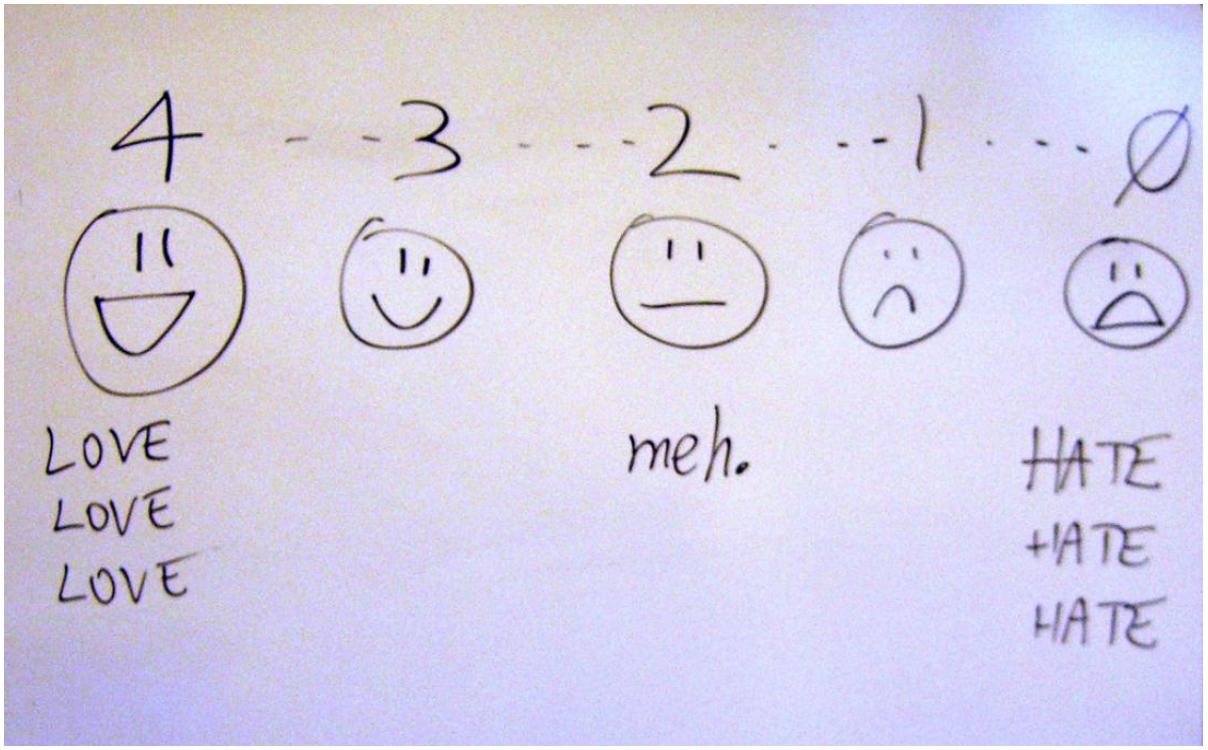
FACULTY OF HUMANITIES
UNIVERSITY OF COPENHAGEN

PhD thesis

Lorna Wildgaard

Measure Up!

The extent author-level bibliometric indicators are appropriate measures of individual researcher performance



Main academic advisor: Peter Ingwersen

Secondary advisors: Jesper Wiborg Schneider & Birger Larsen

Submitted: 31/08/2015

Institute name:	Faculty of the Humanities
Name of department:	Royal School of Library and Information Science
Author:	Lorna Wildgaard
Title/Subtitle:	Measure Up! The extent author-level bibliometric indicators are appropriate measures of researcher performance.
Subject description:	Author-level bibliometric indicators are used by researchers in self-evaluation and by administrators alike to measure academic activities and indicate the impact of the researcher. This thesis examines what these indicators are actually measuring and if the indicators are at all appropriate.
Cover picture:	BillsoPHOTO, “Evaluation Scale”, December 10th, 2009 via Flickr, Creative Commons Attribution and Share Alike.
Academic advisors:	Professor Peter Ingwersen; Jesper Wiborg Schneider, Senior Researcher PhD; Professor Birger Larsen
Committee members	
Chairman:	Jeppe Nicolaisen, Lector, PhD, RSLIS
Opponents:	Henk F. Moed, Professor Rodrigo Costas, Researcher, CWTS
Submitted:	31th August 2015
Word Count:	42,921 words in PhD body and 53,456 words in collected papers, excluding references and appendices.

“Studies have shown that accurate numbers aren’t any more useful than ones you’ve made up.”
Dilbert, 05/08/2008 copyright Scott Adams.

“If scientometrics is a mirror of science in action, then scientometricians’ particular responsibility is to both polish the mirror and warn against optical illusions”.

M. Zitt (2005). Facing diversity of science. A challenge for bibliometric indicators. Measurement, 3(1):38-49

Acknowledgements

This thesis is the result of work funded by a three year PhD scholarship from the Royal School of Library and Information Science (RSLIS), an institute under the Faculty of the Humanities, Copenhagen. The research project received additional financial support from the ACUMEN Fp7 project. But financing can only get a research project so far. Without the support of my colleagues at the RSLIS and ACUMEN partners the project will have ground to a halt long ago. ACUMEN would have fallen apart at the seams without the coordination skills and patience of Laura de Ruiter, Clifford Tatum and Fleur Praal who created the templates for submitting reports and read all the EU protocols and endless administrative documentation to reduce them to a readable amount. A special thank you to Paul Wouters and Frank van der Most for their professional support, criticisms and encouragement throughout and after the ACUMEN project.

A big thank you to the technical support and administrative staff at RSLIS for patiently and cheerfully helping with my endless questions about university procedures, protocols and administration, especially Ragnhild Riis whose energy and knowledge is a true gift to RSLIS; Johanne Maiblom and Susanne Avecado who juggle the coordination of the teaching programmes and student administrative responsibilities seemingly effortlessly, and the student counselors for their support and involvement in diverse projects to get the students excited about the information science education particular thanks to Charlotte Knagh Trojahn, and the students in the Bibliometric And Data Analysis Student Society who keep me on my biblometric toes. Since the very first day as a PhD student, my fellow PhD students at RSLIS have invested interest and shown support, which I want to thank them for: Sara, Rikke, Ole, Sille, Jesper, Niels-Ole and Maria you have made me laugh and motivated me enormously.

To Jesper Schneider, Peter Ingwersen and Birger Larsen, my supervisors, thank you for opening my eyes to bibliometrics and sharpening my interest in research evaluation. Peter and Birger, thank you for your confidence you have shown in me throughout the entire PhD project, your inspiration and your good humour. No one could wish for a better supervisor than Jesper, who took over as main supervisor in the final critical year of the PhD project. Thank you for your collaboration and constructive feedback. You have an impressive enthusiasm for research evaluation and a wonderful ability to make statistics understandable, and not least appropriate. I cannot describe how much our meetings have motivated and inspired me. Thank you.

Further I would like to thank Henk Moed and Rodrigo Costas for their willingness to join my PhD assessment committee together with Jeppe Nicolaisen. I know a large amount of work to fit in to your busy schedules. I appreciate your time, effort and questions. Here's looking forward to an entertaining PhD defense!

Finally, a special thank you to my husband Kim, and our children: Zander and Balthazar. Kim: thank you for helping me in all sorts of ways, especially your undoubting faith and support in me to complete the PhD education. You confirm that my work as a researcher is worthwhile and worth prioritizing. I love you and am greatly looking forward to what our next 20 years together will bring. To you Zander and Balthazar, the best boys any one could wish for – you guys are a constant reminder of what really is important in life.

Summary

This combined PhD work concerns the appropriate development and appropriate application of author-level bibliometric indicators (ALI). The main objective of the thesis is to gain knowledge of the extent ALI are appropriate measures of a researcher's performance at the individual level. The motivation is that evaluations effect people and a basic ethical principle should be that ALI, based on things that appear measureable, first should be theoretically and operationally defined before the mathematical robustness of ALI is defended or the ALI is applied in practice. ALI essentially figure out retrospectively, using limited data, how much and where a researcher publishes, and how much other researchers use his or her work. Numbers representing this usage increasingly underpin research policy and are an established part of research evaluation. But bibliometricians urge caution: even after decades of use, we do not really understand what citation and publication data are and what we do with them at the individual level. As ALI are increasingly influencing research activities, this lack of understanding needs to be explored. These considerations led to the three research questions:

1) What are the characteristics of ALI of academic performance?
2) To what extent are ALI, appropriate in the evaluation of researchers from different disciplines and different academic seniorities?
3) To what extent are the concepts being measured defined in indicator construction?

In order to explore the research questions, publication, citation and demographic data on 750 researchers active in Astronomy, Environmental Science, Philosophy and Public Health were sourced. 51 ALI were calculated for each researcher using this data as well as 18 publication and citation counting indicators. The construction of these ALI and relations between them was explored in 7 research papers and in an empirical analysis of the concepts operationalized in ALI presented in the PhD body. The main contributions of the thesis are:

1) a detailed methodological and theoretical analysis of the construction of ALI,
2) demonstration of the appropriateness of ALI in researcher rankings and in different disciplines;
3) recommendation of a set of ALI that are theoretically and methodologically robust.

The PhD work contributes to the development of guidelines for evaluation using ALI as tools to objectively and informatively measure the research performance of individual researchers.

Resumé (Danish summary)

Denne PhD omhandler udvikling og hensigtsmæssig anvendelse af bibliometriske indikatorer på forfatter niveau (BIFN). Det primære formål med afhandlingen er at afdække om BIFN er en god metode til at måle forskeres præstationer på individuelt niveau. Bibliometriske evalueringer har betydning både for forskeres finansiering og ansættelser. BIFN bør derfor være klart definerede, teoretisk begrundede og være demonstreret operationelt anvendelige før indikatoren introduceres til måling af forskere. BIFN viser, retrospektivt og ofte ved brug af begrænsede data, hvor meget en given forsker publicerer, hvor der publiceres og i hvilken grad andre forskere benytte deres arbejder. Tal der beskriver benyttelsen af en forskers arbejder er beskrevet forskningspolitisk og er nu en essentiel del af administration og forskningsevaluering. Bibliometrikere er dog bekymrede over denne brug idet der selv efter årtiers brug ikke er opstået konsensus om hensigtsmæssighed af BIFN. Da BIFN i stigende grad påvirker forskningen i samfundet, bør denne tvivl undersøges og validitet af BIFN testes. Følgelig forsøges i denne afhandling at besvare tre primære spørgsmål:

1. Hvilke karakteristika besidder bibliometriske indikatorer benyttet til at måle akademisk præstation?
2. I hvilken grad er BIFN anvendelig til at evaluere akademikere med forskellige anciennitet og inden for forskellige videnskabelige discipliner?
3. I hvilken grad er BIFN operationaliseret som målbare variabler i forbindelse med konstruktionen?

Publikationer, citationer og demografiske data fra 750 aktive akademikere indenfor astronomi, filosofi, folkesundhedsvidenskab og plante- og miljøvidenskab blev benyttet. For hver akademiker blev 51 BIFN og 18 publikations- og citationsindikatorer beregnet og analyseret i forhold til de tre forskningsspørgsmål. Via syv publikationer præsenteres:

1. en detaljeret metodologisk og teoretisk analyse af konstruktion af BIFN.
2. en demonstration og analyse af hensigtsmæssig brug af BIFN via rangordning af akademiker inden for forskellige akademiske discipliner.
3. en anbefaling af et teoretisk og metodologisk robust sæt af BIFN

Afhandlingen bidrager med ny viden i form af retningslinjer til brug af BIFN når BIFN ønskes benyttet som en objektivt og informativ metode til at måle individuelle akademikeres præstation.

List of figures

List of tables

Chapter 1: Introduction

Note: Throughout this thesis I use the preferred term "researcher" to describe scientists and scholars. The term "scientist" is used when referring specifically to a person who is an expert in a science, especially physical or natural sciences. "Scholar" is a broader term, and can be used generally for anyone who has profound knowledge of a particular subject in the Humanities but it is not used for the sciences. Whereas a "researcher" could either be a scientist or a scholar.

Bibliometrics is the application of mathematics and statistical methods to books and other media of communication to analyse the structure of science, measure science and to indicate the production, citations and collaboration of researchers, institutions and countries (De Bellis, 2014; Pritchard, 1969). Today therefore there are several ways to characterize bibliometric indicators of individual researcher achievement. There are *Altmetrics* which is a data and technology driven broad category of metrics that capture the various parts of use a researcher's work can have but as yet have no theoretical basis (Zahedi et al., 2014). They emphasize how often the researcher's work is viewed, recommended or downloaded, discussed in science blogs, journal comments, on social media, saved in social bookmarking services, cited in scholarly literature, and offered through commercial vendors including ImpactStory[1] and Altmetric.com[2]. Several publishers have started providing such information to readers, including BioMed Central and Elsevier. Symbolic capitalism scores that allow companies to view the social credit of the researcher based on the amount of social media mentions over time, which emphasize the importance of networks and the influence the researcher has in this network, including for instance the Klout score[3]. *Esteem indicators* are marks of respect from the research community that indicate an individual's research reputation, including counts of awards, fellowships of learned societies, prizes, honours and named lectures, keynote and plenary addresses at conferences, positions in national and international strategic advisory bodies, industrial advisory roles, editorial roles, and conference organisation. Esteem indicators are used in systems for assessing the quality of research in higher education institutions such as REF[4] and ERA[5]. *Conventional ALI (ALI)* allow a mathematical estimation of the impact or relative standing of individual researchers and their contribution to moving science forward (De Bellis, 2014; De Bellis, 2009). A researcher's papers published in journals or books by academic and scientific publishers

[1] https://impactstory.org/
[2] http://www.altmetric.com/
[3] https://klout.com/corp/score
[4] http://www.ref.ac.uk/
[5] http://www.arc.gov.au/era/era_2015/era_2015.htm

are counted and combined with the amount of times these papers are cited. These counts are typically normalized for the field the researcher works in and the number of co-authors on the papers. Although relatively new, the now (in)famous *h*-index just introduced in 2005 (Hirsch, 2005), ALI have quickly become widely accessible and thus adapted and implemented by administrators and researchers in evaluations for tenure, promotion, funding and in other political decisions (Aagaard, 2015). Accordingly ALI have been developed by a wide range of interested parties, not just bibliometricians, into an extensive repository of indicators that aim to improve the field of evaluative metrics at the researcher level - aiming to position the researcher in their field, indicate excellence, production, independence or contribution, as well as how research is communicated and its impact (Cronin, 2014; De Bellis, 2009). More recently, ALI are being used to monitor investment of public money in science by documenting at the individual level the productivity of researchers, i.a. (UFM, 2015; Sivertsen, 2009). The mechanical objectivity of bibliometric indicators supplement Peer Review processes that have long been criticised for their subjectivity and bias (Vieira et al., 2014; Bornmann, 2012; Nederhof & van Raan, 1987; Moed, 1985a), while some argue that peer review gives power to the scientific elite, and enforce the gender power structure (Weingart, 2005). Yet ALI have in turn brought their own limitations and bias to researcher evaluation (Bertocchi et al., 2013; Bornmann, 2012).

