, respectively. Focusing on countries appearing at least twice in a set, institutions in the top 20 set, for which the **THE** score is much larger than the **QS** score, are located in The Netherlands, Germany, USA and Taiwan, while universities in the bottom 20 set can be found in Chile, Italy, France and Japan.

These differences are probably caused by the fact that in the **QS** methodology ‘regional weightings are applied to counter any discrepancies in response rates’ (QS Normalization, n.d.), while **THE** does not apply such weighting. Hence, in the top 20 set one finds institutions from countries that have already a sufficient number of institutions in the upper part of the reputation ranking, and in the bottom 20 set universities in countries that are underrepresented in this segment. The outcomes then would suggest that Southern Europe and Northern Europe are considered distinct regions in the **QS** approach.

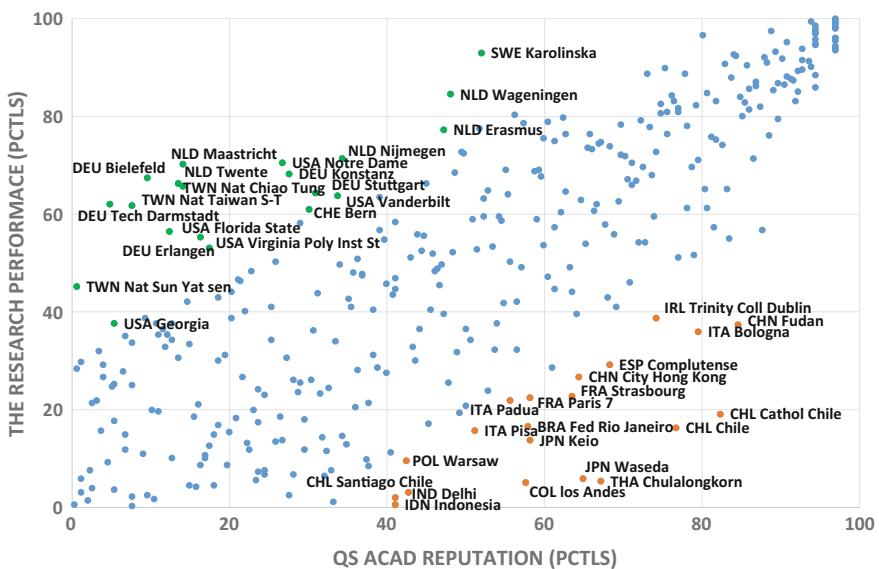


Fig. 18.7 Scatterplot of THE Research Performance versus QS Academic Reputation

18.6.4 ARWU Highly Cited Researchers Versus Leiden Top Publications Indicator

Figure 18.8 is constructed in a manner very similar to Fig. 18.7, but for two different indicators. It gives the names of the top 10 institutions with the largest difference, and the bottom 10 with the smallest difference between **ARWU** and **Leiden** measure. In the top 10 set two institutions from Saudi Arabia appear. Their score on the Highly Cited Researchers linked with these institutions indicator is much higher than ‘expected’ on the basis of the number of highly cited articles published from them.

This outcome illustrates a factor highlighted by Gingras (2014) who found in the Thomson Reuters List of Highly Cited Researchers—the data source of the **ARWU** indicator—a disproportionately large number of researchers linked with institutions in Saudi Arabia, mostly via their secondary affiliations, and who suggested that “by providing data on secondary affiliations, the list inadvertently confirms the traffic in institutional affiliations used to boost institutions’ places in world university rankings”. King Abdulaziz University, the institution Gingras found to be the most ‘attractive’ given the large number of researchers that indicated its name as secondary affiliation, is not in the Top 20 list, but it ranks 28th and would have been included in a top 30 list. The top 10 list includes six Japanese institutions. Whether their score on the **ARWU** Highly Cited Researchers indicator is caused by the same factor is as of yet unclear, and needs further investigation, without which *no* valid conclusions about these institutions can be drawn.

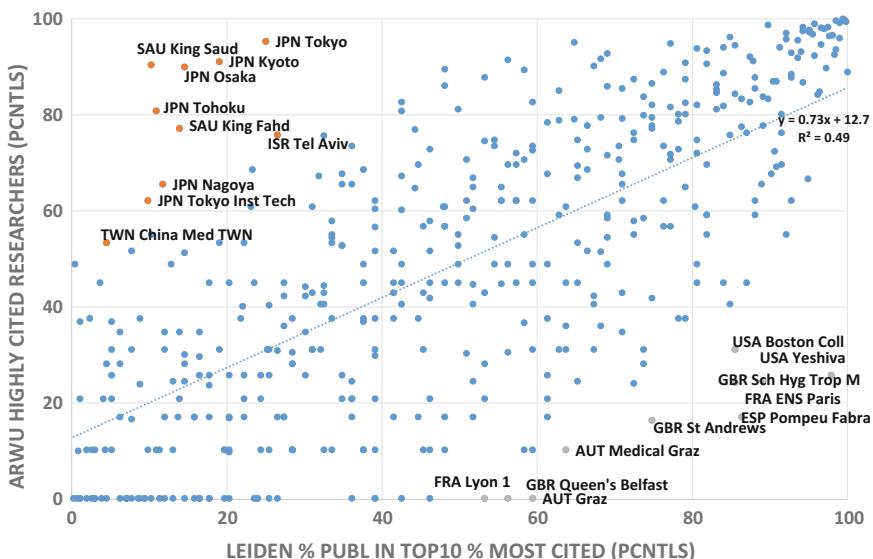


Fig. 18.8 Scatterplot of ARWU Highly Cited Researchers versus Leiden Top Publications indicator

The institutions and countries represented in the bottom 10 set constitute *prima facie* a rather heterogeneous set. However, it includes a number of institutions focusing on social sciences, or located in non-English speaking countries. This suggests that the **Leiden** indicator corrects more properly for differences between subject fields and native languages than the TR List of Highly Cited Researchers does.

It must be noted that the **ARWU** indicator is based on two lists of highly cited researchers, both compiled by Thomson Reuters, a first one in 2001, and a new one in 2013. The **ARWU** 2015 ranking is based on the sum of the numbers in the two lists. But the counts derived from the new list are based exclusively on the *primary* affiliation of the authors, thus substantially reducing the effect of secondary affiliations highlighted by Gingras.

18.7 Discussion and Conclusions

The *overlap* analysis clearly illustrates that there is *no* such set as ‘*the*’ top 100 universities in terms of excellence: it depends on the ranking system one uses which universities constitute the top 100. Only 35 institutions appear in the top 100 lists of all 5 systems, and the number of overlapping institutions per pair of systems ranges between 49 and 75. An implication is that national governments executing a science policy aimed to increase the number of academic institutions in the ‘top’ of the ranking of world universities, should not only indicate the range of the top segment (e.g., the top 100), but also specify which ranking(s) are used as a standard, and argue why these were selected from the wider pool of candidate world university rankings.

Although most systems claim to produce rankings of *world* universities, the analysis of *geographical coverage* reveals substantial differences between the systems as regards the distribution of covered institutions among geographical regions. It follows that the systems define the ‘world’ in different manners, and that—compared to the joint distribution of the 5 systems combined—each system has a proper orientation or bias, namely **U-Multirank** towards Europe, **ARWU** towards North America, **Leiden Ranking** towards emerging Asian countries, and **QS** and **THE** towards Anglo-Saxon countries.

Four entirely different methods were applied to construct *indicator scores* from raw data. **ARWU** and **QS** apply a normalization by the maximum, **THE** uses a percentile rank-based approach but for some indicators an exponential component was added, while **U-Multirank** calculates a distance to the median. This has severe implications for the *interpretation* of the scores. For instance, in the **THE** system 90% of institutions has a Research or Teaching Performance score below 55 or 50, respectively. This means that only a small fraction of institutions ‘profits’ in the overall ranking from a high score of these indicators, reflecting that the distribution of the actual values of the reputation-based component is much more skewed than that for the citation-based indicator. The distribution of **U-Multirank** performance

classes (A-E) among institutions varies substantially between indicators, and, as the definition of the classes is based on the distance to the median rather than on quintiles of a distribution, may strongly deviate from 20%.

ARWU indicators (Awards, Alumni, Articles in Nature and Science, Highly Cited Researchers, and Overall) show the largest *skewness* in their distributions, followed by **THE** indicators (Research and Teaching Performance, Overall), while **QS** indicators (Academic and Employer Reputation and Overall) jointly with the two **Leiden** relative citation impact indicators obtain the lowest skewness values. It follows that the degree of skewness measured in the various systems is substantially affected by the way in which the systems calculate the indicator scores from the raw data.

Several pairs of very similar if not identical indicators from *different* ranking systems rank-correlate only *moderately*, especially those based on student and faculty numbers. The causes of this lack of correlation are as yet unclear and must be clarified. It must be noted that in several systems the role of this type of data is far from being marginal. For instance, in the **QS** citation impact indicator an institution's number of academic staff constitutes the denominator in a citation-per-faculty ratio for that institution. Also, the question should be addressed whether *self-reported* data from institutions are sufficiently accurate to constitute an important factor in the calculation of indicators and rank positions. But even if data is obtained from statistical agencies such as national statistical offices, a thorough investigation is urgently needed as to whether such agencies apply the same definitions and categorizations in the data collection and reporting.

The citation-based indicators from **Leiden**, **THE**, **ARWU** and **U-Multirank** show strong or very strong rank correlations with one another, but correlate only weakly with the **QS** Citation per Faculty indicator. The latter is constructed differently in that an institution's total citation count, corrected for differences in citation levels between disciplines, is divided by the number of faculty employed in an institution. An analysis comparing **QS** and **Leiden** citation indicator scores may suggest that the **QS** citation measure does not only apply a *field* normalization, but also a normalization by *geographical region*, but more research is needed to validate this. The effect of indicator normalization is further discussed below.

A pairwise *correlation* analysis between seven citation-, reputation- or teaching-related indicators from the 5 systems shows for the major part of the pairs moderate or strong—but never very strong—, positive Spearman correlation coefficients (with values between 0.4 and 0.8). The conclusion is that these indicators are related to one another, but that at the same time a certain degree of *complementarity* exists among the various ranking systems, and that the degree of (dis-)similarity between indicators *within* a ranking system is similar to that between measures from *different* systems. The conclusion is that the various ranking methodologies do indeed measure different aspects. There is no single, ‘final’ or ‘perfect’ operationalization of academic excellence.

The analysis on the statistical relation between *two reputation-based indicators*, namely the **QS** Academic Reputation indicator, and the **THE** Research Performance measure, which is largely based on the outcomes of the **THE** reputation survey, reveals the effect of the use of ‘weightings’ to counter discrepancies or unbalances

upon the overall results. This particular case relates to (world) regional weightings. A ranking naturally directs the attention of users to its top, and multiple rankings to multiple tops. But what appears in the top very much depends upon which normalizations are carried out.

This analysis, as well as the analysis of the QS citation-per-faculty measure discussed above, provides an illustration of how the position of institutions in a ranking can be influenced by using proper, effective indicator normalizations. The current author does not wish to suggest that the developers intentionally added a normalization to boost particular sets of institutions or countries, as they provide in their methodological descriptions purely methodological considerations (QS Normalization, n.d.). But the two analyses clearly show how such targeted, effective boosting *could in principle* be achieved technically. When ranking systems calculate complex, weighted or normalized indicators —as they often do—, they should at the same time provide simple tools to show users the actual *effect* of their weightings or normalizations. Figures 18.7 and 18.8 in Sect. 18.7 illustrate how this could be done.