Author-level bibliometric evaluation has been condemned but at the same time is interpreted as a consequence of the science system itself (Wouters, 2014b). Bibliometrics are one of the many evaluation tools that policy makers use to create a cultural hegemony, a governing power that can manipulate the value system of science and the practices of researchers. ALI are fascinating to study for two reasons. Even though at their core they are simple counting models of the number of citations certain publications have received, they have a history of theoretical discussion behind them, arguing for what these counts could imply about a researcher's impact in the scientific community. Their implementation and interpretation are cloaked in the political motives and/or the will and knowledge of the person using them. Two aspects the fledgling alt- and esteem metrics have yet to mature in to. I therefore consider this PhD work an appropriate opportunity to illuminate the contradictions in the construction and implementation of ALI; animate contentions so that end-users of bibliometrics can be better informed and more capable of recognizing more appropriate metrics. Therefore, motivated by the simultaneous evolution in citation database accessibility, the explosion in the number of ALI and their increased use in indicator-based researcher evaluation by administrators and by researchers themselves, the overall aim of this thesis is to gain knowledge of the appropriateness of *conventional* ALI.

1.1 Acknowledgement

The first two years of the PhD project was in collaboration with the European FP7 project ACUMEN (Academic Careers Understood through Measurements and Norms)[6]. ACUMEN provided access to a set of researchers active in the social sciences, natural sciences or humanities and the publication lists of these researchers were used to generate the bibliometric data and hence the foundation for interpreting the results of the bibliometric analyses presented in the papers included in this thesis. Further, the dilemmas facing the implementation of bibliometrics in the daily work of administrators and researchers was investigated through ACUMEN in close contact with these end-users and highly experienced bibliometricians which further identified gaps in bibliometric research and thus further verified the need for this thesis to study the appropriateness of bibliometric indicators and accordingly highly influenced the formulation of the research questions.

Table 1. Work-package structure of ACUMEN collaboration

Work-package	Description	Partner 1	Partner 2
WP1	Evaluation Impact	Estonian Research Council	eHumanities Royal Netherlands Academy of Arts and Sciences (KNAW/DANS)
WP2	Institutional Web Presence	University of Wolverhampton	CSIC Spanish National Research Council
WP3	Researchers Web Presence	Bar-Ilan University, Israel	-
WP4	Gender effects of evaluation	University of Leiden	T.H. Wildau Technical University of Applied Sciences
WP5	New Bibliometric Indicators	Royal School of Library and Information Science, Copenhagen	Humboldt University Berlin Aarhus University
WP6	ACUMEN Portfolio	University of Leiden (administrator)	All partners

The ACUMEN collaboration consisted of 6 work packages (WP) that addressed the main problems in the evaluation of individual researchers, of which I was part of WP5 that investigated bibliometric evaluation, Table 1. WP5 explored the idea that through bibliometrics, bibliographic information could be meaningfully linked to research activities by both individuals under evaluation and in third party evaluations by administrators. The ultimate goal of ACUMEN was to use the combined knowledge from all WPs to develop 1) guidelines for evaluation, that would support the

[6] ACUMEN results in brief: http://cordis.europa.eu/result/rcn/159979_en.html

individual and the evaluator in evaluation situations, 2) a recommended presentation portfolio that would ensure all the researchers' activities are presented and documented and, 3) especially for WP5, recommend a pallet of established or novel evaluation indicators tailored to the research field and the seniority of the researcher.

A brief summary covering the deliverables in WP5 can be found in Section 4.2 of this thesis and the full report as Appendix A. More information about ACUMEN can be found at http://cordis.europa.eu/project/rcn/97240_en.html.

1.2 Structure of the thesis

1.2.1 The thesis body

This PhD work consists of a collection of 7 research papers, preceded by introductory chapters and discussion that combines the findings from each paper with an empirical analysis of the theoretical construction of current indicators. The papers are based on publication and citation data of researchers contacted through the ACUMEN project (which forms the premise of this PhD work). Chapter 1 serves as a short introduction motivating the need for a critical reflection on ALI, introducing the objectives of this thesis and the research questions. In Chapter 2 key concepts are defined and related work presented. Six major themes in author-level bibliometrics are identified. These themes further motivate the need for this PhD work's investigation into the appropriateness of the indicators and also expose a gap between the application of indicators and the theoretical background and construction of indicators. Consequently, this chapter forms the background for the posed research questions. The theoretical framework is presented in Chapter 3 and in Chapter 4 the research contributions are presented, including a brief summary of the ACUMEN WP5 final report (included as Appendix 1). In Chapter 5 the research approach is described. Chapter 6 provides a reflexive empirical analysis, where the conceptualization and validation of ALI are investigated through a supplementary analysis of indicator construction using Gingras' evaluation criteria (Gingras, 2014), operationalizing the theory presented in Chapter 3. It was necessary to conduct this extra analysis, because after reflecting over the results of the 7 papers, a foundational analysis of the validity of indicators was clearly missing, which is essential to be able to answer the research questions. By combining the results of the statistical and theoretical analysis a set of recommended indicators was produced, Section 6.5. Appendix B, e-material, details the composition of the indicators investigated in this PhD work, via this link: **http://tinyurl.com/nj4mvca**. 51 ALI and 18

publication and citation indicators were investigated. Conclusions and concerns in relation to the three research questions are discussed in Chapter 7.

1.2.2 The papers

The empirical investigations resulted in a set of 7 critical papers included in this PhD work: The characteristics of ALI (Papers 1 & 2); the feasibility of using indicators to document an individual researcher's "impact" (Paper 3); the potential psychological effects of indicators (Paper 4); the extent indicators measure the same thing, the dominant characteristics of central indicators and the independence of isolated indicators (Paper 5); the extent indicators rank actual scholarly performance rather than ranking researchers coverage in databases, and the stability of indicators in cross-database comparisons (Paper 6), and finally the extent different indicators are appropriate in demarcating ranked performance in different disciplines (Paper 7).

1.3. Objectives of the thesis

A review of the current state of bibliometric methods and future directions that appropriately capture the impact of researchers from different disciplinary societies has recently been described in Cronin et al (Cronin, 2014). This thesis contributes to this important discussion of the appropriateness of bibliometric methodologies, by investigating the mechanisms within the indicators that are operationalized to produce the numbers that in turn are used as labels of research performance. The objective of this PhD work is to recommend a pallet of author-level bibliometric indicators, but to achieve this objective ALI are empirically analyzed in the 7 papers included in this PhD work and the concepts ALI theoretically and technically operationalize are explored in the PhD body. The current state of disambiguation and agitation surrounding ALI is described to provide a summative background of the current challenges, dilemmas, culture and contradictions that affect the design, intentions and application of ALI. This background is important and motivates the need for this thesis work that, if possible, will cut through this muddle of caveats to recommend appropriate bibliometric indicators or if not possible, risk undermining the credibility of author-level bibliometric indicators.

The objectives are as follows:

- The overarching objective is to gain knowledge of the appropriateness of ALI in the light of the aforementioned background.
- Determine a set of ALI appropriate for application by end-users in author-level evaluation to supplement well-informed peer judgment in decision making processes. Consequently the thesis does not just focus on the famous indicators currently used in evaluation, the *h*-index for example is a commonly requested value used in promotion decisions, but this thesis also investigates the lesser known but equally available indicators with the aim to identify more appropriate metrics. This set will be used in the various tests of indicator applicability conducted in this thesis.
- Investigate the background on which the indicators in the set are designed and the concepts they attempt to measure through theoretical and methodological exploration.
- Evaluate the uniqueness and redundancy between indicators.
- Compare the performance of indicators in researcher rankings across different academic seniorities and disciplines

The underlying rationale for the thesis is that the policy decisions made on ALI affect *people* and that, ethically, it is *necessary* to ensure the appropriateness of the indicators.

1.4 Research Questions

The overall aim of this thesis is to examine the appropriateness of ALI by repeatedly probing what it is an indicator *actually measures* and what it is an indicator is *interpreted to measure*. This knowledge is of increasing importance as a great deal of effort, money and time is invested in the development of meaningful quantitative evaluations of academic performance. The indicator values are used on a policy level to distribute university funds and aggregated to compare universities in rankings, and on the researcher level in everyday evaluations that affect an individual's research, tenure, promotion and funds. Importantly researchers are aware of the consequences of measurement systems and through the ACUMEN project I found that researchers are applying indicators to their curriculum vitae in strategic moves to document their achievements. To explore the appropriateness of ALI and learn more what informs their design, application and interpretation, this thesis builds on previous research in this area: the diverse approaches used by indicator developers, that are too many to reference here individually (please refer to the systematic assessment of the construction of indicators presented in Paper 2 and Appendix B), indicator theory

by Wouters (2014a; 2014b; 1999) and indicator evaluation by Gingras (2014). The overall aim is investigated through three research questions that are explored in the 7 papers which form the base of this thesis. Each paper differs in methodology and the aspect of the research question addressed. Chapter 6 enriches the investigations of the research questions through an empirical analysis of the theoretical and operational construction of indicators.

The amount interest in ALI and their application has increased prolifically since the introduction of the *h*-index, reviewed in Chapter 2, yet only a few popular indicators appear to be implemented in practice. This unbalance between the extensive production of ALI and the extent different indicators used in practice could suggest a lack of knowledge in the evaluation and perhaps also the bibliometric community about which indicators are available and what they are designed to measure, leading to the first research question:

1. What are the characteristics of author-level bibliometric indicators of academic performance?

The first question will not only illustrate the abundance of ALI but also addresses the complexity of ALI and the requirements to data needed to calculate them – which data they need, the accessibility of this data, how they are computed and how transparent the calculations are, and if it is possible to understand what the indicator values express. The indicators are demonstrated using bibliometric data on published articles from four very different scientific domains, where the journal article has very different status and are not necessarily the principle written medium for knowledge diffusion. The unique characteristics of ALI and the complexity of their calculation are documented in Papers 1 & 2. The indicators identified in these studies form the base of all investigations presented in this thesis.

This thesis collects a great variety of bibliometric indicators that claim to measure different aspects of researcher performance as well as different dimensions of impact. There is thus a need to assess the methodology of indicator construction before the performance of the indicator is compared using data from researchers in different disciplines and from different academic seniorities. Changes and variation in numerical values might be due to the mathematical design of the indicator rather than changes in a researcher's bibliometric data. This leads to the second research question:

2. To what extent are the author-level bibliometric indicators, outlined in the exploratory study in Papers 1 & 2, appropriate in the evaluation of researchers from different disciplines and different academic seniorities?

The second question explores both the theoretical and mathematical construction of indicators in the investigation of the disciplinary appropriateness of simple ALI. The selected disciplines were determined by the ACUMEN project to represent a broad pallet of different publication and citation traditions within the natural sciences, social sciences and humanities. The disciplines are Astronomy and Astrophysics (Astronomy), Environmental Engineering and Science (Environmental Science), Philosophy and the History & Philosophy of Science (Philosophy), and Public Health and Health Policy (Public Health). Initially the disciplines were identified through WoS Subject Categories, and authors with papers in these categories were invited to take part in the ACUMEN survey about web presence and make their curriculum vitae (CV) available. Through a painstakingly detailed verification process, each researcher's CV was checked to see if they indeed belonged to these broad disciplines and their specialties identified to enable informed comparisons. Together with the responses from the survey, this unique dataset provides publication and citation information as well as access to the researchers' curriculum vitae, demographic data such as gender, nationality, affiliation and academic seniority. In order to investigate the research question two from different perspectives, the question is divided in to sub-questions that are investigated through the included papers: the appropriateness of ALI in increasing the value of publication information on a researcher's CV from the perspective of the researcher is explored in Paper 3, and the potential psychological effects of ALI and issues in application and interpretation of indicators that should be addressed by both researchers and evaluators in Paper 4. The latter paper draws on lessons learned in the evaluation studies literature. The potential disciplinary appropriateness of different ALI is explored in Paper 5. Paper 6 questions how indicators represent the impact of researchers across two main citation databases by studying the construction of the indicators and what this means for the position of the researcher in rankings. Paper 7 continues to investigate indicator construction and the appropriateness of and differences between ALI applied in the 4 scientific disciplines and academic seniority.

In regards to the validity of performance analysis using ALI, indicators are tricky as they present different bibliometric pictures of researcher performance and the numerical values can be difficult to interpret. From a researcher's perspective, indicators that are interpreted to promote them in the

most flattering might be the most useful, but from the evaluator's point of view, the indicators that are informative to a particular question are the most useful. Which is why, in this thesis, the appropriateness of ALI is explored and not their usefulness, as usefulness can differ greatly dependent on the end-user's needs. However, the commonality in the application of indicators by whomever the end-user, is that the interpretation of the ALI affects our interpretation of researcher performance. So it is a fundamental demand that the indicator is a valid measure. But how is validity addressed during indicator construction? Are the concepts that the indicator is designed to measure defined and operationalized in indicator construction? Together with the first research question, this leads to the third and final question:

3. To what extent are the concepts being measured defined in the construction of author-level bibliometric indicators?

The third question is partially, but not satisfactorily, investigated in the papers investigating research question one. In research question one I begin to analyse the construction of the indicator and if it is clear what the indicator measures through a systematic analysis of their composition. This investigation is completed through an empirical analysis and discussion in Chapter 6 of this thesis work, where retrospectively I examine the extent theoretical concepts of citations and publications are defined and operationalized in indicator construction. Indicators measure grand concepts such as excellence, prestige, contribution and impact, and an examination of the extent these variables are demarcated is imperative for the appropriate application and interpretation of indicators. Is it clear what the indicator is designed to measure and how it measures this?

Together the 7 papers with the empirical analysis in Chapter 6 attempt to answer the three research questions. These answers are summarized and discussed in the conclusion, Section 7.1. The success of the methodologies used to answer the three research questions and used to examine the appropriateness of ALI is also discussed, Section 6.5. The papers use different methodologies for analysing ALI while the thesis body is a summative, reflective supplement to the papers, and completes the findings with a discussion of the concepts used in the construction of indicators. This final analysis in Chapter 6 enriches the main research question with a refreshing look at appropriateness. Without analysis of the concepts used in indicator construction, substantial doubt can be cast on the existence of an actual relationship between indicators and the effect of a researcher's publications and hence their appropriateness.

1.5 Limitations

It should be noted that the thesis does not provide a detailed assessment of potential negative impacts and rebound effects linked to author-level bibliometric evaluation (e.g. agreement between peer-review and bibliometric evaluation or future directions in the implementation of ALI in national evaluation systems.) Neither does it quantify the impact of individual evaluation through prospective analysis to estimate the relative risk of the exposure of author-level bibliometric evaluation in the long term. This thesis does though provide practical examples of deconstructing bibliometric indicators to understand what they measure and contextual interpretations of indicator values in regards to academic discipline and to some extent seniority.

Chapter 2: Background

The aim of this chapter is to provide a concise description of the main background that continues to shape ALI, and at the same time begin to explain issues related to the research questions. For this purpose the three main "concepts" that are most commonly operationalized in ALI are first introduced in Section 2.1. A brief chronology of the topical trends surrounding the development and application of ALI is subsequently presented in Section 2.2, which led to the identification of six recurrent issues that directly motivate the objectives and research questions in this thesis work, Sections 2.2.1 to 2.2.6.

2.1 Conceptual background

ALI use quantitative measures to study the scientific progress, communication and impact of published works attributed to an individual researcher. The emphasis of ALI is on the special interrelations with authors, publications and citations and the application of bibliometric techniques to measure these and account for the effects of other variables such as time, field demarcation and age in the evaluation of the individual researcher. Yet the definitions of these three major concepts can either be hard to distinguish or are defined diffusely. It follows that the labels indicator developers use, even if the labels are the same, may not necessarily have identical meanings. Table 6 presents the definitions of author, publication and citation used by the developers of the ALI studied in this PhD work. In the following, an overview of the technical and operational concepts of author, publication and citation is given.

2.1.2 Defining the concept of Author

An explicit discussion of what is an author is found in Foucault (Foucault, 1979) and Barthes (Barthes, 1977). Foucault considers the author as an ideological product, without a constant form, which changes as society changes. An author is nothing more than a marker in the proliferation of meaning. In response, Barthes argues that the qualities of a work are not reliant on the biographical or personal attributes of the author – the author and the work are unrelated. The meaning or importance of the work depends on the reader not the writer, therefore the work is never used as the author intended (Luukkonen, 1997; Latour, 1987), thus severing the ties between authority and authorship – an approach that would be problematic for ALI that link indications of authority to authors rather than papers. The definitions of author in indicator construction ought to be pragmatic, to clarify whether the concept of author is an attribute of a physical manifestation of a publication

registered in a citation index or if other conceptualizations should be considered in light of the specific consequences for indicator measures these could have. There are consequences for choices made at the conceptual stage of indicator development that lead to different measures. For example academic publishers define "author" as the person responsible and accountable for the scientific work, a status qualified by four criteria: the researcher has provided substantial contribution, revisions, approval and integrity to the work[7]. But these formal statements of what constitutes an author vary in the paradigmatic domain (Bošnjak and Marušić, 2012). Bošnjak and Marušić found that only 53% of the WoS Science Citation Index categorized journals included in their study explicitly defined authorship compared to 6% in Arts and Humanities, implying that just as how patterns of authorship are different in different disciplines. What constitutes an author in one discipline might not be applicable in another. The majority of developers of the ALI studied in this PhD work clearly attribute scientists or researchers who have published the value of "author", 20/51 indicators, Table 6, yet the remaining indicators use more specific terms in their indicators, e.g. a common operationalization of "author" are "Price Winners", "person listed on byline of a published paper" or no definition at all. When developers use multiple terms to operationalize the single concept of "author" the definition might be appropriate in some situations, as in the *x* index, where only researchers with at least 15 publications are considered authors. The point, and this can be applied to conceptualization of publications and citations as well, is NOT that there can be only one definition of a concept, but that developers have to define clearly what they mean and establish what dimensions of the concept we need to understand, what underlies the assumptions of the indicator model and the rules the model sets out to measure the concept. Ferrara and Saline (2012) attempt a definition of "author" that connects their concept of author with concrete observations and determine how the concept may change over time or differ between locations. They build on the early work of Derek de Solla Price who defines authorship as "person or persons working at the research front who have produced a paper at a particular time and therefore it is possible to tell something about the relationships among the people from the papers themselves (Price, 1970). Ferrara and Saline present a model based on bibliographic data in which objects involved in an analysis are called "facts". A person is a fact, and this helps to isolate and describe the different elements that compose the fact itself. These elements are separated in to 1) the measures associated *with* the fact and 2) dimensions involved *in* the fact. So a person is a fact that produces a paper at a given time, measures are the number of papers this person produces and dimensions compose the

[7] Example of the definition of the role of authors and contributors: http://www.icmje.org/recommendations/browse/roles-and-responsibilities/defining-the-role-of-authors-and-contributors.html

fact, i.e. contribution is a dimension of the person authoring the publication. Dimensions can be chosen as criteria for grouping data and analyzing them, such as the institution, country, role in producing the product as recorded in the bibliographic data (corresponding author, first author or last author), and profession of the person. Dimensions are then organized as hierarchies to scale the data up or down along the dimensions and to represent different levels of aggregation. Values for the person change in time and may be taken from different sources, for example the person's role and relevance to the product changes. Therefore an author in this model is a person on a byline of a publication registered in the database used for the analysis. Thus an author contributes with a product of a certain "quality" because the product has passed a formal peer review and is published in a journal or by a book publisher that ascribes to specific indexing policies in a database that in turn only includes important and influential journals. Accordingly an "author" is judged by their peers to be of a certain quality and fulfills predetermined criteria at the paper, journal and database level hence bibliometricians can rationalize modelling them as functions of papers and use them in indications of prestige and excellence within the formalized domain. Likewise, Plume and van Weijen (2014) use the term 'author' to define the occurrence of an individual on an paper. Similarly, Marchant (2009) presents a binary definition where 0 is an author without a paper and 1 is an author with a paper, where the author is a function f from N to N and 1_x is an author with 1 paper having received x citations. Co-authors like affiliation, field normalization, career length, are extensions of author but are ignored in Marchant's definition because authors are operationalized as discrete entities, clearly distinct and delineated from each other, and thus scoring rules satisfy independence. From this point of view it makes sense to add to authors together and it makes sense to multiply an author by an integer, and compare authors given their publication and citation record. This begs the question, based on the premise that authors are researchers listed on the author byline on papers indexed in citation indices, is the concept of what constitutes an author in the first position on an author by-line the same concept as what constitutes the author in the second, third or last position? Or does an author exist, even if the database contains no article published by him? Skupin rationalizes the paradigmatic approach to indicator construction, suggesting authors are like surface temperature, existing everywhere in a knowledge domain, continuously, but with different intensities in different scientific locales, and *sometimes* reacting to conditions in those locales by the physical manifestation of a publication (Skupin, 2009). As a network theorist Skupin considers the author concept is more complicated than Marchant's binary solution, and complications have had consequences for indicator design (Skupin, 2009). He argues that the author concept is uncontested yet different conceptualizations have implications for modelling science. He claims that in the

scientometric literature there is no explicit indication that authors are anything other than discrete entities that are distinct and delineated from each other. He on the other hand considers authors as entities that overlap and interact with each other at different intersections in the knowledge creation process. So authors are discrete but, with no right to exist on their own, as they are always dependent on the existence of other entities. Glänzel (2003) agrees that different strengths of authorship appear in different situations and the concept of "co-authorship" or "contribution" raises the question in how far collaboration is reflected by corresponding indicators which operationalize authorship as a variable that can be fractionalized as in Marchant's definition, or in line with many attempts to mathematically model contribution including arithmetic ('proportional'), (Hooydonk, 1997), geometric (Egghe et al., 2000), fractional counting (Price, 1981) and harmonic counting (Hagen, 2010; Hagen, 2008). It is thus imperative to clarify what constitutes an author. Beaver and Rosen (1978) list 18 reasons for co-authorship, suggesting that because these reasons are motivated differently, they support different conceptualizations of author. More recently echoed by Birnholtz (2006) author is defined and used, amongst others, to produce publications and gain access to expertise, gain credit for contribution, equipment, funds, networks, intellectual interest, education, and advance knowledge. Suddenly the author concept is complex and not binary at all because collaboration, the reason to collaborate and the mission of the researcher as instigator, writer or academic advisor alters the concept "author" dependent on the strength of the author's role in regards to another concept, that of "contribution". Merely operating "author" in formalized representations of science by documenting the position on the author byline as a means to define the author's role, as suggested in (Ferrara & Salini, 2012), does not necessarily indicate the right amount of contribution in authorship as practiced in the paradigmatic domain.

2.1.3 Defining the concept of Publication

The very concept of publication, in the etymological meaning, is making something public[8]. Recently, the concept of "publication" has taken on new perspectives in altmetric indicators, to encompass websites, datasets, and other digital materials presented in varying levels of formality and robustness that have been argued as encroaching upon established publication practices (Bishop, 2015) presenting challenges both in the management of these research "documents" and for bibliometric research evaluation. This section is however limited to the conceptualization and definition of publications in *conventional* ALI, Table 6.

[8] Definition from the Online Etymology Dictionary http://www.etymonline.com/index.php?term=publication

The scientometric literature provides no clear and simple answer to what constitutes a publication and is perhaps being careless (Lazarev, 1996) about the nature of the specific properties under study. He concludes that recognising the relationships between the sociological importance of publications and their representation in the formalised realm of science, depends on the proper documentation of all procedures and techniques used in indicator development, and as Wouters (1999) suggests one would expect a reflection on the nature of the publications being measured. Skupin (2009) agrees: what constitutes a "publication" should be clearly defined to ensure a representative operationalization in the indicator and the extraction of meaningful relationships. A more precise definition within the framework of an appropriate mathematical model has been done by Price (1970), although this is not explicitly used in the scientometric literature on indicator development. Price conceptualizes publications through reference to humanistic philosophy, where a scholarly publication is not a piece of information but an expression of the state of the researcher at a particular time. Authors do not publish facts or theories but a complex of these, thus making a scientific paper a concept in itself more than a hypothesis. Defining a paper as an expression of a person working at the research front, it is then possible to operationalize "papers" to tell something about the relations among the people from the papers themselves using bibliographical references and collaborative authorships as social links. Cawkell (1976) defines papers as the end product of scientific research, supported by Lazarev who defines papers in journals as having the *"properties of being fit for a use in a (professional scientific) activity of representatives of a certain domain for the achievement of their (professional) aims"* (Lazarev, 1996). Since published papers follow a traditional pattern in the main and the structure of the paper is indicated by references to earlier 'building blocks' (Cawkell, 1976; Merton, 1973) the potential for examining science through its literature obviously exists. However, what actually constitutes a paper in the instance of ALI is unclear. In current practice, publications are generally operationalized using the umbrella term "papers indexed in a citation index", Table 6. Perhaps this implicitly references Price´s or Cawkell's early definitions, a fine example of Garfield's idea of obliteration by incorporation (Garfield, 1975) or else we must assume that the concept of publication is so obvious that it needs no definition. One would expect at least the metaphors, synonyms and concepts for "papers" in the citation index to be referenced in the corresponding disciplines to establish the paradigmatic importance of what it is that is being measured by the indicator and hence to argue the superiority of the indicator in capturing what is considered important. Returning to the indicators referenced in Table 6, the practice appears to be that a "paper" is a document that can be labelled as a book, book chapter,

report, thesis, article or review in serials and periodicals, on the condition that the physical form has had the core value of being published in refereed scientific journals or by scientific publishers and a representation of the document is indexed in a citation index. Accordingly the validity of aggregating, comparing and dividing the counts into categories of books, articles, reviews etc. as defined in the citation index can be defended (Glänzel, 2003). This approach is exemplified in the recently proposed Snowball Metrics Recipe Book. Here scholarly output is defined as publications in journals, book series, books, or artefacts, compositions, designs, devices and products, digital media, exhibitions, internet publications, performances, reports and software indexed primarily in institutional output repositories, Scopus, WoS, GS and WorldCat (Colledge, 2014). Colledge stipulates the importance of *"overarching definitions that ensure consistency in data sources to validate that comparisons of output counts are meaningful and do not result in misleading conclusions",* because differences can be caused by distinct coverage as well as performance.

Although the discussion of the concept of publication in indicator development is sparse, the discussion of the *differences* in publication forms between scientific disciplines and the underrepresentation of certain disciplinary publication forms in citation indices and consequently in bibliometric assessments flourishes, notably (Castellani, 2014; Hicks, 2012; Tinkler, 2011; Frandsen & Nicolaisen, 2008). Common sense dictates that not all publications are similar and in the paradigmatic realm not all publications carry evidence of the same weight or importance. Although structural analyses of the scientific paper have been conducted i.a (Suppe, 2015) and disciplinary typologies created (Hjørland, 2006; Ziemski, 1975) publications in indicators share the same label "paper" and we are left to assume that all publications are somewhat equal for the comparison of indicator values and researcher performance. But there is an underlying conceptual space in which papers are cognitively mapped and ranked as a knowledge construction mechanism. In this thesis, publications from Astronomy, Philosophy, Public Health and Environmental Science, were used in the empirical investigations. Environmental Scientists were observed in the data set to publish in conference proceedings, Philosophers publish infrequently and cite deep into the past; Public Health Scientists, quite the opposite and Astronomers having utterly different norms of authorship, commonly producing mega-authored papers. Some functions of publications within Public Health and Environmental Science are used in the following paragraph to exemplify issues surrounding definition and operationalization of publications.

In the pyramid of evidence in the medical sciences, the information the paper communicates also embodies different levels of trust in the legitimacy of the information. Original articles for example publish the results of research, claim, prove, argue and aim at impact on the medical community offering concepts and methods for others to use; conference papers, full papers or abstracts, are often the preliminary stage of a journal article, essays argument for or against a concept, standpoint or opinion, while reviews evaluate, synthesize and contextualize other researcher's publications and attempt to establish a value. Within each publication type, continuing with" review" as an example, is a terminology including such terms or phrases, as "review of the evidence", "comprehensive review", "literature review", "overview" and "systematic review" to name but a few (Grant & Booth, 2009). Grant and Booth identified 14 different review types, each with their own associated methodologies, concluding that there is a need for an agreed set of discrete, coherent and mutually exclusive review types to provide an explicit basis to gain a clear understanding of what counts and what does not count as a review. Skupin stresses that publications should be conceptualized as a set of discrete objects so they can be counted in aggregate. Yet for a single set of discrete objects there are a number of alternative concepts derived from different denominators - all equally valid – but in need of consensual definition (Skupin, 2009). Likewise in civil and environmental engineering (CEE) there is a multiplicity of publication types that all fall under the concept "article" (Dzombak, 2013). Dzombak and Mehta identify 11 different types of article that create the mainstream body of knowledge that lead to increased knowledge sharing and collaboration with multi-sector partners but which typically are not published in the main disciplinary journals indexed in citation indices. The authors claim that in CEE it is not necessarily hypothesis-driven manuscripts with well-defined methodologies and positivistic epistemologies preferred by academic journals that move science forward or serve the needs of practitioners and innovators seeking practical insight to directly advance their work but rather preliminary results presented for a series of open ended questions. Consequently, CEE researchers struggle to carve out and share aspects of their work through papers in authority academic journals that are counted by indicators, as much of the information generated does not lend itself to traditional "articles" (Dzombak, 2013), e.g. typologies presenting methods or models and how to use them, manuscripts about challenges and opportunities calling for action and solutions, descriptions of best practices or informal essays which purpose are to generate discussion on a given topic. On this background operationalizing an indicator as a valid measure of CEE researchers production using "papers in a citation index" as the concept definition is not appropriate, even though it is a reproducible measurement.