The analysis focusing on the number of highly cited researchers reveals possible traces of the effect of ‘secondary’ affiliations of authors in counting the number of highly cited researchers per institution. The **ARWU** team has already adjusted its methodology to counter this effect. But even if secondary affiliations are fully ignored, this indicator can be problematic in the assessment of an institution. How should one allocate (highly cited) researchers to institutions as researchers move from one institution to another—a notion that is properly expressed in the methodology along which **ARWU** calculates its Awards and the Alumni indicator. The analysis has identified other universities showing discrepancies similar to those of Saudi institutions, but the interpretation of this finding is as yet unclear. A general conclusion holds that by systematically comparing pairs of indicators within or across systems, discrepancies may be detected that ask for further study, and help evaluating the data quality and validity of indicators.

The analysis on the correlation between *academic reputation and citation impact* in the **THE** ranking (see Figs. 18.4 and 18.5) shows first of all that two-dimensional scatterplots for a subset of institutions with labelled data points provide a much more comprehensive view of the relative position of individual institutions than the view one obtains by scanning one or more rank lists sequentially from top to bottom. The outcomes of the analysis raise interesting questions. Why are there such large differences between countries as regards the correlation between the two types of indicators? What does it mean if one finds for a particular country that a large subset of institutions has statistically similar citation impact scores, but assumes a wide range of reputation-based scores, or vice versa?

The current author wishes to defend the position that ranking systems would be more useful if they would raise this type of questions, enable users to view the available empirical data that shed light on these questions, and in this way contribute to their knowledge on the pros and cons of the various types of indicators, rather than to scan sequentially through different rankings, or calculate composite indicators assigning weights to each constituent measure.

18.8 Concluding Remarks

Developers of world university ranking systems have made enormous progress during the past decade. Their systems are currently much more informative and user friendly than they were some 10 years ago. They do present a series of indicators, and institutions can be ranked by each of these separately. But the current interfaces hinder a user to obtain a comprehensive view. It is like looking into the outside world through a few vertical splits in a fence, one at the time. In this sense, these systems are still one-dimensional. A system should not merely present a series of separate rankings in parallel, but rather a dataset and tools to observe patterns in multi-faceted data. The simple two dimensional scatterplots—to which easily a third dimension can be added by varying the shape of the data point markers—are good examples.

Through the selection of institutions covered, the definition of how to derive ratings from raw data, the choice of indicators and the application of normalization or weighting methodologies, a ranking system distinguishes itself from other rankings. Each system has its proper orientation or ‘profile’, and there is no ‘perfect’ system. To enhance the level of understanding and adequacy of interpretation of a system’s outcomes, more insight is to be provided to users into the differences between the various systems, especially on how their orientations influence the ranking positions of given institutions. The current chapter has made a contribution to such insight.

Acknowledgements The author wishes to thank two referees of the journal *Scientometrics* for their useful comments on an earlier version of this paper. The author is also grateful to the members of the Nucleo di Valutazione of the Sapienza University of Rome for stimulating discussions about the interpretation and the policy significance of world university rankings.

Chapter 19

Comparing Full Text Downloads and Citations

Abstract This Chapter presents a statistical analysis of full text downloads of articles in Elsevier's ScienceDirect covering all disciplines. It reveals large differences in download frequencies, their skewness, and their correlation citation counts, between disciplines, journals, and document types.

Keywords Corrected paginated proof · Obliteration by incorporation · Obsolescence patterns · Online publication date

19.1 Introduction

In the past decade, many scientific literature publishers have implemented usage monitoring systems based on data including clickstreams, downloads and views of scholarly publications recorded on an article level, that allow them to capture the number of times articles are downloaded in their PDF or HTML formats. This type of data is not only used by publishers as a way to monitor the usage of their journals but also by libraries who wish to monitor and manage the usage of their collections (Duy & Vaughan, 2006). The growing need for this type of monitoring resulted in the launch of COUNTER (Counting Online Usage of Networked Electronic Resources), an international initiative which aimed to set standards and facilitate the recording and reporting of online usage statistics in a consistent, credible and compatible way. Nowadays, COUNTER is an industry standard, used by most publishers and libraries and allows for downloads data to be analyzed and compared more easily by subscribers and publishers alike. This development could be one of the reasons that research in this area has seen such significant growth.

Research on the relationships between citations and downloads has expanded in various studies attempting to understand the relationship between the two as usage phenomenon and as a way to measure research impact. (e.g., Schloegl and Gorraiz, 2011; Gorraiz, Gumpenberger & Schloegl, 2014; Guerrero-Bote & Moya-Anegón,

This chapter is largely based on Halevi & Moed (2014b) and on Moed & Halevi (2016).

2014). Kurtz et al. (2005a; 2005b) published two pioneering papers analyzing usage mainly of the NASA Astrophysics Data System (ADS), and comparing the number of electronic accesses—which they term “reads”—individual articles in astronomy and astrophysics journals with citation counts.

In their review article published in 2010, Michael Kurtz and Johan Bollen describe “Usage Bibliometrics” as the statistical analysis of how researchers access their technical literature, based on the records that electronic libraries keep of every user transaction (Kurtz & Bollen, 2010). They underline that many “classical”, citation-based measures have direct analogs with usage, and that an important approach to validation of usage statistics is to demonstrate the similarities and differences between citation and usage statistics. An important class of usage statistics is based on the number of times articles from publication archives are downloaded in full text format, denoted as “downloads” below. Kurtz and Bollen claim that “the relation between usage and citation has not been convincingly established” (p. 23) and that “direct comparisons over the same set of input documents are rare” (p. 23).

In 2005, Moed published an analysis of the statistical relationship between citations and full text article downloads for articles in one particular journal: Tetrahedron Letters (Moed, 2005b). A main objective of the current chapter is to expand the analyses presented in the 2005 article in the following ways:

- Analyze a much larger set of journals covering all domains of science and scholarship.
- Analyze in more detail download patterns as a function of time;
- Examine the statistical correlation between downloads and citations both at the level of journals and of individual articles;

A full discussion, interpretations of the new findings and their positioning within the framework of the review article by Kurtz & Bollen (2010) will be given in a full article to be published in a later phase. The base assumption underlying this chapter is that a sound statistical analysis of relationship between downloads and citations, and a thorough reflection upon its outcomes, contributes to a better understanding of what both download counts and citation counts measure, or more generally, to more insight into information retrieval, reading, and referencing practices in scientific-scholarly research. It is the very combination of the two types of data that enlarges so to speak the horizon, and provides a perspective in which each of the two types can be positioned. In the quantitative study of research activity and performance, downloads and citations provide complementary data sources. In this article the term “usage” is reserved for the use made of electronic publication archives in the broadest sense, and recorded in the archive’s electronic log files. It includes activities such as downloading in pdf, viewing in html format, browsing through abstracts, and also saving, sharing or annotating documents in reference managers.

19.2 Data Collection

One of the main challenges of analyzing downloads and citations figures lays in the availability and completeness of the data collected. The database used to collect the data, whether citations or downloads, might be incomplete. Thus, for example, downloads collected for Scopus™ covered journals, might not be representative of usage in general, because not all literature searches use Scopus™ as their platform of choice. In addition, Scopus™ citations are biased by incomplete source coverage as complete citations are only available from 1996 forward which is a well-documented limitation of the database. Unlike Scopus, ScienceDirect™ is a very specific source of full text articles which is mostly used to either view or download content. Therefore, usage data is fairly complete in ScienceDirect™.

Downloads versus citations examined in this chapter were aggregated in 3 levels: (1) database (e.g., all ScienceDirect™ articles); (2) journal; (3) individual article. The data was collected in two sets of citations and usage data; one at the level of journals and the second at the level of individual articles.

- Journal Level Data: the first set of data contained all 20,000 peer-reviewed journals covered in Scopus™. For each journal, citation counts for the years 2004–2010 were aggregated per year and per journal. Download counts were aggregated for all 2500 journals covered both in Scopus™ and ScienceDirect™, Elsevier full text database per year and per journal.
- Document Level Data: Citations and counts on a per document basis were collected for all individual document published in 63 ScienceDirect™ journals between 2008 and 2012 covering all domains of science and scholarship. Downloads and citations counts on document level are up to September 2013.

It must be noted that the journals studied are not a random sample from the set of journals in ScienceDirect™. The aim of the selection was to include journals from different disciplines and cover all major disciplines, in order to study differences among disciplines, and also to include journals that were originally sections of one and the same “parent” journal, so that one could even obtain indications of differences within a journal (Table 19.1).

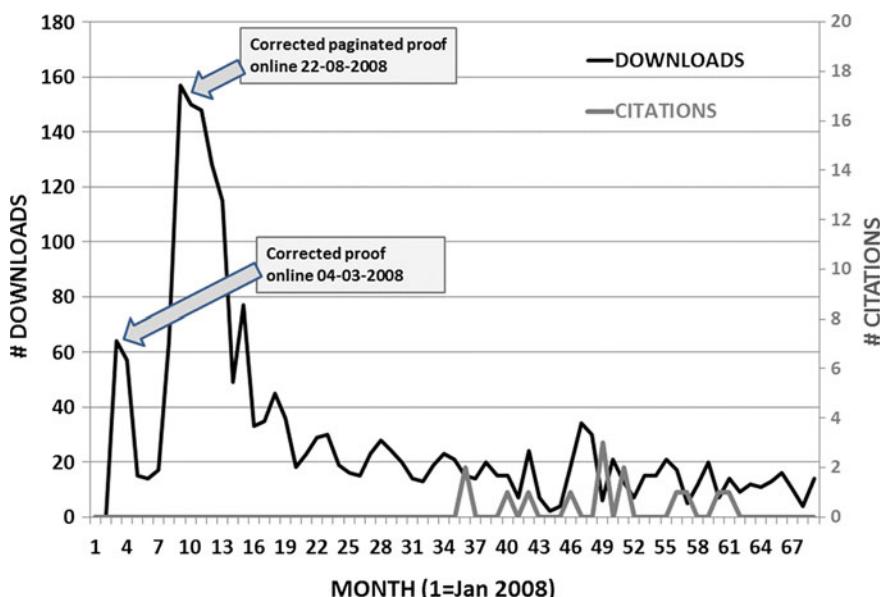
19.3 Results

19.3.1 Downloads Versus Citations of an Individual Article

Figure 19.1 presents for one particular article the number of downloads shown on the left vertical axis and the number of citations shown on the vertical right axis over each month after publication. The article is taken from the Journal of Wind Engineering and Industrial Aerodynamics. Downloading of an article from ScienceDirect is technically possible when the final version of the manuscript

Table 19.1 List of journals analyzed in this chapter

• Annals of Pure and Applied Logic	• Molecular Oncology
• Applied Clay Science	• Ophthalmology
• Applied Surface Science	• Performance Evaluation
• Biochimica et Biophysica Acta - Bioenergetics	• Physica A: Statistical Mechanics and its Applications
• Biochimica et Biophysica Acta - Biomembranes	• Physica B: Condensed Matter
• Biochimica et Biophysica Acta - Gene Regulatory Mechanisms	• Physica C: Superconductivity and its Applications
• Biochimica et Biophysica Acta - General Subjects	• Physica D: Nonlinear Phenomena
• Biochimica et Biophysica Acta - Molecular and Cell Biology of Lipids	• Physica E: Low-Dimensional Systems and Nanostructures
• Biochimica et Biophysica Acta - Molecular Basis of Disease	• Phytochemistry
• Biochimica et Biophysica Acta - Molecular Cell Research	• Phytochemistry Letters
• Biochimica et Biophysica Acta - Proteins and Proteomics	• Plant Physiology and Biochemistry
• Biochimica et Biophysica Acta - Reviews on Cancer	• Plant Science
• Bioinorganic Chemistry	• Poetics
• Cancer Letters	• Powder Technology
• Differential Geometry and its Application	• Stem Cell Research
• Earth and Planetary Sciences Letters	• Surface Science
• European Journal of Cancer	• Tectonophysics
• Fuzzy Sets and Systems	• Tetrahedron Letters
• Journal of Applied Geophysics	• Thin Solid Films
• Journal of Cultural Heritage	• Topology and its Applications
• Journal of Dentistry	• Trends in Plant Science
• Journal of Econometrics	• Water Research
• Journal of Economics and Business	• Journal of Science and Medicine in Sport
• Journal of Hydrology	• Applied Ergonomics
• Journal of Informetrics	• design studies
• Journal of International Economics	• Journal of Historical Geography
• Journal of Logic and Algebraic Programming	• Journal of Phonetics
• Journal of Medieval History	• Child Abuse and Neglect
• Journal of Wind Engineering and Industrial Aerodynamics	• Behavior Therapy
• Limnologica	
• Lingua	
• Materials Science & Engineering A: Structural Materials: Properties, Microstructure and Processing	
• Materials Science & Engineering B: Solid-State Materials for Advanced Technology	
• Materials Science and Engineering C	

**Fig. 19.1** Longitudinal download and citation counts for an individual article

corrected by the authors is made available online. The date at which this occurs is the online publication date. It is important to note the different phases of publication i.e. corrected proof and corrected paginated proof as they are seen to generate different downloads patterns. As can be expected, the corrected proof which became available in March 2008 generated over 60 downloads followed by over 150 downloads when it was paginated in August 2008. At this date, the journal issue in which an article appears is complete, all its documents are online, and the downloading of its articles boosts in the early periods in the article's age which generate the highest downloads figures in an article's life cycle. It could be explained as current awareness activity when readers keeping abreast in their area of research and closely reading content as it becomes available. Variations between the 10th and 40th month are probably due to seasonal influences and academic life cycles.

At this point, no citations are recorded which is expected for a 4 months old article. The first citations of this article appear in the following year, approximately July 2009. These citations can be assumed to be at least partially a result of the article's early heavy downloads followed by steady downloads rate in the months followed. However, it is also important to observe how citations might affect the article's downloads on the periods following the appearance of the first citation. Figure 19.1 shows that downloads rate is increasing in close proximity to the first citation during months 37–40 followed by an additional peak appearing after citations are recorded in months 40–43. These download peaks may be the result of citations, as the latter increase an article's visibility. Although earlier papers (e.g., Moed, 2005b) provide evidence that citations may have a positive effect upon downloads, causality in the relationships between downloads and citations are not further investigated in the current article.

The pattern shown in Fig. 19.1 is a common pattern that can be observed for the overwhelming part of documents analyzed in the current article: full text downloading starts when the corrected proof is online; next, usage increases strongly when the article is paginated, followed by a rapid decline—although it should be noted that the time period between the article's online publication date and the decline phase of its full text downloads varies across journals and disciplines (see Fig. 19.3) and document types (see Fig. 19.4); next, influences of seasonal and academic cycles are visible in the decline period; and finally the monthly number of downloads shows a revival when the article is cited. It should be noted, however, that the time period between an article's online publication date and its first received citation depends upon the journal, discipline and document type as well, and may be much shorter than that of the article represented in Fig. 19.1.

19.3.2 Downloads by User Institution

Figure 19.2 presents data on monthly full text downloads from ScienceDirect that users from 3 academic institutions made between January 2008 and May 2013. The

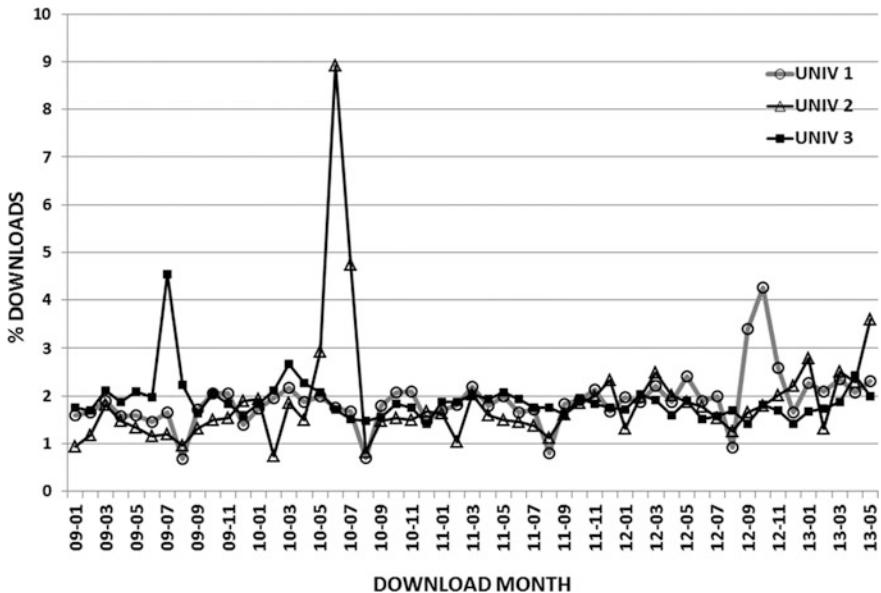


Fig. 19.2 Longitudinal download counts for three user institutions. The vertical axis gives the percentage of downloads in a month, relative to an institution's sum of downloads during the total time period. For University 2 the actual percentage of downloads in July 2010 is 9%, which is 4.5 times the level one would find if the number of an institution's downloads would be constant over time

data show a clear peaky behaviour. University 1 represented in Fig. 19.2 participated in a national research assessment exercise, in which research staff members could submit full text PDF downloads of their best articles to an evaluation agency for assessment by an expert panel, with a submission deadline in October 2012. For the peaks of Institutions 2 and 3 no explanation is available as of yet. Whether or not these peaks are caused by bulk downloading can be examined by grouping the downloaded articles by user session and by journal volume and issue, and determining the number of downloads per session, journal volume or issue. The three institutions were selected as they provide good illustrations of peaky usage behaviour. In a follow-up study the frequency at which this type of behavior occurs across all user institutions will be further analyzed.

19.3.3 Downloads Time Series Per Journal and Document Type

Figure 19.3 shows the average number of downloads per full length article for journals in social, applied, life, clinical medicine, mathematics and humanities

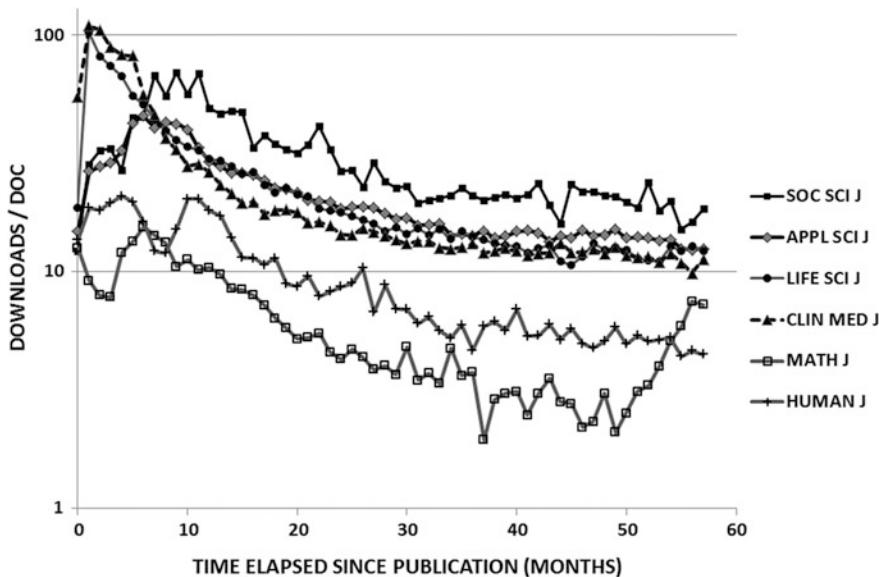


Fig. 19.3 The number of downloads per full length article as a function of the articles' age for 6 journals. The journals cover the subject fields of Social Sciences (SOC SCI), Applied Sciences (APPL SCI), Life Sciences (LIFE SCI), Clinical Medicine (CLIN MED), Mathematics (MATH) and Humanities (HUMAN), respectively. AGE = 1 indicates the months in which the articles were published

sciences over time. The overall phenomenon seen in Fig. 19.3 is that all journals display peak downloads in the 1st months following publications, despite the difference in the amount of downloads which varies considerably between journals. Yet, there are differences among the represented journals in the month in which download counts peak. For instance, for the clinical medicine and life sciences journal downloads peak 1 month after the month in which they were published online, whereas for the applied science and the mathematics journal in the 7th month. Moreover, large differences exist in the decline rates in the various journals. These decline rates themselves tend to decline as the documents grow older. This is consistent with the two-factor models explored by Moed (2005b), and the four-factor models explored by Kurtz et al. (2005b).

Figure 19.4 displays the development of downloads over time for four document types in the set of 63 journals: full length articles (FLA), reviews (REV), short communications (SCO) and editorials (EDI). As can be seen in the graph, reviews, short communications and editorials reach their peak downloads in the 1st month after publication, and full length articles in the 3rd month. Short communications and editorials show the most rapid decline during the first and 24th month after publication. After 2 years, the decline rates of the four types are similar. The level of downloads is highest for reviews, and lowest for editorials, at least in the set of 63 journals analyzed in this section.

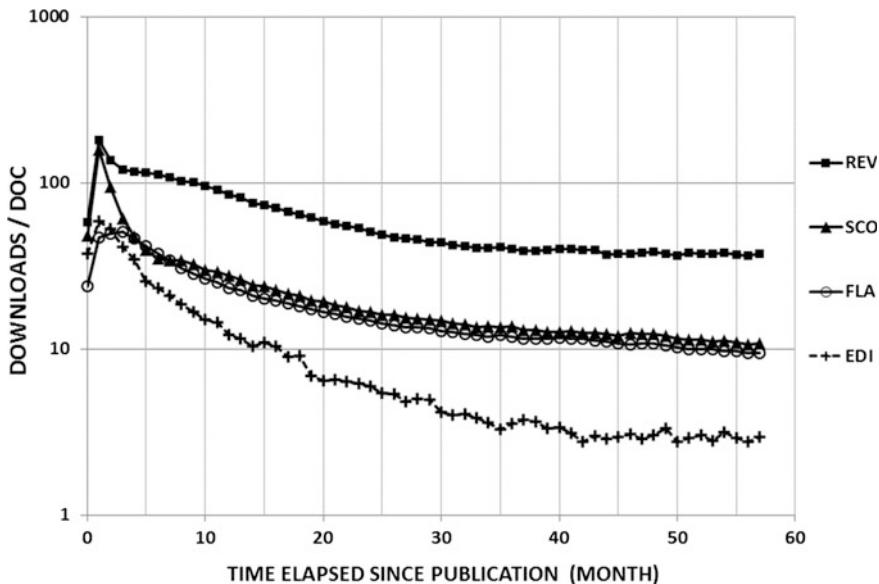


Fig. 19.4 The number of downloads per document type as a function of the documents' age. Data are shown for 4 document types published in the 63 journal set: full length articles (FLA), reviews (REV), short communications (SCO) and editorials (EDI)

19.3.4 Download-Versus-Citation Ratios

Adopting a diachronous approach, Fig. 19.5 presents for documents published during 2008–2009 the ratio of the number of downloads and citations as a function of the documents' age, or, in other words, of the time elapsed since their publication date, expressed in months. In this figure the documents from all journals in the 63 Journal Set are aggregated into one “super” journal. Ratios of downloads and citations are calculated for four types of documents: editorials, full length articles, short communications and reviews. Figure 19.5 clearly shows that the ratio of downloads and citations very much depends upon the type of document and upon the time elapsed since their publication date. For full length articles, reviews and short communications this ratio reaches a value of about 100 after 45 months.