If the real world phenomena "publication" is not encompassed in the concept label, measurement validity is reduced (Watt & van den Berg, 1995). The fact that the scientometric literature is documenting that people are defining and measuring "publication" in different ways in different disciplines, should convince indicator developers to expand their generic definition of "paper" (Watt & van den Berg, 1995) and to include new measurement items that provide a more representative conceptualization even though this could in turn affect the simplicity of the methodology used to collect "papers". Perhaps the dominate concept of "publication" as "papers in citation indices", Table 6, is an example of instrumentalism, i.e., the definition of the concept "publication" is determined not by whether it is literally true or corresponds to reality in some sense, but by the extent to which they are "papers in citation indices" and can help to make accurate empirical predictions or resolve conceptual problems. From this perspective the amount of work it takes to get useful bibliometric data is reduced, a reproducible method is supported, and it is possible to build on other researchers experiments with indicators using data from the same source. Some indicator developers even claim that the representation of a researcher's scientific output in a citation index *is* a good enough approximation of the performance of the whole of a researcher's oeuvre when supported by other measures (Antonakis & Lalive, 2008), others disagree (De Battisti & Salini, 2012; Bar-Ilan, 2008) and Paper 7, arguing the consequences of underrepresentation present different pictures of scholarly impact. Between the vastly different types of publications researchers produce, and the actual representation of a these publications sourced in citation databases and counted in indicators - and hence their appropriateness – indicators are heavily dependent on the formalized representation of science in citation indices.

2.1.4 Defining the concept of Citation

Whole theses and books have been devoted to citation theory and to discussing the meaning of citation, i.a. (Nicolaisen, 2004; Cronin, 1984). I will return to citation and indicator theories in Chapter 3, while in the following provide a concise overview of the competing rational definitions of citation, which is important in indicator development because reference lists and citations are the basis of ALI, and if authors are aware of this, indicator developers are most surely aware of the effects of citation tactics in evaluation and ALI scores (Moed, 2005). Therefore the operationalization of citations is an aspect that should also be considered in indicator construction, and consequently affects the characteristics of ALI.

The citation distribution to a researcher's articles is skewed. Disregarding the effect of age, a typical distribution of citations to a researcher's publications typically reveals a limited number of highly cited articles and a much larger share of uncited or moderately cited articles. This pattern, according to Moed (2005, p.218) can be found for both leading researchers making prominent contributions to their field and for less prominent researchers. The difference is that prominent researchers tend to have higher citation rates to their significant or as Moed calls them "flag" papers, and relatively lower shares of uncited papers than the less prominent researchers. The amount and distribution of the citations to a publication or aggregated to a researcher stems from sociological citing behaviour which has led to many paradigmatic studies of the concept of citation: in knowledge creation, knowledge use, citing behaviour, as well as studies in the semiotics of citation, where the relation between citations as signs and the things they refer to. These perspectives are used with the indicator approach to define and operationalize the concept of citation within the infrastructure of the discipline the indicator is designed to measure (Wouters, 2014a; Lazarev, 1996).

There are a great variety of perspectives on the concept of citation in the scientometric literature (Nicolaisen, 2004; Wouters, 1999). Instrumental in indicator construction is Wouters consideration of the semiotic inversion of the reference into the citation, arguing that references have very different characteristics both textually and behaviourally (Luukkonen, 1997), as rhetorical or reward devices, yet on the other hand citations are all the same and no longer embody the type of reference that produced it. This, Wouters claims, is because in bibliometric analysis, citations are operationalized as markers of use by other persons, created in an indexing process where the original references are decontextualized from the original text, the number of links to other references registered, collected and counted as citations, co-citations, or bibliographic coupling links within a formalized system. As the citation is now measurable, addible, divisible and comparable, they become attributes of the cited text and the cited author and can be used in indicators as proxy measures of quality or impact or both (Wouters, 2014a). The concept of "citation" is open for interpretation, of which there are numerous, as it is impossible to exclusively link these markers of behavior to a specific behavioural characteristic with respect to citing unless "*one re-translates the citation to the reference as is done in reference analysis*" (Wouters, 2014a; Wouters, 1999). Moed, however, disagrees (2005). He claims that references and citations are not theoretically distinguishable and cannot be separated in operationalization. A citation is not just the product of the citation indexer, as Wouters claims, but also of the scientist and reflects some form of cognitive influence.

Operationalizing the concept of citation and interpreting its meaning has resulted in a great deal of sociological knowledge of the citing behavior of researchers as well as a great diversity of conceptualizations of citation that vary in the meaning attached to the citation because the interpretations are governed by paradigmatic and social norms (Cronin, 2000; MacRoberts & MacRoberts, 1996). Citing (and not-citing) is a "complex social-psychological behavior" (MacRoberts & MacRoberts, 1996) yet others claim that because citation count is heavily affected by factors other than scientific utility, it is essentially arbitrary (Leimu & Koricheva, 2005). But the latter view can be contested: do not the skewed distributions and the fact that 10% of papers attract 60% of the citations testify to some regularity in citing behavior that indicators should be able to measure in research evaluation? The long history of debating what a citation is has led us to where we are today. The conceptualization and operationalization of "citation" remain problematic in indicator construction. Table 6 illustrates the various definitions used in indicator construction, terms like "popularity", "quality" and "reward". "Impact" or "broad impact" are often used, but perhaps as Moed (2005) suggests, *citation impact* would be more appropriate, as it infers the methodology along which the indicator measures impact. However, developers of the indicators studied in this PhD opt for "no definition" as a common method to circumvent the citation concept, Table 6.

Choosing not to define citations may be an effect of bibliometrics lacking a citation theory that encompasses a theoretical foundation for citation analysis, a clear justification to the use of Science and Technology indicators in science policy and an explanation for researchers' citing behavior (Riviera, 2012; Wouters, 1999; Luukkonen, 1997; Leydesdorff, 1987; Cronin, 1984; Cozzens, 1981; Cronin, 1981) Citation theorists are at least agreed that the presence of a citation *may* signify that author A has been influenced by the work of author B, but it cannot, on its own, say anything about the extent or strength of the influence (Martyn, 1964). Variations in existing assumptions about citations, references and indicators, Chapter 3, claim major differences in interpretation, and as such give "citation" new labels in conceptualization and interpretation, e.g. "indicators of use", "value", "persuasion" or "influence". Citations may well be operationalized to reflect an influence, effect or a strong impression of cited documents on citing authors, but such an influence of a document is just a consequence of its value (Lazarev, 1996). Such a view builds further on the work of Ravetz (1971), who defines citation as a form of reward or income, and the Mertonian Normative Citation Theory (Merton, 1973) where citations are defined as the scientist's way of acknowledging an

intellectual debt to other scholarly works. Even more pragmatic is Singleton (1976) who operationalizes citations as 'a quantitative and 'computer manipulable' measure of *something or other'*. Small (1978) on the other hand considers citations as markers or symbols, while Lindsey (1978) defines citations as quality sensor machines, which can be used, with varying degrees of confidence, to "estimate the quality, impact, originality, penetration or visibility of individual and corporate performance within and across disciplines".

Operationalizing the concept of citation at the individual level can be particularly problematic because according to the "Average Mantra" (Nicolaisen, 2004, p.48) citations define influence *on average,* so there needs to be a substantial amount of citations before the average makes any sense. Related to this, is the core issue that not all sources used to write an article are cited, authors only cite a fraction of their influences and accordingly acknowledging these points can lead to the following realization: it is futile to use citations as measure impact, influence and quality at the individual level (White, 2001). For example, if a researcher frequently cites a specific paper, it is not known for sure if he/she has been strongly influenced by it but we can observe that the document is used repeatedly. On the other hand, one might be strongly impressed by some paper, but not use it actively and, therefore, would not cite it and the "influence" of the document would not be reflected. Citations are then argued to be only informative about influence if used statistically in the aggregate thus making ALI redundant (Small, 1987; Nederhof & van Raan, 1987, p.326). In response the Social Constructivist perspective (Nicolaisen, 2004, p.51) argues citations as rhetorical devices used to manipulate the reader into supporting the author's argument as negotiations between scientists take place in the course of scientific practice, i.a. (Latour, 1987). Latour highlights that citations can be positive or negational, essential to the referencing text or perfunctory, whether they concern concepts or techniques or neither, whether they provide background reading, alert readers to new work, provide leads, etc., (Luukkonen, 1997; Latour, 1987). The heterogeneous and chaotic operationalization of citations is then understandable. Scientific documents are a collective process (Latour, 1987) that sell a product and have little to do with intellectual debt (Gilbert 1977, p.113).These findings challenge the validity of Merton's claim that citations are a "*recognition of intellectual debts and original research findings*", (Merton, ibd. Garfield, 1979, p. vii-xi). Further, persuasion is logically not the major motivation to cite, but selection of the most useful papers is. Authors thus cite over the entire scale of reputation and quality and do not favour high end names, disqualifying in this view the operationalization of citations as a proxy measure of quality (White, 2004). Recently, Erikson & Erlandson (2014) presented a full taxonomy of motives to cite,

concluding that it might be misleading to treat all citations as identical concepts in quantitative citation analysis and they should be weighted differently in indicators. As citations are then understood to represent different concepts *and* these concepts are not equal entities, some citations are worth more than others. How then, if we adhere solely to the sociological approach can indicators divulge anything useful at all about researcher performance unless citations are sorted in to conceptual typologies and weighted before computation? This would make indicator construction and the requirement to data preparation very complex, but no one said bibliometric assessment had to be easy. Back in 1979 Eugene Garfield, in attempt to improve the transparency of citation and legitimize the use of citation indicators, defined citation by limiting its function. He proposed citation as "markers" within the formalized realm of science representation and through his index that documented formalized scientific communication, he made citations countable. A smart move by a deveolper of the ISI citation database, because claiming no other correlation between citations and the real world enables the development of bibliometric indicators despite of the lack of a consenusal theoretical citation framework (Garfield, 1979, p.246). Garfield regarded citations as "footprints in the landscape of scholarly achievement" i.e. citations document the passing of ideas (Cronin, 1981), but are only one among a multitude of indications of how scientific information is used within the framework of documented science communication (Glänzel & Schoepflin, 1999). This definition circles this overview back to the present sociological debate of the semiotics of reference and citation, as discussed at the start of this section.

2.2 Topical Background

There is an unprecedented amount of ALI. Papers 1 & 2 describe the characteristics of over 100 such indicators but there are certain to be more. Understandably, there is also an unprecedented amount of social and political investment in ALI, as these indicators claim to document objectively grand concepts like the quality, effect or excellence of a researcher's work, see for example the definitions of what the indicators are designed to measure in Table 6. But what are the topical trends in the scientometric literature that have shaped the development of ALI as we know them today? To answer this question a systematic search of Web of Science[9] and Scopus[10], the two main citation indices commonly used to study research production in academic fields, was conducted on the 11th of December 2014. Both databases were accessed through the University Library, Copenhagen. The

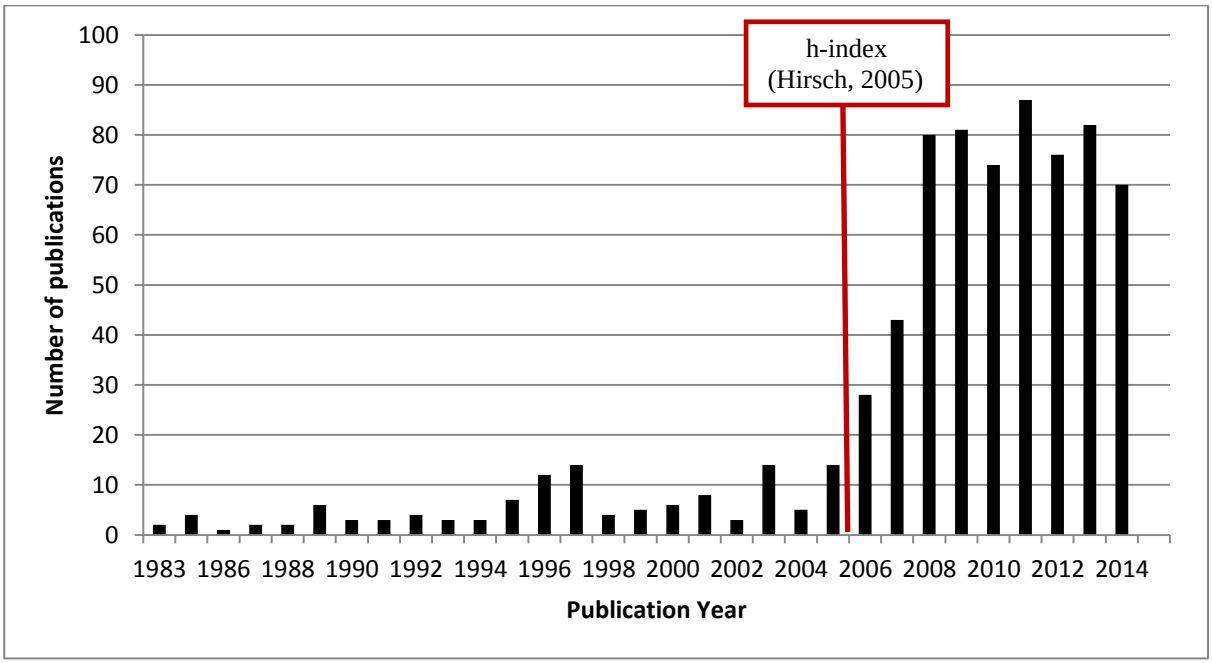
[9] "Overview - Web of Science" Thomson Reuters. 2010. Retrieved 2014-12-15 http://thomsonreuters.com/thomson-reuters-web-of-science/

[10] Overview – Scopus *Scopus Info*. Elsevier. 2014. Retrieved 2014-12-15 http://www.elsevier.com/online-tools/scopus/content-overview

search was supplemented with the references from Papers 1 and 2 to include articles on indicator construction and performance from sources and document types not included in the citation databases. The titles of the retrieved documents were manually filtered at the title/abstract level to include only publications on the development and application of indicators of *author*-level performance. The resulting set consists of 749 publications, ranging from 1983 to 2014, Figure 1. The publications showed distinct topical trends and these are presented chronologically in the following paragraphs.