Figure 19.6, however, shows large differences in this ratio among the 63 journals. It displays on the vertical axis the ratio of downloads and citations for the aggregate of full length articles published in the 63 Journal Set, and on the vertical axis the ratio of the skewness values of the download and citation article distribution, respectively, further discussed in the next section. Each symbol represents a particular journal. Distinct symbols indicate the main discipline covered by a journal. Figure 19.6 shows that journals in social sciences and humanities tend to have large downloads ratios versus citations, and several mathematics periodicals relatively low ratios. Clinical medicine journals show large variations.

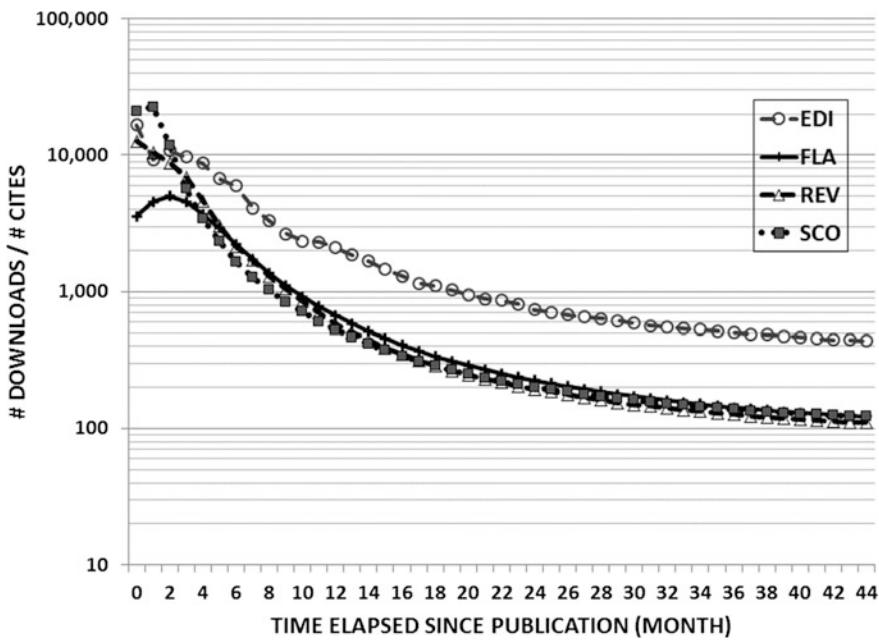


Fig. 19.5 Ratio of downloads and citations of documents as a function of their age (63 Journal Set). *EDI* Editorials; *FLA* Full Length Article; *REV* Review; *SCO* Short Communications

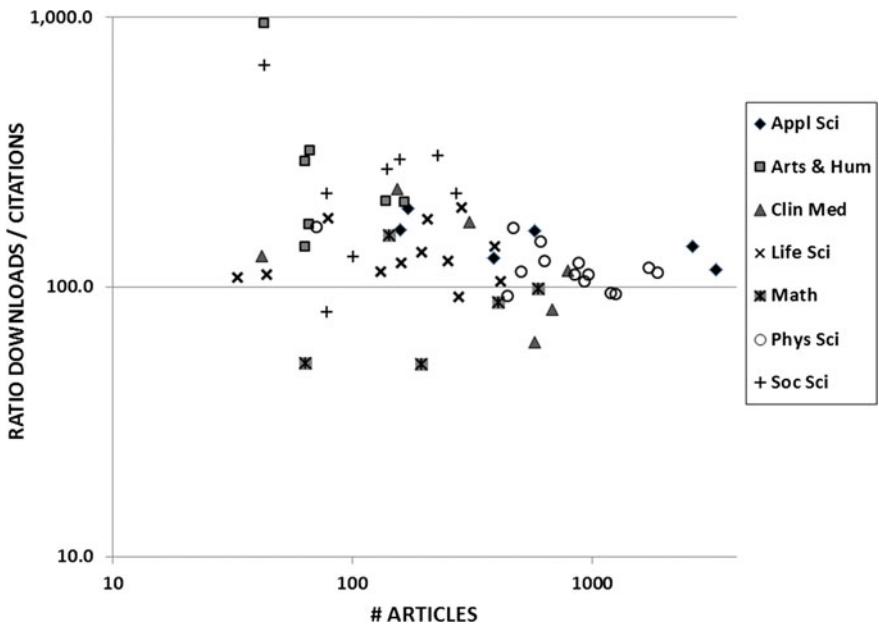


Fig. 19.6 Ratio of mean and skewness of the article download and citation distribution for 63 journals set (full length articles (FLA) only)

19.3.5 Statistical Correlations Between Downloads and Citations at the Journal and Article Level

Figure 19.7 presents an analysis at the journal level. It is based on download counts in the year of publication and citations in the third year after publication and shows the Pearson correlations per discipline. Spearman rank coefficients per discipline tend to be somewhat lower than the Pearson values, due to the skewness of the underlying distributions, but the overall picture presented in Fig. 19.7 does not change if the former type is plotted rather than the latter. Analyzing the correlation per discipline between a journal's average number of downloads per article against the number of cites per article, Fig. 19.7 shows that in the areas of biochemistry & molecular biology, neuroscience and veterinary sciences downloads and citations are highly correlated followed by chemical engineering, pharmacology and immunology. Disciplines which display the lowest correlation coefficients between downloads and citations are arts & humanities and health professions. The factors responsible for these differences in correlation must be further studied. For instance, the low correlation in Arts & Humanities may be due to the fact that the citation database used does not cover the publication output in this domain sufficiently well, and particularly misses citations in and to books.

Scatterplot of downloads versus citation counts of articles in an applied science journal. The diagonal represents the linear regression line. It shows that the articles that are frequently downloaded (tentatively defined as those with more than 2000

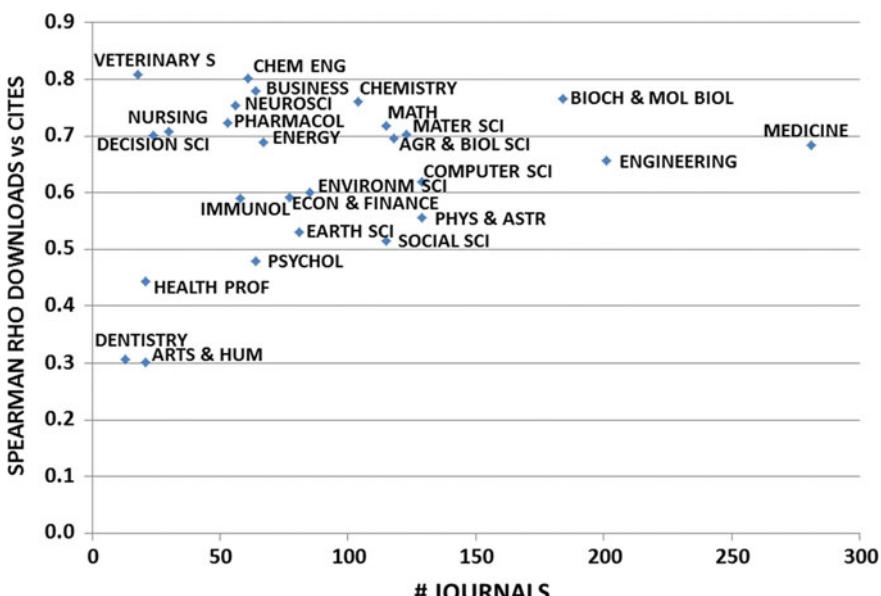


Fig. 19.7 Correlation between downloads and citations at the journal level by discipline

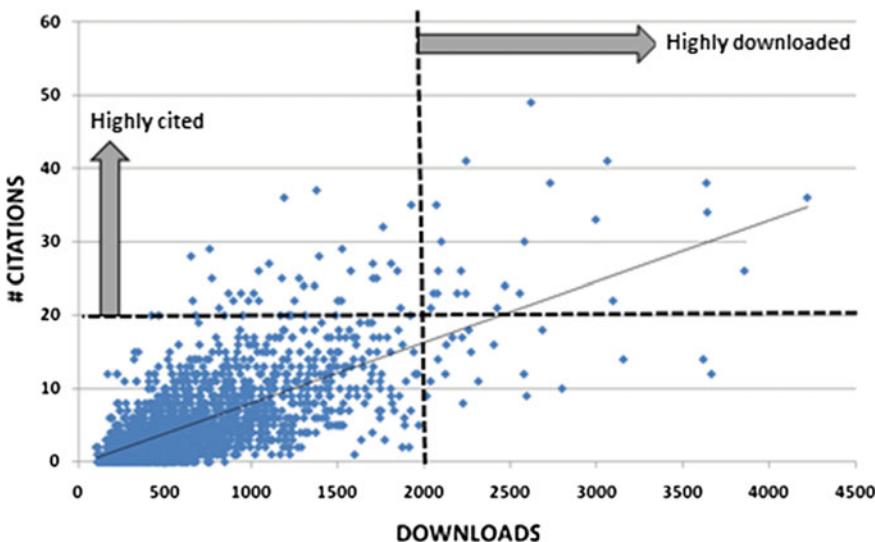


Fig. 19.8 Downloads versus citation counts for a journal in applied sciences

downloads) almost all have a minimum citation count of about 10. In other words, among the articles cited less than 10 times, there are no highly downloaded articles. This is so to speak one side of the correlation coin. But apart from this observation, the citation counts of the highly downloaded articles show a strong scatter. Such a scatter is even more clearly visible among the download counts for articles that are highly cited (tentatively, more than 20 times). But all these highly cited articles have a download rate that exceeds 500 (Fig. 19.8).

19.4 Discussion and Conclusions

19.4.1 Analyses by User Country and Institution

The fact that seasonal and academic cycles are reflected in longitudinal download patterns is not surprising. What is of interest is the peaky behavior at the level of user institutions, and the apparent lack in many cases of solid explanations for such behavior. Even if the overall contribution of number of downloads made in peak months across institutions is perhaps only a few per cent of the total number of downloads, more understanding of the cause of outliers is desirable. A combined qualitative-quantitative approach is the most promising, in which interviews with librarians at institutions is complemented with a more detailed analysis of the underlying usage patterns. Typical questions that should be addressed are: is downloading in peak months a form of bulk downloading, in which large numbers

of documents are downloaded issue by issue, journal by journal, in a single user session. Bulk sessions can be identified, for instance, by calculating the average number of downloads per used journal in a session. This parameter tends to obtain extremely high values if complete journal issues or (annual) volumes of a journal are downloaded article-by-article in one single user session.