In the 1980s publications on quantitative author-level assessment primarily debated the correlation between bibliometrics and peer judgment of scientific output, i.a. (Porter et al., 1988; Nederhof & van Raan, 1987; Moed, 1985a; Koeing, 1983) and the differences in citation and publication behavior between researchers in different disciplines, i.a. (Nederhof et al., 1989; Moed, 1989; Moed, 1985b). In the 1990s bibliometric analyses beyond the limitations of traditional citation indices began to be explored, i.a. (Garfield, 1998; Reed, 1995; Mendez et al., 1993). The advance in citation indices led to a vast amount of bibliometric analysis of production in specific scientific specialties, though still primarily the hard sciences i.a. (Xia et al., 1999; Bordons et al., 1995; Peters & van Raan A.F.J, 1994; Plomp, 1994). At the same time bibliometricians were investigating possible fruitful relationships between indicators and science and technology (Banerjee, 1998; Schmoch, 1997); the role of citations in communication theory, building on the earlier work by Cronin (1984), on the sociological interpretations of citations to documents (Leydesdorff & Van den Besselaar, 1997; Luukkonen, 1997). Logically the discussion of author-level bibliometrics as a paradigm blossomed, (Glänzel & Schoepflin, 1994), specifically how to standardize bibliometric terminology and indicators (Katz, 1996; Ravichandra Rao, 1996; Lazarev, 1996; Vinkler,1996), how to define what authorship and collaboration is and how to operationalize authorship bibliometrically, (Katz & Hicks, 1997; Sen, 1997; Herbertz, 1995; Logan, 1991).
Studies on collaboration constitute a recurrent theme to the present day, i.a. (Abbasi et al., 2014; Liu & Fang, 2012; Galam, 2011), increasingly from an evaluationalist and research policy perspective, where the extent inter-institutional and international collaboration is beneficial for research production and individual career trajectories (Abramo et al., 2014). Ultimately, this period ended with a call for an unified indicator theory as a steering framework for future indicator development and application in part to ensure good evaluation practices (Wouters, 1999; Rousseau, 1998; Leydesdorff, 1998).

Figure 1. The growth of interest in ALI



From the year 2000 to the introduction of the *h*-index by Hirsch in 2005, a shift in the discussion of ALI becomes apparent. Discussions of theory and operationalizing citations were replaced with vigorous debates on the use of bibliometric indicators as tools used by governmental agencies, as proxy measures of innovation and productivity in management and funding policy strategies (Russell & Rousseau, 2002; van Leeuwen et al., 2001). Part of this discussion in the bibliometric community concerned the validation of indicators as appropriate measures of individual researcher performance (Cameron, 2005; Costas & Bordons, 2005; Burrell, 2001; Aksnes et al., 2000) and their practical application in science policy and evaluation which was seen to be becoming increasingly institutionalized (Boyack & Börner, 2003; Rowlands, 2003; Bordons & Gomez, 2003). Amongst others, Hicks (2004) and Wiberley Jr (2003), argued particularly the inappropriateness of quantitative indicators in the assessment of researchers in the Humanities and Social Sciences. Consequently, the bibliometric literature grew increasingly concerned with the potential of the recently established Google Scholar database as a supplement or replacement to traditional citation indices (Noruzi, 2005; Jasco, 2005a; Jasco, 2005b).

Throughout the pre *h*-index period, there was a reluctance in the bibliometric field to address ALI because of the inherent size dependency between citation numbers and publication numbers and cumulative effects (van Raan, 1997). Aggregation at the individual level being very low, was argued and consensually agreed to lead to statistical problems and distort indicator values (van Raan, 1996). ALI were not considered to contribute with useful information to a global view of the scientific output of researchers whereas a combination of indicators that quantified the production of researchers, e.g. the total number of published papers, the impact of their publications e.g. the average number of citations per paper and the impact factor of the journals where these papers were published, relative citation rates and so on, did (Martin, 1996). However, after the introduction of the *h*-index, which was a type of indicator never seen before, ALI started to receive a lot of attention and follow-up work from indicator developers from various fields due to the *h*-index's ability to balance the quantity of publications with impact and rank scientists (Alonso et al, 2009). As Figure 1 shows, the literature on ALI exploded, and bibliometricians were quick to cast off their previous reluctance to develop ALI and embrace the challenges of individual metrics, (Panaretos & Malesios, 2014; Wildgaard et al., 2014; Kosmulski, 2013)[11]. van Raan (2006) commented on how quickly the *h*-index attracted attention from the scientific world, policy makers and media. The *h*-index was legitimized by its quick acceptance as a useful measure by the leading scientific journals Nature (Ball, 2005) and Science, and its potentials for ranking researchers in a fair way. Yet in the same journals, caution was just as quickly advised, as "everyone knows that most citation measures, while alluring, are overly simplistic" (van Rann, 2006).

The good properties of *h* and its many adaptations that extend and attempt to overcome the drawbacks of the initial *h* proposal (Alonso et al, 2009; Marchant, 2009) were explored using different researcher profiles (Schreiber, 2013; Schreiber et al., 2012; Schreiber, 2008) and different arithmetic functions, i.a. (Jin et al, 2007; Sidiropoulos, 2007; Egghe, 2006; Kosmulski, 2006; Miller, 2006). The strong influence of research policy continues to shape the literature concerning ALI. The advantages and limitations of ALI in assessment and policymaking continue to be debated, i.a.(Vieira et al., 2014; Bornmann et al., 2008a; van Leeuwen, 2006) as well as the application of bibliometric indicators, especially their suggested application as tools to measure scholarship in hiring, reappointment, tenure, promotion and funding decisions (Južnic et al., 2010; Holden et al., 2005). Accordingly, a flux of guidelines and standards were proposed, aiming to steer meaningful evaluations i.a. (Bornmann & Werner, 2014; Bach, 2011; Schmoch et al., 2010; Sandström & Sandström, 2009), particularly as new sources of data, new types of Altmetric

[11](Wildgaard et al., 2014) is included as Paper 1 in this thesis.

indicators and new perspectives on individual evaluation created new challenges and demands on ALI (Ortega, 2015; Bartoli & Medveta, 2014). Currently, a strong trend in the literature is if indicators need to be adapted for gender (Eloy et al., 2013; Sandström & Hällsten, 2007), seniority (Egghe, 2013; Kosmulski, 2009), career trajectories (Pillay, 2013; Costas et al., 2010a) and cross- and intra- disciplinary comparisons (Harzing et al., 2014; Claro & Costa, 2011; Namazi & Fallahzadeh, 2010; Costas et al., 2009). Moreover, bibliometric indices are being used alongside traditional input-output indicators of the investment in science (Chen et al., 2014; Lepori et al., 2011; Bornmann & Mutz, 2009; Iivari, 2008).

Not surprisingly, with the increased interest in ALI in research policy and individual evaluation, throughout this post *h*–index era the discussion of the validity of ALI (Gaster & Gaster, 2012), the reliability of bibliometric evaluation at the individual level is a prominent (Browman & Stergiou, 2014; van Leeuwen, 2014), and importantly the effect interpretation using different statistical methods has on the perception of the researcher in evaluations (Vieira & Gomes, 2011; Costas et al., 2010b). Consequently, the quality of data and coverage of citation indexes or other data sources is another major topic i.a. (Franceshini et al., 2013; Meho & Rogers, 2008), as the influence of the scope of the citation index on author-level indicator values is not inconsequential in rankings i.a. (Harzing, 2013; Minasny et al., 2013; Bar-Ilan, 2008; Frandsen & Nicolaisen, 2008; Schreiber, 2008).

In recent years, the bibliometric literature appears to be turning increasingly introspective, and investigating if the *appropriate* methodology is being used to explain and predict trends in bibliometric analyses. In earlier years this topic appears to have been peripheral, (Waltman & van Eck, 2012; Prathap, 2012; Bornmann & Daniel, 2007; van Raan A.F.J, 1998; Glänzel & Schubert, 1992), but more recently (Schneider, 2014; 2013a) has readdressed the core methodologies broadly used in the bibliometric literature. Drawing on knowledge from the field of statistics, he strips away the overreliance of the bibliometric community on confidence intervals and significance levels, to remind us that bibliometrics is not a pure science that cannot detach from the fact that its object of study is produced in a social system for the sake of the statistical method. Figuring out what concepts are being operationalized, where the data comes from, what is missing and how the results should be interpreted is more important than sophisticated statistical calculations (Schneider, 2014; 2013a). The call for a unifying indicator theory is revisited as the lack of a theoretical frame for interpreting the sociological conceptualizations indicators analyze. This still constitutes a real

problem (Riviera, 2012). Thus the literature again discusses the need for a theoretical and methodological framework to underpin understanding which methods or best practices can be applied to explain the relationships between formalized representations of science and pragmatic science, and limit related claims to legitimation of specific indicators (Wouters, 1999). Basically, to support the extent ALI are appropriate measures in evaluation.

Six major research issues affecting the development and implementation of ALI were identified in the background review, which at the present time still remain unresolved. These issues motivate the objectives and research questions in this thesis work. They appear to directly affect the professional development of indicators, i.e. Research Questions 1 and 3: *the characteristics of ALI* and *the extent the concepts being measured are defined in indicator construction*; and affect the implementation of bibliometric methodologies and the interpretation of ALI, i.e. Research Question 2: *the appropriateness if indicators in the evaluation of individual researchers from different disciplines and seniorities*. The six issues are discussed in the following sections, 2.2.1 to 2.2.6.

2.2.1 The ambivalence of the bibliometric community

Initially ALI were advocated as blasphemous, and bibliometricians advocated that they should be avoided (Weingart, 2005). The scientific community met author-level bibliometrics with hostility and the threat of legal action if they were implemented (Nørretranders, 2007, p.126; Weingart, 2005, p.126; Garfield, 1979). Bibliometricians were particularly concerned that the use of author-level bibliometric analysis would damage the good reputation of bibliometrics (Schoepflin & Glänzel, 2001) because ALI were "insufficient means of evaluation that lead to erroneous conclusions", (Seglen, 1996; Le Pair, 1995). There was a lack of confidence in the results of ALI because for generations of bibliometricians the strength of bibliometrics was that results were based on the analysis of aggregated data, where "the biases and deficiencies of individual citers are repaired to a tolerable degree by the combined activity of the many" (White, 2001), where deficiencies are reduced to "random noise" (Cawkell, 1976) and "…references can be used on the aggregate as an indicator of influence" (Small, 1987). So when Hirsch, a physicist, introduced the *h*-index (Hirsch, 2005) the bibliometric community praised his efforts (Bornmann et al., 2008a; Egghe, 2006) but was quick to heavily criticize *h* because of the dominant view in bibliometrics citations distributions at the individual level are highly skewed. Only in the analysis of large data files are vagaries in referencing behaviour and the effects of skewness cancelled out. What can be obtained from microanalysis of the individual? (White, 1990). However, fired by the immediate

interest in the potentials of *h* as a combination of production and impact in one easily calculable numerical value, the possibilities of peer-to-peer ranking, as well as the timely demand for performance indicators in the management of universities, bibliometricians quickly embraced the challenges of ALI. Many variants of *h* and *h*-independent alternatives were developed, each claimed more robust, valid and sophisticated than the next. Please refer to the overview in Paper 1 and Paper 2. This development was met by an appalled scientific community who quickly voiced their disapproval of indicators. Giving credence to simplistic metrics like the *h*-index was voiced as damaging, encouraging a gaming mentality[12] and supporting universities to pressurize staff into increasing their indices, and encouraging research policies that monitor research output at the individual level (Dahler-Larsen, 2012; Collini, 2012). But even though the fear of reliance on numbers in research policy and administrative decisions is real enough, (Bishop, 2014; Schneider & Aagaard, 2012) it cannot be attributed to ALI alone. Dismissing author-level bibliometrics and re-adopting the stance against these indicators was not seen as a solution by bibliometricians, whereas a need for standards in indicator development and interpretation was.

2.2.2 Lack of Standards

The proliferation in indicators has not resulted in guidelines steering the development and application of ALI, but appears to have escalated the lack of consensus among different bibliometric research camps about how to defend bibliometric standards, debated already in 1990s by (Glänzel, 1996; Katz, 1996; Ravichandra, 1996; Vinkler, 1996; Glänzel & Schoepflin, 1994). With no advisory boards, common standards or contextual assessments, "indicators published are mostly incomparable, which in fact impedes the development of the field and makes the users of scientometric results mistrustful", (Vinkler, 1996). Consequently, standardization of data, methods, indicators and their presentation is urgently needed. For instance, Vinkler continues, the time periods applied should be standardized across fields and subfields in calculating citation and publication indicators.

[12] In the defense of bibliometrics, gaming will be a problem for any method of allocating money or statistically estimated relationships as the basis for policy rules. Goodhart's law basically states that when you *"attempt to pick a few easily defined metrics as proxy measures for the success of any plan or policy, you immediately distract or bait people into pursuing the metrics, rather than pursuing the success of the policy itself"*[12]. The answer to gaming, though, is to be aware of how this might be achieved and to block obvious strategies, not to dismiss any system that could potentially be gamed (Bishop, 2014).

Standards regarding the ethical aspects of evaluative bibliometrics have recently been proposed again, at the plenary session at the 14th STI conference[13], for example, later published in (Hicks et al., 2015). The ACUMEN collaboration, WP5, Appendix A, p.196, provides a practical guide to limiting the consequences of the use bibliometric indicators: from the researcher's own perspective and from the evaluator's perspective, and especially addresses principles in improving the informed use, calculation, interpretation and contextualization of indicator values. Bach (2011) also proposed ethical standards for the bibliometric evaluation of individual researchers, calling for more work in standardizing the concept of author, studies into the fit of indicators with the purpose of evaluation, studies into data quality and guidelines into how to interpret and contextualize the numbers produced by indicators. Later Furner suggested a conceptual framework to study the ethics of evaluative bibliometrics that could, if produced, inform decision-making in the distribution of rewards (Furner, 2014). Meanwhile, Bornmann et al (2008b) took a mathematical approach in their standards of good practice for analyzing bibliometric data, presenting and interpreting the results, focussing in contrast to Bach and Furner on appropriate statistical analyses of citation counts. Common for all these suggestions to the formulation of standards, is that they warn of the dangers of the ease of generating bibliometric indicators and accordingly the ease of in-proper use.

The lack of standards could be a symptom of greater, unresolved issues in the bibliometric community, from the continued disagreement of what a citation is and motivations to cite, Chapter 2, Section 2.1.4 and Chapter 3, to foundational disagreements around what bibliometrics actually measure, amongst many others (Hicks, 2012; Bollen, 2009; Kermarrec et al, 2007) Perhaps standards are impossible, as they would be detrimental for the development of ALI as a research field, and be the same as enforcing a formal and global hegemony. Yet it is invalid to assume that a disciplinary *framework* for ALI would only be sound when consensus among practitioners is reached. Because various methodological and theoretical positions exist does not mean that no methodological and theoretical foundation exists at all. A framework would be useful in identifying the issues requiring attention and stimulating further work in ALI. However, speculation aside, *guidance* on how to develop and apply ALI is possible and is the responsibility of the bibliometric community, as discussed at the aforementioned STI conference. The need for guidance has become even more urgent due to the demand for objective data and the use of indicators in the management of universities.

[13] https://www.youtube.com/watch?v=1rm63gsc3oI

2.2.3 The inconsistency of author-level indicators

ALI require the collection of a set of variables for a researcher within a given specialty, e.g. the number of publications, number of citations, number of co-authors and number of years active as a publishing researcher. A relationship between these variables is represented numerically through mathematical expressions that use all or a selected sub-section of these variables. As each indicator uses different arithmetic functions in different combinations to calculate the relationship between these variables, they each produce a different number that tells a different bibliometric story of a researcher's accomplishments (De Bellis, 2014). Bibliometric analysis of similar researchers using ALI can thus result in similar numerical values that do not reliably indicate meaningful differences of competitive effort and productivity between researchers, and even less so the academic "quality" the individual. It means that from the outset no one indicator can consistently capture the multi-dimensional aspects of research activity and impact (Colledge, 2014; Bach, 2011; Schmoch et al., 2010; Costas et al., 2010a; Bornmann & Mutz, 2009; Bornmann et al., 2008a; Glänzel, 2003). Indicators could then be used to supplement each other: dependent on the indicators and what they measure, there is a greater possibility for depicting the multi-dimensional aspects of the researcher in the context of their surroundings. But more than that, as each indicator is built on a unique combination of arithmetic functions, one specific indicator could measure more consistently than others the rates of publication and citation data that are typical for a discipline. In principle we could then use this knowledge to our advantage to recommend disciplinary appropriate indicators, which is what this PhD work attempts to do. But the relationships between indicators remain unclear, because their consistency is also time and space dependent (Schmoch et al., 2010; Alonso et al., 2009; Bornmann & Mutz, 2009; Tol, 2009; Waltman & van Eck, 2009; van Raan, 2006). An exploration into consistently dominant indicators in disciplinary rankings directly motivate Research Questions 1 and 2, and hence the investigations in Papers 3, 5, 6 and 7.

The mathematical counting method is a key in investigations of the inconsistencies of ALI, as it is by the counting up method change and comparisons in researcher performance is indicated, (Vanclay, 2015; Costas & Bordons, 2007a). The applied arithmetic can be crude, sophisticated, robust, simple or complex, but however mathematically consistent the indicator is, it can still at some level be affected by inconsistencies in the underlying data. Studies have shown that even with perfect data, mathematical inconsistencies can still be present in indicators (Waltman & van Eck, 2012). Any indicator used to measure non-definitive concepts such as "a citation" or "impact" (Wouters, 2014a; Wouters, 1999; Leydesdorff & Van den Besselaar, 1997) and calculated on small

amounts of imprecise data is immediately suspect as the indicator is so malleable (Weingart, 2005). A single digit difference in an indicator value between researchers may be due to error, the time window chosen, the type of publication used to communicate the results and the publication date of the paper and the time it takes specific publication types to attract citations. Citation distributions are highly skewed and the long tails of the distribution affect calculation of average values. A related phenomenon is the concept of criticality. At some critical point, a paper achieves enough citations that other citations to it seem to accelerate. Using bibliometrics the paper, in citation terms, takes off on a different and higher trajectory (Pendlebury, 2008). The challenge in and possibilities of determining this point was noticed by de Solla Price in *Big Science, Little Science* and sociologists of science discussed the point of cumulative advantage, i.a Robert Merton, Jonathan and Stephan Cole. Defining this point mathematically later attracted the attention of physicists, i.e. the *h*-index (Hirsch, 2005), *alternative h* (Batista et al., 2006) and the *generalized hf* (Radicchi et al., 2008). Yet even though this is a mathematically elegant approach to ALI development, an inconsistency becomes apparent in scholar rankings. Seven years after the introduction of the *h*-index, Waltman and Eck demonstrated its core inconsistency, in that it performs in rankings in a counterintuitive way (Waltman, 2012). With consistent indicators, it is sure that if two authors show the same relative or absolute performance they do not rise or drop rank positions (Riviera, 2012), this is not the case of the *h*-index and Moed (2005) was quick to point out after the introduction of *h*, that authors with very different citation distributions can have the same *h*-index. Consequently indicators continue to be investigated to see if they suffer the same inconsistencies as *h* and if different mathematical manipulations can be used to overcome this. Please refer to the 42 ALI that adapt *h* or are inspired by *h* identified in Papers 1 and 2. The further development of *h* clearly illustrates that the discussion of the mathematical inconsistencies of indicators is a major issue within the bibliometric community. And if I refer again to the indicators referenced in Papers 1 and 2, we observe that the criticisms of indicators and consequent suggestions to improvements are communicated through mathematical theorems, proof and equations i.a. (Eck & Waltman, 2008; Wan et al., 2007) that are, in my opinion, inaccessible to end-users who are typically outsiders to the field of indicator construction but use bibliometrics in their daily activities. These end-users, such as reviewers, hiring committees and grant panels, are directly affected by their inconsistencies yet they use bibliometrics to support decisions.

2.2.4 Exogenous variables

The count of publications, authors and citations are controlled for within ALI models, or they are endogenous. Other factors that are not controlled for would therefore be exogenous, they are determined outside of the indicator model. The exogenous variables set arbitrary external conditions on the ALI and create difficulties for estimating more realistic model behaviour.

At the level of the individual researcher there are great many exogenous variables that affect the computation of indicators, which is why investigations into relevant normalizations, e.g. field, source, university or group, are so important to yield accurate results when using different families of indicators at the individual level, (Waltman & van Eck, 2013). Normalization of citation scores using reference sets based on WoS Subject Categories (WCs) has become an established practice in evaluative bibliometrics to establish disciplinary benchmarks. Leydesdorff and Bornmann (2014) point out that WCs were developed decades ago for the purpose of information retrieval and evolved incrementally with the database; the classification is machine-based and partially manually corrected. They show that WCs do not provide sufficient analytical clarity to carry bibliometric normalization in evaluation practices because of "indexer effects" that in turn create artificial expectations and unrepresentative benchmarks in evaluation, as in the *hf*-index (Radicchi, et al, 2008) and the *n* index (Namazi and Fallahzadeh, 2010). Dividing science into clearly delineated fields is artificial; fields are by no means homogeneous and can be divided into sub-fields that differ in publication and citation practices, and field classification systems are defined at the journal level rather than the publication level (Waltman and van Eck, 2013). Alternatives using source normalization i.e. the average of the ratios of a publication's actual number of citations and its expected number of citations have been suggested, i.a. (Lundberg, 2007; Bornmann, 2010; Gingras and Larivière, 2011; Moed, 2010), likewise fractional counting (Leydesdorff and Bornmann, 2011) and a priori normalization (Glänzel, Schubert, Thijs, & Debackere, 2011) and directly as an ALI in the *IQP*-index (Antonakis and Lalive, 2008) which uses the journals researcher publishes in to establish specialty-specific citation rates. Other approaches attempt to avoid the effects of exogenous variables in the model, and normalize using characteristics of the endogenous variables, e.g. by using the number of authors per paper as a distribution variable (Carabone, 2011), or by multiplying the fractional publication count by a weighted factor so some publication types count more than others (Antonakis & Lalive, 2008) and not forgetting the *h*-index that limits the citation count to publications that have a minimum threshold of *n* citations (Hirsch, 2005). Normalizing the indicator enables end-users to arrive at a degree of belief in the resulting value and faith in that they have correctly applied the indicator and held exogenous variables constant.

For any particular paper, a great many known exogenous variables can occur and will determine what happens in the calculation of the indicator (Aksnes, 2009): funding, location, research topic, language, time, age of researcher even gender has been suggest to directly affect indicator scores, further investigated in the ACUMEN project. But it is primarily the large variation in the relative importance of quality versus visibility dynamics of the researcher in the database that concerns the appropriateness of ALI (Aksnes, 2009). These are listed here: the specialty of the researcher and indexing policy of the database (Archambault & Larivière, 2010), the percentage of a researcher's articles that appear in journals indexed in citation database (Jasco, 2005a), the age of the references and ratio between new publications in the field and total number of publications affect concepts of currency and use (Russell & Rousseau, 2002), disciplinary averages based on journal categories or research areas defined by the database do not represent the specialty of the individual researcher (Papers 7 and 8), there are large individual and disciplinary differences in publication and citation rate (Hicks, 2004), international knowledge development rather than national knowledge development is represented in citation indices (Russell & Rousseau, 2002), the contribution of the researcher to technological, societal or industrial advances are typically not included in the computation of indicators, (Tinkler, 2011), technical factors such as spelling mistakes, errors in reference lists affect accuracy of citation counts to individual articles and are not averaged out at the individual level and, citation circles and self-citations can affect the amount of received citations and how they are counted (Moed 2005;2010).

2.2.5 Commercialization

The wide spread interest from the scientific and administrative community in bibliometric methods and the increasing reliance on good indicator scores for income has created a commercial interest in satisfying end-users of bibliometric indicators. Policy makers, administrators, funding agencies and research councils have either built up their own institutions to collect and process data on performance of own researchers or use commercial institutes or groups to do bibliometric analyses for them, take the Center for Science and Technology Studies (CWTS) as an example. Other providers offer tailor-made tools that allow anyone to identify their own impact relative to their peers and compute their own ALI, e.g. Google Scholar author citation tracker[14], InCites[15] or free

[14] https://scholar.google.com/intl/en/scholar/citations.html
[15] http://thomsonreuters.com/en/products-services/scholarly-scientific-research/research-management-and-evaluation/incites.html

software specifically designed to retrieve and analyze academic citations. Harzing's POP[16], for example, uses Google Scholar and (since release 4.1) Microsoft Academic Search to obtain the raw citations, analyse these, and present a list of ALI. The commercial value of indicators has developed bibliometrics, transforming a "hobby-like" field to a demand-pull field (Miquel, 1994), and the demand appears dominated by science policy and business (Gläser & Laudel, 2007) who's actors use indicators in evaluations that range from the strong and highly regulated to the weak, secretive and unregulated (Whitely, 2007). Bibliometricians claim that indicators are being used as grading systems with no scientific basis and the numbers are taken at face-value (Gingras, 2014; Gläser & Laudel, 2007). They have become a *product* to be sold, a *solution* that provides modulated methods and data to fit local constraints, benchmarks and comparisons to competitors and priced to fit a variety of audiences and funding agencies (Miquel, 1994). Researchers, it appears, use indicators to market themselves as "micro-brands", tacking their movements and reputation over time (Nørretranders, 2007). They combine social credit scores with conventional bibliometric indicators of impact to increase their commercial "value" (Cronin, 2014, p.12). Gingras (2014) argues that the need for numbers has resulted in a loss of critical sense in the employment of indicators and commercial vendors pray on the psychological and sociological influences of performance evaluation that affect the end-users – the need to be up-dated on changes in indicator scores, connected, visible, cited and the need to perform- marketing their indicators as documentation of the researcher's influence and power (Nørretranders, 2007). The methodology behind such indicators and how they are calculated has repeatedly been criticized in the bibliometric field, initiated by Leydesdorrf and Opthof (2010a;2010b), continued recently by (Lopez-Cozar et al., 2014) because of the indicators persistent lack of transparency, blend of different types of publications, citations and media mentions from different sources (which could or could not be weighted or counted fractionally in the algorithm) and the individual's capacity to manipulate with the data. Like the Colonel's blend of spices or the Coca Cola recipe, the construction of the indicators and algorithms to collect data are a trade secret and the rapid commercialization and appetite for bibliometrics beyond the bounds of professionalism could threaten the methodology of bibliometrics. The end-user is not able to reflect upon the mathematical or operational mechanisms in the construction and calculation of the indicator and likewise bibliometricians can only reflect upon how the indicators are *probably* calculated. Now indicators are applications that can be subscribed to, and the resulting numbers interpreted by end-users with little or no professional background in the field, with little or no knowledge of the modalities of bibliometrics. However the

[16] http://www.harzing.com/pop.htm

acceptance of results from both commercial providers and the amateur implementation of indicators could be due to the omnipresence of indicators. The need to evaluate and produce numbers overrides the usefulness and validity of the indication (Haustein & Larivière, 2015) and at the same time, the growing competition for appointments and funding drives researchers to use quantitative indicators to demonstrate the superiority of their research compared to their peers, the influence they have in their networks and the strategies they use to disseminate science in other media than journals and books, including blogs, presentations, editorials in newspapers, radio-spots, etc. Likewise the growing competition for appointments and funding drives administrators to use these indicators as criteria to judge a researcher's present and future academic performance and influence. The fear in the bibliometric community is that bibliometrics will be de-professionalized and lose their character as methodological tools designed to supplement other bibliometric indicators and forms of assessment (Wilsdon, 2015; Gläser & Laudel, 2007). Gläser and Laudel suggest a remedy to decrease interest in commercial bibliometrics by increasing the validity of professional indicators, which is the major motivation of this PhD work. They conclude that improvements in validity would require science policy and other stakeholders to invest in professional bibliometric evaluations, bibliometric education, guidelines and ethical standards and also collaborate on the creation of an open citation database that is quality controlled to overcome the described problems.

2.2.6 Institutionalization of ALI

The institutionalization of indicators through science policy and research evaluation has become an important area for the development and application ALI (Cronin, 2014; Gläser & Laudel, 2007; Aksnes, 2005; van Leeuwen, 2005; Aksnes et al., 2000). Glänzel and Schoepflin (1994) proposed the institutionalization of bibliometrics would lead to the perception of bibliometrics being dominated by science policy and business interests which could cause a fragmentation and diminishing quality of bibliometric studies. They predicted a shift away from basic and methodological research towards applied bibliometrics. Likewise Russell (1994) wrote of the expected trend of the production of data without sound theoretical and methodological foundation being used to back-up policy decisions will implicate the bibliometricians supplying the original data. Indications of production, distribution and use of scientific and scholarly papers have indeed become embedded in organizational and social systems that combine innovation process management strategies with approaches to create, monitor, manage and improve science. Indicators are implemented as part of digital era-governance techniques to strengthen the relationship between the institution, research and society and standardize best practices (de Vries, 2010).