19.4.2 Downloads Time Series Per Journal and Document Type

Perhaps the main observation of the outcomes presented in this article is that they show such large differences among journals, subject fields, and types of document. It must be underlined again that the journals studied are not a random sample from the total population of journals in ScienceDirect. The aim of the selection was to include journals from different disciplines and cover all major disciplines and include sectionalized journals as well. Our outcomes thus show how large the variability across journals and subject fields can be. The analyses at the journal level presented in the current chapter show that, adopting a diachronous approach, during the first 4 years after online publication date, all journals show a delay in downloads, in the sense that the average number of downloads per month increases after the month of publication and reaches its peak after 2–8 months, depending upon the journal. Such a behavior is qualitatively similar to that of citation obsolescence: both processes show a delay.

Moed (2005b) used in a diachronous approach a two factor model based on monthly rather than annual usage counts. Although the model showed a reasonable fit when applied in 2005 to Tetrahedron Letters, a journal publishing on a monthly basis short communications with a relatively short life cycle, download obsolescence patterns per journal reveal that a two factor model tends to be inappropriate. Full length articles, reviews, short communications and editorial have different download obsolescence patterns; their differences are similar to those found for citations. The ratio of the number of the number of downloads per review to that per article is similar to the same ratio for citations. And short communications mature more quickly than full length articles both in terms of downloads and citations.

19.4.3 Download-Versus-Citation Ratios

Findings in this chapter illustrate that the actual ratio of downloads and citations strongly depends upon the age of the used articles. It must be noted, however, that the rate of decline decreases over time, and that the value of downloads per citation ratio stabilizes to some extent after 3 years or so to a value of approximately 100. The conclusion is that, after 4 years following the online publication date, the

number of downloads of the articles in a journal is two orders of magnitude higher than the number of citations. This result applies both for full length articles, reviews and short communications. For editorials, however, the ratio is a factor of 2 higher than it is for the other document types.

19.4.4 Statistical Correlation Between Downloads and Citations

Large differences in the degree of linear correlation were found among subject fields at the journal level, the Pearson correlation coefficients varied between around 0.3 in the humanities to 0.9 in molecular biology. Intuitively one might conjecture that subject fields in which the correlation is high tend to be very specialized fields, such as molecular biology and biochemistry, in which the main users or readers of publications are the researchers active in that field, in other words, fields in which the author and the reader populations tend to coincide. Fields in which the reader population is probably much wider than the research community—including for instance interested readers from other disciplines of publications made by humanities and social science researchers, or practitioners (engineers or nurses) using technical information from engineering and nursing journals—the correlation is lower. But the analysis did not define or measure more precisely the degree of overlap between author and user population, so that rigorous testing of the hypothesis that the degree of correlation between downloads and citation counts is positively related to this overlap, has not been carried out, due to a lack of information about the user or reader population.

19.4.5 Factors Differentiating Between Download and Citations

Despite the fact that in all journals analysed in the study download and citation counts per article positively correlate, the following factors differentiate between downloads and citations.

- *Usage leak.* Not all full text uses of a publisher archive's documents may be recorded in the archive's log files.
- *Citation leak.* Not all relevant sources of citations may be covered by the database in which citations are counted.
- *Downloading* the full text of a document does not necessarily mean that it is fully *read*.
- *Reading and citing populations may be different.* For instance, industrial researchers may read scientific papers but not cite them as they do not publish papers themselves.

- *Number of downloads depends upon type of document.* For instance, editorials and news items may be heavily downloaded but poorly cited compared to full length articles.
- *Downloads and citations show different obsolescence functions.* Download and citation counts both vary over time, but in a different manner, showing different maturing and decline rates.
- *Downloads and citations measure different aspects.* Short term downloads tend to measure readers' awareness or attention whereas citations result from authors' reflection upon relevance.
- *Downloads and citations may influence one another in multiple ways.* More downloads may lead to more citations. But the reverse may be true as well. Articles may gain attention and be downloaded because they are cited.
- *Download counts are more sensitive to manipulation.* While citations tend to be regulated by the peer review process, download counts are more sensitive to manipulation.
- *Citations are public, usage is private.* While citations in research articles in the open, peer reviewed literature are public acts, downloading documents from publication archives is essentially a private act.

References

- Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics*, 101, 1129–1144.
- Abramo, G., & D'Angelo, C. A. (2016). A farewell to the MNCS and like size-independent indicators Original Research Article. *Journal of Informetrics*, 10, 646–651.
- Acharya, A., Verstak, A., Suzuki, H., Henderson, S., Iakhiaev, M., Chiung, C., Lin, Y., et al. (2014). Rise of the Rest: The Growing Impact of Non-Elite Journals. <http://arxiv.org/pdf/1410.2217v1.pdf>.
- Ackers, L. (2005). Moving people and knowledge: Scientific mobility in the European Union1. *International Migration*, 43, 99–131.
- Adie, E., & Roe, W. (2013). Altmetric: Enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26, 11–17.
- Adler, R., Ewing, J., & Taylor, P. (2008). *Citation statistics. A report from the International Mathematical Union*. www.mathunion.org/publications/report/citationstatistics0.
- Aguillo, I. F., Granadino, B., Orteag, J. L., & Prieto, J. A. (2006). Scientific research activity and communication measured with cybermetric indicators. *Journal of the American Society for the Information Science and Technology*, 57, 1296–1302.
- Albert, M. B., Avery, D., Narin, F., & MacAllister, P. (1991). Direct validation of citation counts as indicators of industrially important patents. *Research Policy*, 20, 251–259.
- Alberts, B. (2013). Impact factor distortions. *Science*, 340, 787.
- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web. Methodological approaches to ‘webometrics’. *Journal of Documentation*, 53, 404–426.
- Altmetric.com (2014). www.altmetric.com.
- Anonymous, Nature. (2005). Ratings games. Editorial. *Nature*, 436, 889–890.
- AUBR (2010). Assessment of University-Based Research Expert Group (AUBR). *Assessing Europe's University-Based Research* (p. 151). K1-NA-24187-EN-N, European Commission, Brussels. <http://ec.europa.eu/research/era/docs/en/areas-of-actions-universities-assessing-europeuniversity-based-research-2010-en.pdf>.
- ARWU (2016). *World University Ranking 2016*. <http://www.shanghairanking.com/ARWU2016.html>.
- Bar-Ilan, J. (2014, June). Evaluating the individual researcher—adding an altmetric perspective. *Research Trends*, 37. <https://www.researchtrends.com/issue-37-june-2014/evaluating-the-individual-researcher/>.
- Barjak, F., Li, X., & Thelwall, M. (2007). Which factors explain the web impact of scientists' personal homepages? *Journal of the American Society for Information Science and Technology*, 58, 200–211.
- Becker Model (n.d.). <https://becker.wustl.edu/impact-assessment>.
- Benedictus, R., & Miedema, F. (2016). Fewer numbers, better science. *Nature*, 538, 453–455.
- Bishop, D. (2013). *The Matthew effect and REF2014*. <http://deevybee.blogspot.lt/2013/10/the-matthew-effect-and-ref2014.html>.

- Björk, B.-C., & Solomon, D. (2012). Open access versus subscription journals: A comparison of scientific impact. *BMC Medicine*, 2012(10), 73.
- Björk, B.-C., & Paetau, P. (2012). Open access to the scientific journal literature—status and challenges for the information systems community. *Bulletin of the American Society for Information Science and Technology*, 38, 39–44.
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55, 1216–1227.
- Bollen, J., & Van De Sompel, H. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59, 136–149.
- Bonacorsi, A., & Daraio, C. (2003). Age effects in scientific productivity. The case of the Italian National Research Council (CNR). *Scientometrics*, 58(1), 47–88.
- Bonacorsi, A., & Daraio, C. (2003). A robust nonparametric approach to the analysis of scientific productivity. *Research Evaluation*, 12, 47–69.
- Bonacorsi, A., & Daraio, C. (2004). Econometric approaches to the analysis of productivity of R&D systems. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.). *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*. Dordrecht (the Netherlands): Kluwer Academic Publishers, 51–74.
- Bonacorsi, A., & Daraio, C. (2005). Exploring size and agglomeration effects on public research productivity. *Scientometrics*, 63, 87–120.
- Bonacorsi, A., Daraio, C., & Simar, L. (2007). Efficiency and productivity in European Universities. Exploring trade-offs in the strategic profile. In: A. Bonacorsi, & C. Daraio (Eds.), *Universities and Strategic Knowledge Creation. Specialization and Performance in Europe* (p. 508). Edward Elgar Publisher, Cheltenham (UK). ISBN: 9781847201102, EID: 2-s2.0-36749059342.
- Borner, K., Klavans, R., Patek, M., Zoss, A. M., Biberstine, J. R., et al. (2012). Design and update of a classification system: The UCSD map of science. *PLoS ONE*, 7(7), e39464. doi:[10.1371/journal.pone.0039464](https://doi.org/10.1371/journal.pone.0039464).
- Bordons, M., Morillo, F., and Gómez, I. (2004). Analysis of cross-disciplinary research through bibliometric tools. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 437–456). Dordrecht (the Netherlands): Kluwer Academic Publishers.
- Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, 45, 197–245.
- Bornmann, L., de Moya-Anegón, F., & Leydesdorff, L. (2013). The new excellence indicator in the World Report of the SCImago Institutions Rankings 2011. arXiv preprint arXiv:1110.2305.
- Boyack, K. W., Small, H., & Klavans, R. (2013). Improving the accuracy of co-citation clustering using full text. *Journal of the American Society for Information Science and Technology*, 64, 1759–1767.
- Bozeman, B., Dietz, J. S., & Gaughan, M. (2001). Scientific and technical human capital: An alternative model for research evaluation. *International Journal of Technology Management*, 22, 716–740.
- Braam, R. R., Moed, H. F., & van Raan, A. F. J. (1991). Mapping of science by combined co-citation and word analysis, I: Structural Aspects. *Journal of the American Society for Information Science*, 42, 233–251.
- Braam, R. R., & van den Besselaar, P. (2014). Indicators for the dynamics of research organizations: A biomedical case study. *Scientometrics*, 99, 949–971.
- Bradford, S. C. (1953). *Documentation* (2nd ed.). London: Lockwood.
- Breschi, S., and Lissoni, F. (2004). Knowledge networks from patent data. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 613–644). Dordrecht (the Netherlands): Kluwer Academic Publishers.

- Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional repositories: Exploration of costs and value. *D-Lib Magazine*, 19(1–2), 0001.
- Calero-Medina, C., López-Illescas, C., Visser, M. S., & Moed, H. F. (2008). Important factors in the interpretation of bibliometric rankings of world universities. *Research Evaluation*, 17, 71–81.
- Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information*, 22, 191–235.
- Callon, M., Law, J., & Rip, A. (1986). How to study the force of science. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 3–18). London: MacMillan Press.
- Calvanese, D., De Giacomo, G., Lembo, D., Lenzerini, M., & Rosati, R. (2007). Tractable reasoning and efficient query answering in description logics: The DL-Lite family. *Journal of Automated Reasoning*, 39, 385–429.
- Carpenter, M., & Narin, F. (1983). Validation study: Patent citations as indicators of science and foreign dependence. *World Patent Information*, 5, 180–185.
- Chen, H., Roco, M. C., & Son, J. (2013). Nanotechnology public funding and impact analysis: A tale of two decades (1991–2010). *IEEE Nanotechnology Magazine*, 7, 9–14.
- Cicchetti, D. V. (1991). The reliability of peer review for manuscript and grant submissions: A cross-disciplinary investigation. *Behavioral and Brain Sciences*, 14, 119–186.
- Cole, S., & Cole, J. R. (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review*, 32, 377–390.
- Cole, J. R., & Cole, S. (1971). Measuring the quality of sociological research: Problems in the use of the Science Citation Index. *The American Sociologist*, 6, 23–29.
- Cole, S., Cole, J. R., & Simon, G. A. (1981). Chance and consensus in peer review. *Science*, 214, 881–886.
- Cole, S. (1989). Citations and the evaluation of individual scientists. *Trends in Biochemical Sciences*, p. 9 a.f.
- Counter. (n.d.). Journal usage factor: Results, recommendations and next steps. <http://www.uksg.org/sites/uksg.org/files/JournalUsageFactorReport080711.pdf>.
- Cozzens, S. E. (1989). What do citations count? The Rhetoric First model. *Scientometrics*, 15, 437–447.
- Cronin, B. (1984). *The citation process. The role and significance of citations in scientific communication*. London: Taylor Graham.
- Cronin, B. (2014). Meta Life. *Journal of the American Society for Information Science and Technology*, 65, 431–432.
- Cronin, B., & Sugimoto, C. (Eds.). (2014). *Beyond bibliometrics. Harnessing multidimensional indicators of scholarly impact*. Cambridge, MA: MIT Press.
- Dahler-Larsen, P. (2013). Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review*, 16, 969–986.
- Daniel, H.-D. (2004). *Guardians of Science: Fairness and reliability of peer review*. Weinheim, Germany: Wiley-VCH.
- Daraio, C., & Simar, L. (2007). *Advanced robust and nonparametric methods in efficiency analysis Methodology and Applications*. New York: Springer. ISBN 978-0-387-35155-1.
- Daraio, C., Bonacorsi, A., & Simar, L. (2015). Efficiency and economies of scale and specialization in European universities. A directional distance approach. *Journal of Informetrics*, 9, 430–448.
- Daraio, C., Lenzerini, M., Leporelli, C., Moed, H. F., Naggar, P., Bonacorsi, A., et al. (2016). Data integration for research and innovation policy: An Ontology-Based Data Management approach. *Scientometrics*, 106, 857–871.
- De Bruin, R. E., Braam, R. R., & Moed, H. F. (1991). Bibliometric lines in the sand. *Nature*, 349, 559–562.
- DORA. (2009). San Francisco Declaration on Research Assessment. <http://am.ascb.org/dora/>.
- Duy, J., & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32, 512–517.

- EU Council. (2000). Lisbon European Council 23-23 March 2000. Presidency Conclusions. http://www.europarl.europa.eu/summits/lis1_en.htm.
- EU Council. (2007). Council of the European Union. Council Resolution on modernising universities for Europe's competitiveness in a global knowledge economy. 16096/1/07 REV 1. http://www.consilium.europa.eu/ueDocs/cms_Data/docs/pressData/en/intm/97237.pdf.
- EUROCRIS. (2013). *Website of the European Organization for International Research Information*, Retrieved November 25, 2013.
- Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, 29, 109–123.
- Evaluation. (n.d.). <https://en.wikipedia.org/wiki/Evaluation>.
- Fetterman, D. M. (1994). Empowerment evaluation. *Evaluation practice*, 15, 1–15.
- Fetterman, D., & Wandersman, A. (2007). Empowerment evaluation: Yesterday, today, and tomorrow. *American Journal of Evaluation*, 28, 179–198.
- Fralinger, L., & Bull, J. (2013). Measuring the international usage of US institutional repositories. *OCLC Systems and Services*, 29, 134–150.
- Francis Bacon. (n.d.). In Wikipedia. Retrieved August 25, 2014 from http://en.wikipedia.org/wiki/Francis_Bacon.
- Garfield, E. (1964). The Citation index—a new dimension in indexing. *Science*, 144, 649–654.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178, 471–479.
- Garfield, E. (1979). *Citation Indexing. Its theory and application in science, technology and humanities*. New York: Wiley.
- Garfield, E. (1983a). How to use citation analysis for faculty evaluation, and when is it relevant. Part 1. *Current Contents*, 44, 5–13, October 31, 1983. In: *Essays of an Information Scientist* (vol. 6, pp. 354–362). Philadelphia: ISI Press.
- Garfield, E. (1983b). How to use citation analysis for faculty evaluation, and when is it relevant. Part 2. *Current Contents*, 45, 5–13, November 7, 1983. In: *Essays of an Information Scientist* (vol. 6 pp. 363–372). Philadelphia: ISI Press.
- Garfield, E. (1994). The application of citation indexing to journals management. *Current Contents*, 33, 3–5.
- Garfield, E. (1998, February 14). Mapping the World of Science. Lecture presented by Eugene Garfield at the 150 Anniversary Meeting of the AAAS, Philadelphia, PA.
- Garfield, E. (2001). From: *From bibliographic coupling to co-citation analysis. via algorithmic historio-bibliography. A citationist's tribute to Belver C. Griffith*, presented at Drexel University, Philadelphia, PA on November 27, 2001 by Eugene Garfield.
- Geisler, E. (2000). *The metrics of science and technology*. Westport, CT, USA: Greenwood Publishing Group.
- Gibbons, M., Limoges, C., Nowotny, H., Schwartzman, S., Scott, P., & Trow, M. (1994). *The new production of knowledge: The dynamics of science and research in contemporary societies*. London: Sage. ISBN 0-8039-7794-8.
- Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, 7, 113–122.
- Gingras, Y. (2014). How to boost your university up the rankings. *University World News*. 18 July 2014 Issue No: 329. <http://www.universityworldnews.com/article.php?story=20140715142345754>.
- Gipp, B. & Beel, J. (2009). Citation Proximity Analysis (CPA)—A new approach for identifying related work based on Co-Citation Analysis. In B. Birger Larsen, & J. Leta (Eds.), *Proceedings of the 12th International Conference on Scientometrics and Informetrics (ISSI'09)*, 2, 571–575. Rio de Janeiro (Brazil).
- Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics*, 35, 167–176.
- Glänzel, W., Katz, S., Moed, H., & Schoepflin, U. (eds.) (1996). Proceedings of the workshop on bibliometric standards Rosary College, River Forest, Illinois (USA) Sunday, June 11, 1995. *Scientometrics*, 35, 165–166.

- Glanzel, W., & Schubert, A. (2004). Analysing scientific networks through co-authorship. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 257–276). Dordrecht (the Netherlands): Kluwer Academic Publishers.
- Glanzel, W. (2008). The multi-dimensionality of journal impact. *Scientometrics*, 78, 355–374.
- Glanzel, W. (2009). High-end performance or outlier? Evaluating the tail of scientometric distributions. *Scientometrics*, 97, 13–23.
- Glanzel, W. (2010). On reliability and robustness of scientometrics indicators based on stochastic models. An evidence-based opinion paper. *Journal of Informetrics*, 4, 313–319.
- Glanzel, W., & Moed, H. F. (2013). Thoughts and facts on bibliometric indicators. *Scientometrics*, 96, 381–394.
- Glanzel, W., Thijs, B., & Debackere, K. (2016). Productivity, performance, efficiency, impact—What do we measure anyway? Some comments on the paper “A farewell to the MNCS and like size-independent indicators” by Abramo and D’Angelo. *Journal of Informetrics*, 10, 658–660.
- Gonzalez-Pereira, B., Guerrero-Bote, V. P., & Moya-Anegón, F. (2010). A new approach to the metric of journals’ scientific prestige: The SJR indicator. *Journal of Informetrics*, 4, 379–391.
- Gorraiz, J., & Schloegl, C. (2008). A bibliometric analysis of pharmacology and pharmacy journals: Scopus versus Web of Science. *Journal of Information Science*, 34, 715–725.
- Gorraiz, J., Gumpenberger, C., & Schloegl, C. (2013). Differences and similarities in usage versus citation behaviours observed for five subject areas. In *Proceedings of the 14th ISSI Conference*, (vol. 1, 519–535). http://www.issi2013.org/Images/ISSI_Proceedings_Volume_1.pdf.
- Gorraiz, J., Wieland, M., Gumpenberger, C. (2016). Individual Bibliometric Assessment University of Vienna: From Numbers to Multidimensional Profiles. arXiv preprint arXiv:1601.08049.
- Guerrero-Bote, V. P., & Moya-Anegón, F. (2014). Relationship between Downloads and Citations at Journal and Paper Levels, and the Influence of Language. *Scientometrics*, 101, 1043–1065.
- Halevi, G. (2014, Sep.). 10 years of research impact: Top cited papers in Scopus 2001–2011. *Research Trends*, issue 38. <https://www.researchtrends.com/issue-38-september-2014/10-years-of-research-impact/>.
- Halevi, G., & Moed, H. F. (2012). The technological impact of library science research: A patent analysis. In E. Archambault, Y. Gingras, & V. Larivière (Eds.), *Proceedings of 17th International Conference on Science and Technology Indicators*, 1, 371–380. Montréal: Science-Metrix and OST.
- Halevi, G. & Moed, H. F. (2014a). International scientific collaboration. In: *Higher Education in Asia: Expanding Out, Expanding Up. The rise of graduate education and university research* (pp. 79–92). Montreal (Canada): UNESCO Institute for Statistics. Ref: UIS/2014/ED/SD/2 REV. ISBN: 978-92-9189-147-4.
- Halevi, G. & Moed, H. F. (2014b). Usage patterns of scientific journals and their relationship with citations. In E. C. M. Noyons (Ed.). *Proceedings of the STI Conference* (pp. 295–301). Leiden.
- Haunschild, R., & Bornmann, L. (2017). How many scientific papers are mentioned in policy-related documents? An empirical investigation using web of science and Altmetric data. *Scientometrics*, 110, 1209–1216.
- Haustein, S. (2016). Grand challenges in altmetrics: Heterogeneity, data quality and dependencies. *Scientometrics*, 108, 413–423.
- Harding, A. (n.d.). What is the difference between an impact and an outcome? Impact is the longer term effect of an outcome. <http://blogs.lse.ac.uk/impactofsocialsciences/2014/10/27/impact-vs-outcome-harding/>.
- Hazelkorn, E. (2011). *Rankings and the reshaping of higher education*. Palgrave MacMillan: The Battle for World-Class Excellence.
- Hicks, D. (2009). Evolving regimes of multi-university research evaluation. *Higher Education*, 57, 393–404.

- Hicks, D. (2010). Overview of models of performance-based research funding systems. In *Performance-Based Funding For Public Research In Tertiary Education Institutions: Workshop Proceedings*. OECD Publishing.
- Hicks, D. (2012). Performance-based university research funding systems. *Research Policy*, 41, 251–261.
- Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, 520, 429–431.
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *PNAS*, 102, 16569–16572.
- Holton, G. (1978). Can science be measured? In Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Toward a metric of science: The advent of science indicators* (pp. 39–68). New York: John Wiley.
- Huang, M., Chen, S., Lin, C., & Chen, D. (2014). Exploring temporal relationships between scientific and technical fronts: A case of biotechnology field. *Scientometrics*, 98, 1085–1100.
- Hunter, D. E. K. (2006). Using a theory of change approach to build organizational strength, capacity and sustainability with not-for-profit organizations in the human services sector. *Evaluation and Program Planning*, 29, 193–200.
- Hunter, D. E. K., & Nielsen, S. B. (2013). Performance management and evaluation: Exploring complementarities. In S. B. Nielsen & D. E. K. Hunter (Eds.), *Performance management and evaluation. New Directions for Evaluation*, 137, 7–17.
- Jonkers, K., & Zacharewicz, T. (2016). *Research Performance Based Funding Systems: A Comparative Assessment*. JRC Science Policy Report. Publications Office of the European Union. ISBN: 978-92-79-57732-1.
- Katz, J. S., & Hicks, D. (1997). Desktop Scientometrics. *Scientometrics*, 38, 141–153.
- Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14, 10–25.
- King, D. A. (2004). The scientific impact of nations. *Nature*, 4310, 311–316.
- Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, 62, 2147–2164.
- Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., & Murray, S. (2005). Worldwide use and impact of the NASA Astrophysics Data System Digital Library. *Journal of the American Society for Information Science and Technology*, 56, 36–45.
- Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Murray, S. S., et al. (2005). The bibliometric properties of article leadership information. *Journal of the American Society for Information Science and Technology*, 56, 111–128.
- Kurtz, M. J., & Bollen, J. (2010). Usage bibliometrics. *Annual review of information science and technology*, 44, 1–64.
- Larivière, V., & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, 5, 392–399.
- Leiden Indicators. (n.d.). Indicators. <http://www.leidenranking.com/information/indicators>.
- Lenzerini, M. (2011). *Ontology-based data management. CIKM, 2011*, 5–6.
- Leydesdorff, L., Bornmann, L., Mutz, R., & Ophof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the Association for Information Science and Technology*, 62, 1370–1381.
- Leydesdorff, L., & Bornmann, L. (2011). Integrated Impact Indicators (I3) compared with Impact Factors (IFs): An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology*, 62, 2133–2146.
- Leydesdorff, L., Rafols, I., & Chen, C. (2013). Interactive overlays of journals and the measurement of interdisciplinarity on the basis of aggregated journal–journal citations. *Journal of the Association for Information Science and Technology*, 64, 2573–2586.

- Leydesdorff, L., Bornmann, L., Comins, J. A., & Milojevic, S. (2016). *Citations: Indicators of Quality?* The Impact Fallacy: Frontiers Research Metrics & Analytics. doi:[10.3389/frma.2016.00001](https://doi.org/10.3389/frma.2016.00001).
- López-Illescas, C., De Moya-Anegón, F., & Moed, H. F. (2008). Coverage and citation impact of oncological journals in the Web of Science and Scopus. *Journal of Informetrics*, 2, 304–316.
- Luther, J. (2002). White paper on electronic usage statistics. *The Serial Librarian*, 41, 119–148.
- MacRoberts, M. H., & MacRoberts, B. R. (1987). Testing the Ortega hypothesis: Facts and artifacts. *Scientometrics*, 12, 293–296.
- Marshakova Shaikevich, I. (1973). System of document connections based on references. *Scientific and Technical Information Serial of VINITI*, 6, 3–8.
- Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36, 343–362.
- Martin, B. R., & Irvine, J. (1983). Assessing basic research: Some partial indicators of scientific progress in radio astronomy. *Research Policy*, 12, 61–90.
- McCain, K. W. (2012). Assessing obliteration by incorporation: Issues and Caveats. *Journal of the American Society for Information Science and Technology*, 63, 2129–2139.
- Merton, R. K. (1957). Priorities in scientific discovery: A chapter in the sociology of science. *American Sociological Review*, 22, 635–659.
- Merton, R.K. (1996). The Matthew Effect in Science, II: Cumulative advantage and the symbolism of intellectual property. In: Merton, R.K. (Ed.), *On Social Structure and Science* (pp. 318–336). Chicago: The University of Chicago Press, 318–336. Also in *ISIS*, 79, 607–623, 1988.
- Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in Scientometrics. *European Journal of Operational Research*, 246, 1–19.
- Moed, H. F., Burger, W. J. M., Frankfort, J. G., & van Raan, A. F. J. (1985). The Use of Bibliometric Data for the Measurement of University Research Performance. *Research Policy*, 14, 131–149.
- Moed, H. F., de Bruin, R. E., & van Leeuwen, Th N. (1995). New bibliometric tools for the assessment of national research performance: Database description, overview of indicators and first applications. *Scientometrics*, 33, 381–422.
- Moed, H. F., & van Leeuwen, Th N. (1996). Impact factors can mislead. *Nature*, 381, 186.
- Moed, H. F., Glänzel, W., & Schmoch, U. (2004). (eds.). *Handbook of Quantitative Science and Technology Research. The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht (the Netherlands): Kluwer Academic Publishers, 800 pp.
- Moed, H. F. (2005a). *Citation Analysis in Research Evaluation* (p. 346) Dordrecht (Netherlands): Springer. ISBN 1-4020-3713-9.
- Moed, H. F. (2005). Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology*, 56, 1088–1097.
- Moed, H. F. (2008). UK research assessment exercises: Informed judgments on research quality or quantity? *Scientometrics*, 74, 141–149.
- Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4, 265–277.
- Moed, H. F., de Moya-Anegón, F., López-Illescas, C., & Visser, M. (2011). Is concentration of university research associated with better research performance? *Journal of Informetrics*, 5, 649–658.
- Moed, H. F. (2012). The use of big datasets in bibliometric research. *Research Trends*, Issue 30, September 2012. <http://www.researchtrends.com/issue-30-september-2012/the-use-of-big-datasets-in-bibliometric-research/>.
- Moed, H. F., & Halevi, G. (2014). A bibliometric approach to tracking international scientific migration. *Scientometrics* 101, 1987–2001. Author copy available at <https://arxiv.org/ftp/arxiv/papers/1212/1212.5194.pdf>.

- Moed, H. F., & Halevi, G. (2015). Multidimensional Assessment of Scholarly Research Impact. *Journal of the American Society for Information Science and Technology*, 66, 1988–2002. Author copy available at: <http://arxiv.org/abs/1406.5520>.
- Moed, H. F. (2016a). *Altmetrics as traces of the computerization of the research process*. In: C.R. Sugimoto (Ed.), Theories of Informetrics and Scholarly Communication (A Festschrift in honour of Blaise Cronin). ISBN 978-3-11-029803-1. Berlin/Boston: Walter de Gruyter. Author copy available at <http://arxiv.org/ftp/arxiv/papers/1510/1510.05131.pdf>.
- Moed, H. F. (2016b). Comprehensive indicator comparisons intelligible to non-experts: The case of two SNIP versions. *Scientometrics*, 106, 51–65. Author copy available at <http://arxiv.org/ftp/arxiv/papers/1510/1510.05128.pdf>.
- Moed, H. F. (2016c). Iran's scientific dominance and the emergence of South-East Asian countries as scientific collaborators in the Persian Gulf Region. *Scientometrics* 108, 305–314. Author copy available at <http://arxiv.org/ftp/arxiv/papers/1602/1602.04701.pdf>.
- Moed, H. F. (2016d). Toward new indicators of a journal's manuscript peer review process. *Frontiers in Research Metrics and Analytics*, 1, art. no 5, doi: 10.3389/frma.2016.00005. <http://journal.frontiersin.org/article/10.3389/frma.2016.00005/full>.
- Moed, H. F. (2016e). Towards new scientific development models and research assessment support tools. In: Proceedings of the OECD Blue Sky Conference, Ghent, 19–21 September 2016. <https://www.oecd.org/sti/038>—Moed Paper final version 250716 submitted to OECD Blue Sky Conf.pdf.
- Moed, H. F., Bar-Ilan, J., & Halevi, G. (2016). A new methodology for comparing Google Scholar and Scopus. *Journal of Informetrics*, 10, 533–551. Author copy available at <http://arxiv.org/ftp/arxiv/papers/1512/1512.05741.pdf2017>.
- Moed, H. F., & Halevi, G. (2016). On full text download and citation distributions in scientific-scholarly journals. *Journal of the American Society for Information Science and Technology*, 67, 412–431. Author copy available at <http://arxiv.org/ftp/arxiv/papers/1510/1510.05129.pdf>.
- Moed, H. F. (2017a). To truly understand peer review we need to study it directly. *Research Europe*, January, 8.
- Moed, H. F. (2017b). A critical comparative analysis of five world university rankings. *Scientometrics*, 110, 967–990. Author copy available at <https://arxiv.org/ftp/arxiv/papers/1611/1611.06547.pdf>.
- Moya-Anegón, F., Guerrero-Bote, V. P., Bornmann, L., & Moed, H. F. (2013). The research guarantors of scientific papers and the output counting: A promising new approach. *Scientometrics*, 97, 421–434.
- Moya-Anegón, F., López-Illescas, C., & Moed, H. F. (2014). How to interpret the position of private sector institutions in bibliometric rankings of research institutions. *Scientometrics*, 98, 283–298.
- Narin, F. (1976). *Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity*. National Science Foundation: Washington D.C.
- Narin, F., & Olivastro, D. (1992). Status report: Linkage between technology and science. *Research Policy*, 21, 237–249.
- Narin, F. (1994). Patent bibliometrics. *Scientometrics*, 30, 147–155.
- Narin, F., Hamilton, K. S., & Olivastro, D. (1997). The increasing linkage between US technology and public science. *Research Policy*, 26, 317–330.
- Narin, F., Breitzman, A., & Thomas, P. (2004). Using patent citation indicators to manage a stock portfolio. In: Moed, H. F., Glänzel, W., & Schmoch, U. (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*. (pp. 553–568) Dordrecht (the Netherlands): Kluwer Academic Publishers.
- Nederhof, A. J., Luwel, M., & Moed, H. F. (2001). Assessing the quality of scholarly journals in linguistics: An alternative to citation-based journal impact factors. *Scientometrics*, 51, 241–265.