Such management strategies are incentive oriented and some sort of indicators are needed to measure progress and knowledge (as a product) dissemination. Yet there is considerable scepticism among researchers, universities and other stakeholders about the use of bibliometric indicators in research evaluation (Wilsdon et al, 2015). Dahler-Larsen suggests part of this skepticism could be due to the results of bibliometric evaluation being used in an information function as well as an allocation function and operationalize the interconnections between research and researchers at an individual level (Dahler-Larsen, 2012). The indicators in his view are regarded as means of control rather than means of recognition that feed scientific innovation and progress. This indicator fixation results in goal displacement, where researchers focus on indicators being a proxy measure for quality rather than safe-guarding quality itself (Dahler-Larsen, 2012; van Leeuwen, 2005), which means:

> "what started as an intention of objective measurement of scientific production has become a paradox, as rather than inciting innovation in research or stimulating fruitful collaboration, the indicators have resulted in being goals in themselves." (Goodhart, 1975).

The indicators are supposed to be objective and used as a supplement to Peer Review panels, case studies and other qualitative evaluation methods but because of limited resources bibliometric analysis can even remove evaluation by Peer Review panels (Haustein & Larivière, 2015; van Leeuwen, 2005). Even though an informed choice of selected indicators can complement decision-making, it is not currently feasible to assess research outputs using indicators alone (Wilsdon et al, 2015). As Cronin points out, the indicators can lead to an oversimplification of what scientific output and impact are, there is the prospect of monitoring and control of research groups and individuals to generate performance data on demand, the possibility of producing hierarchies of difference and categories of normal/abnormal scientific behavior (Cronin, 2000). Changes in publication and citation behavior may increase a researcher's ALI score but can in the end distort scientific progress (Haustein & Larivière, 2015; Martin, 2013) and break with the basic trust norms and ethics of scholarly practice (Merton, 1973). It has been suggested by Fanelli and Ioannidis that the soft sciences are more vulnerable to adapting behaviour to fit indicators and expected benchmarks rather than the hard sciences (Fanelli & Ioannidis, 2013). Haustien and Larivière (2015) agree, those whose production is being counted may produce more according to the criteria

that the indicator uses to count publications rather than criteria that make sense for science. Many researchers thus feel that the primary purpose of the indicators are as control and reward mechanisms by third parties external to their specialty and institution to increase production (Aagaard, 2015; Emmeche, 2014; Schneider & Aagaard, 2012), and this monitoring suggests a lack of confidence in their work. The result could mean that a redefinition of what it takes to be a successful academic is happening (Emmeche, 2014).

The institutionalized *need* to measure and account for investment in science, such as number of publications, citations and *h*-like indices, may create increasingly perverse incentives in the research sector where much of what is most valued resists simple quantification. Moreover there an on-going discussion that the numbers indicators produce are in danger of being used as a shortcut in evaluation; they have become so institutionalized, they are used as a substitute for thinking (De Bellis, 2009; Egghe & Rousseau, 1990; Garfield, 1985). Too often, poorly designed ALI are "dominating minds, distorting behaviour and determining careers." (Wilsdon et al, 2015). Weingart wrote that systems that link investment in science and scientists with notions of public accountability would fuel the demand for "off the shelf" indicator packages, and he already suggested 10 years ago the healthy skepticism had given way to an uncritical embrace of bibliometric measures and to an inappropriate use in research evaluation (Weingart, 2005). Indicators, he claims, are no longer used as recognition mechanisms that describe the links between knowledge in scientific domains, institutions and authors, or link the communication of ideas and results to understand the structure of science (Cawkell, 1976), but are used to recognize the performance of the individual within the system. The rationale behind the institutionalization of bibliometrics is that the pivotal point of science is publishing results and having these results used in practice or cited and developed in other articles (UFM, 2015; Colledge, 2014; DU., 2009; Aksnes, 2005). Yet publication-based indicators are typically used by policy makers and administrators to measure production and the use of science to decide who gets money or rewards, without regard for citation-based indicators which attempt to measure the type of recognition the work is getting (Weingart, 2005). The behavioural incentives and issues behind writing a paper and citing a paper are vastly different and need to be accounted for as such in bibliometric research evaluation using appropriate indicators.

2.3 Summary

Chapter 2 has introduced some important characteristics concerning ALI, the major concepts of author, publication and citation that are operationalised in the indicator model and six recurrent issues that inform their appropriate application and interpretation.

ALI are attractive because in their mathematical composition they can be characteristically simple and they are used to measure objectively for example a researcher's *impact*, *excellence* or *quality*. Yet their conceptual composition and interpretation is of a highly complex character. They are conceptual models that reify abstractions of authors, publications, and citations in the real world or a formalized representation of the world, which are in turn argued as physical or social constructs (or both), Sections 2.1.1 to 2.1.3. Due to the multiple terminologies used to label these concepts, clear definitions of how they are operationalised in the indicator model is essential to help the end-user know, understand and simulate the subject the model represents. Therefore without analysis of the concepts operationalized in indicator models, substantial doubt can be cast on the existence of an actual relationship between indicators and the effect of a researcher's publications. An investigation into this relationship is continued in the empirical analyses in Papers 3, 5, 6, 7, and 8, and in Chapter 6 to further explore the extent ALI are appropriate in the evaluation of researchers from different disciplines and different academic seniorities (RQ2) and in the extent the concepts being measured are defined in indicator construction (RQ3). The operationalization of the concept of authors, publications and citations into variables that can be measured is fundamental in the construction of ALI, as these are three major variables that are not inconsequential in the scientific communication and knowledge production processes.

The background review introduced how quickly the discussion and development of ALI flooded the indicator market and the speed in which the ALI were accepted and applied in author-level assessment by commercial suppliers and by institutions. Six major themes were identified: the continuing ambivalence of the bibliometric community to ALI; the lack of a framework supporting indicator construction; the mathematical inconsistencies in ALI that bring in to doubt the indicators' ability to say anything reliable about the relationship between a researcher and the measured perception of his or her performance, or the stability of the indicator on small, highly skewed amounts of bibliometric data; exogenous variables that affect the indicator model; and finally the effect commercialization and institutionalization have had on the appropriate development and application of ALI. The conflicts, challenges and potentials raised in these issues motivate why it is necessary to address in this PhD the extent bibliometric indicators are appropriate measures of

individual researcher performance. Consequently it is essential to undertake an honest reflection of the appropriateness of ALI, which could possibly contribute to undermining the legitimacy of author-level bibliometric analysis. However, the overall aim of the thesis is to recommend appropriate ALI, not discredit the bibliometric field. To recommend indicators this PHD work continues in the following chapters with both a theoretical discussion of what ALI "indicate" and an empirical investigation in to the validity of ALI.

Chapter 3: Theoretical assumptions

The field of scientometrics uses many different terms to label the concepts ALI measure and there are likewise many different variables affecting their application and interpretation, as reviewed in Chapter 2. As I discussed, this is because ALI are used to model a specific reflected property of a construct being measured, by directly associating the proportion of citations to publications with the notion of the bigger concept, be it "use", "excellence" or "impact". The interpretation of ALI is of major importance in exploring the appropriate use of ALI which is the main objective of this PhD. Therefore this chapter discusses the theoretical assumptions of citations in indicator analysis and a rationale for thinking about ALI. This discussion will inform the RQ1: the characteristics of ALI and RQ3: the extent the concepts being measured are defined in indicator construction. Which is what makes this chapter particularly interesting – connecting theory with practice.

The first section introduces differences between theoretical assumptions of citation in reference, citation and indicator analysis, Section 3.1. Section 3.2 discusses the rationale of using citations to quantify the use or impact of a researcher's work and, Section 3.3 introduces implications using citations can have in ALI development. This last section focuses on matters related to the importance of a theoretical foundation in the appropriate use of indicators.

3.1 Citations as links to the effects of publications and authors in ALI

The goal of ALI is to operationalize publications and citations so they can measure concepts such as "impact" or "effect". The assumption is that these concepts can be measured using variables found in the specific fields of bibliographic records, primarily publications and their corresponding citations. Accordingly, ALI make these variables operational by counting them and combining them with arithmetic functions to explore the strength of the shared relation between these variables and the concept they aim to measure. ALI are typically referred to as "hybrid" indicators, that aim to capture both productivity and impact in a single figure (Franceschet, 2009). They combine publication analysis that focuses on productivity metrics (number of papers, papers per academic year) with citation analysis which applies impact metrics to study occurrence and co-occurrence counts of references and documents, (total number of citations, number of citations per academic year). Further, citation analysis is concerned with understanding the function and meaning of a citation and/or reference, and why researchers cite each other in the first place. Which means citation theorists may split the meaning of references and the meaning of citations into two distinct issues. This has been thoroughly discussed in the literature (Wouters, 1999; Moed, 2005) and Table 2 presents the possible distinctions between references and citations based on

different theorists' comparisons of the act of referencing a work and measuring citations in a citation analysis.

Table 2. Views on what is measured by references and citations

Author	*References conceived as*	*Citations measure*
Garfield, Salton	Descriptors of document content	
Garfield	Manifestations of scholarly information flows	Utility (quality of formal use).
Small	Elements in symbol-making process	Highly cited items as content symbols.
Merton, Zuckerman	Registrations of intellectual property and peer recognition	Intellectual influence
Cole and Cole		Socially defined quality.
Gilbert	Tools of persuasion	Authoritativeness.
Cronin	The character and composition of reference lists reflect authors' personalities and professional milieux.	It is unclear what citations measure; the interplay between institutional norms and personal considerations should be studied first.
Martin and Irvine	References reflect both influence, social and political pressures, and awareness.	Differences in citation rates among carefully selected matched groups (partially) indicate differences in actual influence.
Zuckerman	Referencing motives and their consequences are analytically distinct	Citations are proxies of more direct measurements of intellectual influence.
Cozzens	References are at the interest of the reward, rhetorical and communication system but rhetoric come first.	Recognition, persuasiveness and awareness generate a certain portion of variation in citation counts.
White	Inter-textual relationships mainly reflect straightforward acknowledgement of related documents.	Co-citation maps provide an ariel view and measure a historical consensus as to important authors and works.
van Raan	References are partly particularistic but in large ensembles biases cancel out.	The upper part of the distribution of a "thermodynamic" ensemble of many citers measures top research.
Wouters	The reference is the product of the scientist.	The citation is the product of the indexer. Validity of citations cannot be grounded merely in reference behavior.
Moed	Citations and references cannot be considered theoretically distinguishable. The citation is not just the product of the citation indexer but is also the manifestations of intellectual influence of the scientist.	
Albarrán		Citations are income. In distributions their value can fall above or below a critical citation line.

Source: The table is adapted from Moed (2005, p.194).

The view of citations in the evaluative context of ALI does not aim at capturing the motives of individuals, but rather their consequences at an aggregate level as discussed in (Moed, 2005, p.221). This embodies a shift in perspective from that of psychology of why researchers reference a particular work or considerations of why they cite a particular paper or author, towards what

researchers jointly express about the structure and performance of scholarly activity (Moed, 2005, p.210).

3.2 Rationale for using citations in ALI

Citation and reference theory may in practice be indivisible. Citation and reference theories are concerned with the meaning of a citation or reference, because researchers apply different aspects of a cited work and it follows that a cited work may be used by very different networks of researchers, the motivation to cite may vary over time, and the use of the content of the cited work may be applied in different contexts. Even if the difference between the two aforementioned theories is only semantic or one of emphasis (de Neufville, 2010) the fact is that in ALI development and application we *do* differentiate between the truth and the practical usefulness of the theoretical meaning of the act of citing and the act of referencing. Practicality is an important characteristic of the rationale for using concepts of citation and/or reference in ALI.

The basic Mertonian rationale for using citations to measure links between publications is that researchers must cite the work they draw from and citations are embedded in the reward system of science where quality is being rewarded, therefore citations become indicators of research published in papers that have been approved as passing peer judgements of quality by the peer review system.

> "For if one's work is not being noticed and used by others in the system of science, doubts of its value will arise."
> (Merton 1977, 54-55)

In other words, citations are interpreted to reflect some measureable aspect of an implied scientific quality that is used as a proxy measure for quality in total (Bornmann et al., 2008b; Nørretranders, 2007; Cole & Cole, 1973). Citations are embedded in how scientists socially and cognitively construct their work: they cite to persuade: to advance interests, defend claims, convince others (Gilbert, 1977). Zuckerman replies to Gilbert, that even if a citing author intends to persuade, the reference may still express intellectual influence, not one or the other. Cozzens (1982) built further on Zuckermans work, and detailed a rhetoric-first model that separated the motives of citation into "reward" and "rhetoric". The multidimensionality of citations has important implications for the analysis and interpretation of citation based indicators, as citation count indicate "impacts" not just "impact" of publications. Martin and Irvine define "impact" as a measurable aspect of quality and accordingly differentiate between research quality, importance and impact (Martin & Irvine, 1983).

They recommend the preferred term "citation impact" rather than "impact" because citation rates constitute just one indicator of one type of impact of a work and the citation impact of a publication limits interpretation to the "actual influence on surrounding research activities at a given time", (Martin, 1996; Martin & Irvine, 1983). Defining just what is meant by "impact" when using citations provides a rationale for ALI to measure the authority of a set of papers and transpose this to the authority of the researcher. But citations are in themselves an imperfect measure, as they are influenced by so many other factors and by linking to the notion of "citation impact" rather than "quality" in indicator development and application, indicator developers may rationalize the continued use of citations. Thus, whatever the motive to cite, authors cite *some sort* of cognitive worth of sources and "citation impact" captures an aspect of this "worth". From this perspective for example, even if a paper is found to be methodologically flawed and thus lesser "quality" – but it has passed through peer review - it can still stimulate future research and stimulate scientific progress, and the amount it has been cited can be a justified indicator of its worth. Furthermore, a prestigious researcher that is highly visible will attract citations, even though the "quality" of the research may be no greater than that of lesser known, less visible researchers. Zuckerman argues that the citation is still predominantly a measure of intellectual influence and worth, and not visibility dynamics (Zuckerman, 1987). But twenty years later Aksnes (2009) claims that the "effects of visibility dynamics are not insignificant compared to those of quality dynamics". He proposes that the passage of time effects the concept of citation, and visibility dynamics cannot be disregarded, relegated or separated from quality dynamics, just like the reference cannot be separated from the citation.