- Nielsen, M. (2011). Reinventing Discovery: The New Era of Networked Science. Princeton University Press.
- Nielsen, S. B., & Hunter, D. E. K. (eds.) (2013). Performance management and evaluation. *New Directions for Evaluation*, 213, 1123.
- Nielsen, S. B., & Hunter, D. E. K. (2013). Performance management and evaluation: Exploring complementarities. In: S. B. Nielsen, & D. E. K. Hunter (Eds.), Performance management and evaluation. *New Directions for Evaluation* (vol. 213, pp. 7–17).
- Noyons, E. C. M., Buter, R. K., van Raan, A. F. J., Schmoch, U., Heinze, T., Hinze, S., et al. (2003). *Mapping excellence in science and technology across Europe Nanoscience and nanotechnology*. Leiden: CWTS.
- O'Connell, C. (2013). Research discourses surrounding global university rankings: Exploring the relationship with policy and practice recommendations. *Higher Education*, 65, 709–723.
- OECD. (2010). Performance-based funding for public research in tertiary education institutions: Workshop proceedings. *OECD Publishing*. doi:[10.1787/9789264094611-en](https://doi.org/10.1787/9789264094611-en).
- OECD Glossary (n.d.). <http://www.oecd.org/dac/dac-glossary.htm>.
- Open Science. (n.d.) In Wikipedia. Retrieved August 22, 2014 from http://en.wikipedia.org/wiki/Open_science.
- Paiva, C. E., Lima, J. P. S. N., & Paiva, B. S. R. (2012). Articles with short titles describing the results are cited more often. *Clinics*, 67, 509–513.
- Paruolo, P., Saisana, M., & Saltelli, A. (2013). Ratings and rankings: Voodoo or science? Series A: Statistics in Society. *Journal of the Royal Statistical Society, Series A*, 176, 609–634.
- Pendlebury, D. (2017). Private communication to the author.
- Pinski, G., & Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12, 297–312.
- Poggi, A., Lembo, D., Calvanese, D., De Giacomo, G., Lenzerini, M., & Rosati, R. (2008). Linking data to ontologies. *Journal on Data Semantics*, X, 133–173.
- Price, D. J. D. (1965). Networks of scientific papers. *Science*, 149, 510–515.
- Price, D. J. D. (1970) Citation measures of hard science, soft science, technology, and nonscience. In: C. E. Nelson, & D. K. Pollock (Eds.), *Communication Among Scientists and Engineers* (pp. 3–22). Lexington, MA, USA: D.C. Heath and Company. Available at <http://www.garfield.library.upenn.edu/essays/v4p621y1979-80.pdf>.
- Price, D. J. D. (1980). The citation cycle. In: B. C. Griffith (Ed.), *Key papers in information science* (pp.195–210). White Plains, NY: Knowledge Industry Publications. Reprinted in: Garfield, E., Essays of an Information Scientist, (vol. 4, pp. 621–633), 1980. Available at <http://www.garfield.library.upenn.edu/essays/v4p621y1979-80.pdf>.
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A Manifesto. Available at: <http://altmetrics.org/manifesto/>.
- QS Normalization (n.d.). Faculty Area Normalization. Technical Explanation. http://content.qs.com/qsiu/Faculty_Area_Normalization_-_Technical_Explanation.pdf.
- QS Methodology (n.d.). <http://www.topuniversities.com/university-rankings-articles/world-university-rankings/qs-world-university-rankings-methodology>.
- Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., & Stirling, A. (2012). How Journal Rankings can suppress Interdisciplinary Research—A Comparison between Innovation Studies and Business & Management. *Research Policy*, 41, 1262–1282.
- Rasmussen Neal D. (2012). Social media for academics. A practical guide. Chando Publishing Scial Media Series. Oxford (UK): Chando Publishing. ISBN 978-1-84334-681-4 (print). 978-1-78063-319-0 (online).
- Rauhvargers, A. (2011). *Global rankings and their impact*. EUA Report on Rankings 2011. Brussels, Belgium: The European University Association. 79 pp.
- Rauhvargers, A (n.d.). *Rankings criteria and their impact upon universities*. http://www.unica-network.eu/sites/default/files/Rauhvargers_UNICA_IRO.pdf.

- Reedijk, J., & Moed, H. F. (2008). Is the impact of journal impact factors decreasing? *Journal of Documentation*, 64, 183–192.
- REF (Research Excellence Framework) (2012). Panel Criteria and Working Methods. http://www.ref.ac.uk/media/ref/content/pub/panelcriteriaandworkingmethods/01_12.pdf. Last Accessed March 28, 2014.
- Research Methods Knowledge Base (n.d.). <http://www.socialresearchmethods.net/kb/intreval.php>.
- Salmi, J. (2009). *The challenge of World class universities*, World bank, Washington, pp. 136. Retrieved 12 Jan. 2011 from: <http://siteresources.worldbank.org/EDUCATION/Resources/278200-1099079877269/547664-1099079956815/547670-1237305262556/WCU.pdf>.
- Sarlemijn, A. (Ed.). (1984). *Tussen academie en Industrie (Between Academy and Industry). Casimir's visie op wetenschap en research management (Casimir's view of science and research management)*. Amsterdam: Meulenhoff.
- Sarli, C., & Holmes, K. (n.d.). *The Becker Medical Library Model* for Assessment of Research impact model. Retrieved November 25, 2013.
- Schloegl, C., & Gorraiz, J. (2010). Comparison of citation and usage indicators: The case of oncology journals. *Scientometrics*, 82(3), 567–580.
- Schloegl, C., & Gorraiz, J. (2011). Global usage versus global citation metrics: The case of pharmacology journals. *Journal of the American Society for Information Science and Technology*, 62, 161–170.
- Schmoch, U. (2004). The technological output of scientific institutions. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems*. Dordrecht (the Netherlands): Kluwer Academic Publishers.
- Schmoch, U. (2007). Double-boom cycles and the comeback of science-push and market-pull. *Research Policy*, 36, 1000–1015.
- Schubert, A., Glanzel, W., & Braun, T. Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major science fields and subfields 1981–1985. *Scientometrics*, 16, 3–478.
- Shin, J. C., Toutkoushian, R. K., & Teichler, U. (2011). *University Rankings Theoretical Basis, Methodology and impacts on Global Higher Education*. USA: Springer. ISBN 978-94-007-1116-7.
- Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 24, 265–269.
- Soh, K. (2013). Misleading university rankings: Cause and cure for discrepancies between nominal and attained weights. *Journal of Higher Education Policy and Management*, 35, 206–214.
- Soh, K. (2015). What the Overall doesn't tell about world university rankings: Examples from ARWU, QSWUR, and THEWUR in 2013. *Journal of Higher Education Policy and Management*, 37, 295–307.
- Soh, K. (2015). Multicolinearity and indicator redundancy problem in world university rankings: An example using times higher education world university ranking 2013–2014 data. *Higher Education Quarterly*, 69, 158–174.
- Sugimoto, C. R., Larivière, V., Ni, C., & Cronin, B. (2013). Journal acceptance rates: A cross-disciplinary analysis of variability and relationships with journal measures. *Journal of Informetrics*, 7, 897–906.
- Survey Committee Biochemistry (1982). *Verkenningscommissie Biochemie. Over Leven*. The Hague (Netherlands): Staatsuitgeverij.
- Taylor, M. (2013). Exploring the boundaries: How altmetrics can expand our vision of scholarly communication and social impact. *Information Standards Quarterly*, 25, 27–32.
- THE Ranking Methodology (n.d.). <https://www.timeshighereducation.com/news/ranking-methodology-2016>.
- Thelwall, M., & Kousha, K. (2014). Research Gate: Disseminating, communicating, and measuring scholarship? *Journal of the Association for Information Science and Technology*, 66, 876–889. doi:10.1002/asi.23236.

- Thelwall, M. (2014b). A brief history of altmetrics. *Research Trends*, issue 37 (Special issue on altmetrics, June). Available at <http://www.researchtrends.com/issue-37-june-2014/a-brief-history-of-altmetrics>.
- Theory of Change (n.d.). https://en.wikipedia.org/wiki/Theory_of_change.
- Tijssen, R. J. W., Visser, M. S., & Van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, 54, 381–397.
- Todeschini, R., & Baccini, A. (2016). *Handbook of Bibliometric Indicators. Quantitative Tools for Studying and Evaluating Research*. Weinheim (Germany): Wiley-VCH. ISBN 978-3-527-33704-0.
- Torres-Salinas, D., & Moed, H. F. (2009). Library Catalog Analysis as a tool in studies of social sciences and humanities: An exploratory study of published book titles in Economics. *Journal of Informetrics*, 3, 9–26.
- Torres-Salinas, D., Robinson-García, N., Aguillo, I. (2016). Bibliometric and benchmark analysis of gold open access in Spain: Big output and little impact. *El Profesional de la Información*, 25, 17–24. <http://hdl.handle.net/10481/39608>.
- UNESCO (2014). Higher Education in Asia: Expanding Out, Expanding Up. ISBN 978-92-9189-147-4 licensed under CC-BY-SA 3.0 IGO. Montreal: UIS. <http://www.uis.unesco.org>.
- UNICEF Glossary (n.d.). https://www.unicef-irc.org/publications/pdf/brief_4_evaluative_reasoning_eng.pdf.
- Vinkler, P. (2010). *The evaluation of research by scientometric indicators*. Oxford (UK): Chandos Publishing.
- Van Leeuwen, T. N., Moed, H. F., Tijssen, R. J. W., Visser, M. S., & van Raan, A. F. J. (2001). Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. *Scientometrics*, 51, 335–346.
- Van Noorden, R. (2013). Scientists join journal editors to fight impact-factor abuse. Nature News Blog (Nature Publishing Group).
- Van Raan, A. F. J. (Ed.). (1987). *Handbook of Science and Technology Indicators*. Amsterdam: Elsevier.
- Van Raan, A. F. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36, 397–420.
- Van Raan, A. F. J. (2004a). *Measuring Science*. In: H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 19–50). Dordrecht (the Netherlands): Kluwer Academic Publishers, 19–50.
- Van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, 59, 461–466.
- Van Raan, A. F. J. (2005). Fatal attraction. Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62, 133–143.
- Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., et al. Approaches to understanding and measuring interdisciplinary scientific research (IDR): A review of the literature. *Journal of Informetrics* 5, 14–26.
- Waltman, L., van Eck, N. J., van Leeuwen, Th N., & Visser, M. S. (2013). Some modifications to the SNIP journal impact indicator. *Journal of Informetrics*, 7, 272–285.
- Waltman, L., van Eck, N. J., & Wouters, P. (2013). Counting publications and citations: Is more always better? *Journal of Informetrics*, 7, 635–641.
- Waltman, L. R. (2016). A review of the literature on citation impact indicators. *Journal of Informetrics*, 10, 365–391.
- Waltman, L., & Traag, V. A. (2017). Use of the journal impact factor for assessing individual articles need not be wrong. arXiv preprint arXiv:1703.02334, 2017.
- Wan, J., Hua, P., Rousseau, R., & Sun, X. (2010). The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics*, 82, 555–566.

- Wandersman, A., Keener, D. C., Snell-Johns, J., Miller, R. L., Flaspohler, P., Livet-Dye, M., et al. (2004). *Empowerment evaluation: Principles and action. Participatory community research: Theories and methods in action.* Washington, DC: American Psychological Association.
- Wandersman, A., & Snell-Johns, J. (2005). Empowerment evaluation: Clarity, dialogue, and growth. *American Journal of Evaluation*, 26, 421–428.
- Webcenter for Social Research Methods (n.d.). Research Methods Knowledge Base. Evaluation Research; Introduction to Evaluation. <https://www.socialresearchmethods.net/kb/intreval.php>.
- Webometrics Ranking, n.d. Methodology. <https://www.webometrics.info/en/Methodology>.
- Wildgaard, L. (2015). *Measure Up! The extent author-level bibliometric indicators are appropriate measures of individual researcher performance*, pp 80–83. Ph.D. Thesis, University of Copenhagen.
- White, H. D. (1990). Author co-citation analysis: Overview and defense. In C. L. Borgman (Ed.), *Scholarly Communication and Bibliometrics* (pp. 84–106). Newbury Park: Sage.
- White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, 60, 1083–1096.
- Wilsdon, J., et al. (2015). The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management. doi: [10.13140/RG.2.1.4929.1363](https://doi.org/10.13140/RG.2.1.4929.1363). Available at: http://www.hefce.ac.uk/media/HEFCE/2014/Content/Pubs/Independentresearch/2015/TheMetricTide/2015_metric_tide.pdf.
- Zellner, C. (2003). The economic effects of basic research: Evidence for embodied knowledge transfer via scientists' migration. *Research Policy*, 32, 1881–1895.
- Zitt, M., & Bassecular, E. (1998). Internationalization of scientific journals: A measurement based on publication and citation scope. *Scientometrics*, 41, 255–271.
- Zitt, M., & Small, H. (2008). Modifying the journal impact factor by fractional citation weighting: The audience factor. *Journal of the American Society for Information Science and Technology*, 59, 1856–1860.
- Zitt, M. (2011). Behind citing-side normalization of citations: Some properties of the journal impact factor. *Scientometrics*, 89, 329–344.
- Zuccala, A., Guns, R., Cornacchia, R., & Bod, R. (2015). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology*, 66, 1333–1347.
- Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. *Scientometrics*, 12, 329–338.