If a source is frequently cited the *worth* of the source is growing in influence, which is indicated in high citation rates thus making these sources authoritative. Authoritative sources tend to be authoritative because of their influence upon practitioners in the field, and this is reflected in their high citation rates. Aggregating citations across publications, as in ALI, rationalizes citations as suitable proxies of the performance of a researcher's work in assessments. When Garfield launched the citation index, a bibliographic system for communicating and evaluating science, he showed that citations could be used to trace the development of ideas by using references as indicators of document content, and from the point of view of the cited document an expression of utility based on ways in which and how frequently they are cited and co-cited and these links could be used to define fields, networks and influential journals (Moed, 2005; Garfield, 1979; Garfield, 1964). Inspired by Robert Merton and Manfred Kochen, Garfield considered a citation as an associative

measure of "intellectual transaction" (Merton, 1973) and by Small's concept of citation as "concept symbols of information" (Small, 1978):

> "[….] citations symbolize the conceptual association of scientific ideas as recognized by publishing research authors. By the references they cite in their papers, authors make explicit linkages between their current research and prior work in the archive of scientific literature. [….] *These* explicit references imply that an author has found useful a particular published theory, method, or other finding." (Garfield, 1994). The quote is shortened, italics denote changes.

Garfield's rationale is that citations can be aggregated in to measures of "utility" through linking citations extrated from a publication's reference list, indexed in his science citation index, directly with the real world referencing practices of researchers (Garfield, 1970, p. 670). As the now named WoS citation index is still to this day the authorative source for bibliometric data, and used prolifically in indicator development, Table 6, it is worth considereing how WoS rationalize the use of citations in bibliometric indicators. It does not stray far from Garfield's original definition:

> "Citation counts are a formal acknowledgement of intellectual debt to earlier patents and previously-published scientific research papers. They are an important indicator of how new patents are linked to earlier patents and scientific papers[17]."

The different assumptions of what a citation means may challenge the interpretation of indicator scores. Therefore in the development of ALI, citations can advantageously be reserved to explicitly test some assumption (Moed 2005, s.195) and the interpretation of the meaning of the citation may be postponed (Wouters, 1999). The value of the citation is not then determined by whether they are literally true or correspond to reality in some sense, but by the extent to which they help to make accurate empirical predictions or to resolve conceptual problems. As I understand it, instrumentalism holds a central role in the development and interpretation of ALI. The term citation in this perspective is properly reserved for a measure that explicitly tests some assumption, hypothesis or theory; for in citation analysis, these underlying assumptions, hypothesis or theories usually remain implicit (Holton, ibd. Moed 2005, p.195). Cole and Cole's approach was to use

[17] Definition from the Glossary of Thomson Scientific terminology, available at: http://ip-science.thomsonreuters.com/support/patents/patinf/terms/#C

citation analysis as research tool, e.g. to make the underlying assumptions explicit by testing hypotheses related to social stratification and related issues (Cole & Cole, 1973; Cole & Cole, 1967). However the use of citations as sociological research tool should not be directly transferred to their application in an evaluative context, which is the context of ALI in which the research performance of individuals is assessed. The sociological approach reveals a structure, the evaluative approach leads to statements about the performance of a particular researcher in the research system where indicators operationalize aspects of an abstract, theoretical concept "research quality". The validity of the ALI can be empirically tested by 1) correlating them to other more direct measures of the concept, and by 2) examining the relationships among variables. Though I would not go so far as to adopt an instrumentalist point of view, which calls into question whether it even makes sense to think of theoretical terms as corresponding to external reality (de Neufville, 2010), I will adopt the perspective that ALI may be thought of primarily as tools for solving practical problems in evaluation and the theories of citation and reference may then be used to facilitate appropriate selection and application of relevant approaches in the implementation and interpretation of the ALI. This was suggested by Wouters in the reflexive indicator theory (Wouters, 1999a; Wouters, 1999b).

Wouters proposed that a reflexive indicator theory would explain how different theories of citation and reference are related to one another, and this can be applied in the development and application of ALI:

> "[…] it is a theory about the indicators themselves, starting from the analytical distinction between the reference and the citation" (Wouters, 199a, p.576)

He suggests that reference behaviour is from the perspective of the citing documents and their author, while citation counts are from one document to another. The citation produced by the author is not identical to the citation as a product of the indexer. In this sense, the reference belongs to the cited text, but in the citation index, the references are no longer organised according to the documents in which they were contained, but according to the documents they point to (Moed, 2005; Wouters, 1999a; Wouters, 1999b). They become attributes of the cited instead of the original text. Wouters' approach builds further on two concepts of information that indicators contain: Citations as a formalized representation, rationalized as a concept of information in a formal entity from which all meaning is purged. The second is the paradigmatic concept of references, focusing on meaning and embracing the concept of information as defined by Bateson as "any difference that

makes a difference". In Wouters' view the values ALI produce constitute a "formalised" science representation of citations to publications that initially neglects meaning, because citations during the indexing process become disparate from references. In order to be useful one has to allocate meaning to the citation again, but the main point is that this attribution of meaning can be postponed. Meaning can be re-attributed through the different citation theories. Thus the reflexive indicator theory guides the process of translating research into citations and references into practice; understanding and/or explaining what influences the outcomes of ALI by drawing on classic theories, e.g. Merton, Zuckerman, Cozzens, and by drawing on implementation theories, e.g. Garfield, White or van Raan.

> "Because of the emergence of the formalized representations, stimulated by the creation of SCI, multiple relations have been created between the formalized and the paradigmatic representations of science (and technology). Every existing science or technology indicator theory is the embodiment of one possible type of relation within the domain of all possible relationships. Encompassing all this is not a sociological theory, but simply this proposal: to recognize the two different domains, to position each indicator theory accordingly and to establish their interrelations" (Wouters 1999, p212-213).

The primary rationale of using citations in ALI is that the effect of knowledge becomes measureable in an objective way and also "query-able". Here I agree with Wouters (1999) that to interpret these representations one needs to attribute meaning again. Moed (2005, p.201) disagrees with Wouters that citations and reference are two analytically *independent* research problems, i.e on one hand the study of patterns in the citing behaviour of scientists, social scientists and scholars and on the other the theoretical foundation of citation analysis. He attempts to extend Wouters' theory, by focusing on the appropriate use of citation indicators in research evaluation. Moed (2005) stresses the importance of identifying which factors account for the skewness of citation counts amongst papers and how these factors are related to research performance. Moed's consideration is important, especially if the increased use of ALI is heading in the direction of an evaluative-economic position, later suggested by Albarrán et al (2011). In Albarrán's opinion, the evaluation culture has brought us to a situation where instead of individuals we now have papers and instead of dollar-signs we have citations. In this new rationale, citations are interpreted as income distribution instead of dollars. Once we take this step, we can predict that the measurement of the low impact

researcher could coincide with indicators of economic poverty, because they have citation impact below a crucial citation line, whereas the measurement of high-impact researchers will be identified with notions of economic affluence. This last example illustrates clearly how attributing a definition of what citations mean can have serious implications for the development, application and interpretation of ALI.

3.3. Implications for ALI

The meaning attributed to citations and references has direct implications for the development of ALI. Cozzens identified citations as part of two systems: the reward system of science, adhering to citation etiquette and the rhetorical system where authors strategically reference other works. The rhetorical system, she claims, is dominant as it also praises colleagues, and includes authorship skills that can lead to promotion and grants and awards. Indicators aimed at measuring the reward system should be constructed in such a way that the effects of the rhetoric system are taken into account. With reference to Cozzens' rhetoric-first model, (Cozzens, 1989), Moed doubts that separating the rhetoric and the reward aspect in separate indicators is possible:

> "[…] even if some rhetoric and communication factors can be separated, there are doubts that this could be done with reference to the reward and the rhetoric systems, as citations reflect both aspects at the same time." (Moed, 2005, p. 215).

The inherent multidimensionality of citation limits the separation of communicative, reward and rhetoric factors and challenges the development of indicators that aim to measure just one of these three attributes. Citations reflect many dimensions at the same time (Leydesdorff, 1998; Leydesdorff & Van den Besselaar, 1997). Leydesdorff noted that different intrinsic functions of citation may be technology-specific, different interactions with the technology and the source generates the variation and the variation may change over time with the use of the document. Various interpretations could be equally valid at the same time and combining indicators into a reflexive model as suggested by Wouters (1999a;1999b) may provide a foothold for understanding. But he is clear that the only way to identify the dynamics of indicators in performance analysis is to test theoretical assumptions. The indicators cannot be thought of as given from above or detached from the theoretical framework, or as unable to undergo changes in actual use. Moed (2005, p.55) agrees that indicators should preferably be developed in response to and as aids in the solution of

interesting questions and problems. Holton also advocated plurality in theoretical development, allowing for a diversity of models and corresponding indicators:

> "The absence of any explicit theory to guide the making and use of indicators may not be good; but the adoption of a single one is likely to be worse." (Holton, ibd. Moed, 2005, p.57).

How should indicator developers theoretically ground ALI as notions of impact, significance, or influence and account for the insights obtained from ALI? Cronin (1984) has two suggestions: an internalistic approach, that devises indicators of distributions and quantities and the externalistic approach, which approaches indicators of social contexts, indicators of the processes authors use to compile their reference lists. Two separate approaches to indicator development could result in one disjunct approach to indicator development, where the robustness of the mathematical foundations of indicators are argued separately from the theoretical robustness of the indicator, see the paper by (Sidiropoulos et al., 2007) describing the *ht*, *hn*, *hc* indicators. However, the fact that citations are a function of so many influencing factors, socio-cognitive behaviour as well as "scientific quality", makes it difficult to cleanly separate an internalistic and externalistic approach to indicator development and interpretation. A high citation count, for example, may not equate equally with high quality or low citation count with low quality. Thus, to claim (Bornmann & Werner, 2014; Bach, 2011; Bornmann et al., 2008b; Martin & Irvine, 1983) externalistic theories must also be considered in the internalistic indicators and vice versa. Only then will the relationship between the citation and social organisation of science be contextualized (Leydesdorff, 1998).

Another implication for ALI development is the increased awareness of individual researchers in the use of citation indicators in research policy (Collini, 2012; Dahler-Larsen, 2012). Bibliometric analysis is one of several research techniques used to evaluate researchers and research techniques are not theoretically neutral. Interpretation is based on implied theory of publishing, citing, referencing, social interactions, and how the indicator is used can influence outcome in many ways not to mention effects on the individual (Paper 4). The most important resource for researchers is their knowledge, and an indicator theory such as Wouters' (1999) has the potential to support the development of indicators that capture the effects of a researcher's knowledge in a systematic manner (Preece, 2011). At the individual level it is not generally an option to rely on large aggregates of data to cancel out the individual vagaries in citation behaviour and effects of data

incompleteness (White, 2001; Small, 1987; Cawkell, 1976) and anyway Moed reminds us of the importance of not rest on the assumption that errors and violations of norms can be concealed and neutralised by using large datasets (Moed, 2005). It is fundamental in indicator development to understand, as also claimed by Zuckerman (1987), if such phenomena as violations and errors are randomly distributed among all subgroups of scientists, or whether they systematically affect certain subgroups (Moed, 2005, p. 216). In a study by Cozzens (Cozzens, 1982), for example, two contexts of citation were identified in a small aggregate of bibliometric data: 1) as a general contribution and 2) as a specific contribution. As these two contexts can be of unequal size and interpretations of citation indicators should be relative to social construction of science relative to the individual and his or her specialty.

Therefore the main implication for ALI, is to identify indicators that work together to present a balanced picture of the researcher's multidimensional impact and interpret indicator values by matching like with like. This would involve calculating many ALI rather than a single one, and operationalizing different meanings of citation; for example, indicators that measure the internalistic qualities, i.e. indicators that measure the lower and/or upper ends of the distribution and average based indicators, together with indicators that capture externalistic properties, e.g. the age of the citations, percent of uncited papers. Combining partial indicators in this way involves describing the biases and errors in any of the indicators used, where the indicators converge or diverge, but this still does not guarantee that the outcome is free from bias (Martin & Irvine, 1893, p.87). Dahler-Larsen (2012) argues the effects of database and indicator construction are not random influences, but constitutive effects, and these biases and deficiencies cannot be repaired to a tolerable degree at the individual level, as investigated in Paper 6, and may be an ever present bias in ALI (White, 2001). Finally, Wouters' reflexive indicator theory does not assume the primacy of one theory, but creates theoretical openness by proposing a framework in which each approach finds its proper place. Likewise the development of science indicators in the multidimensional context of attributing meaning to citations does not necessarily result in consensus upon what ALI are measuring and which ALI ought to be applied in a research policy context. Kahneman (Kahneman, 2011) suggests we maintain faith in the measurements indicators propose no matter how abstract they are, because the indicator culture is sustained by a community of like-minded believers. This rhetorical approach to the theory and concept of indicators preserves the uncertainty of findings (Starbuck, 2006, p.78) and consequently there is never closure and ambiguity always persists (Starbuck, 2006; Wouters, 1999a). As a result we have a theoretically underdeveloped understanding of what the results of these indicators mean. The call for an indicator theory is itself indicative of the urgency to explore

more systematically the relations between the design and use of scientometric methods and qualitative approaches in research assessment (Moed, 2005; Leydesdorff, 1987). But as of yet an indicator theory has not been accepted (Hicks et al., 2015; Wouters, 1999a; Leydesdorff, 1987). Wouters concludes this lack of acceptance is due to many social scientists holding vested interests in specific theoretical positions that would become redundant in indicator theory (Wouters, 1999a; Wouters, 1999b). On the other hand Starbuck (2006) reasons that social scientists are not this contrived, but are unaccustomed to projecting their ideas onto shared frameworks and they would have to learn new ways of thinking and speaking to enable a theoretical framework. Some social scientists have expressed doubts about the validity of theoretical propositions of any kind, questioning if shared certain behaviours attaining to authorship or citing in some situations, predict behaviour in all circumstances. Starbuck (206, p.167) surmises that the simplest road to a theoretical framework and rationalization and operationalization of concepts within the said framework would require explicit actions by key journals to act as professional gatekeepers. He calls for journals to refuse to publish indicator studies or proposals for new indicators that do not adhere to or reaffirm baseline positions. We have not yet reached the extreme situation Starbuck suggests, but the idea of affirming baseline positions is the major concern of this PhD work.

The theoretical discussion in this chapter, the discussion of concepts in Chapter 2, the empirical analyses in the 7 papers, previewed in the next chapter, together with the validation analysis in Chapter 6, will inform RQ1: the characteristics of ALI, and RQ3: the extent the concepts being measured are defined in indicator construction. Further, the discussion in these chapters contributes to the overall objective of this PhD work on the extent ALI are appropriate and if the character of their models indeed capture the performance of the researchers the indicator purport to measure. Or, to play devils-advocate, if indeed these indicators are artistically-crafted illusions of validity that are supported by a powerful, professional bibliometric culture (Kahneman, 2011).

Chapter 4:
A preview of the research contributions

The previous three chapters reviewed fundamental issues that can affect the development, application and interpretation of ALI. How the concept of author, publications and citations are operationalized in ALI was discussed, and the assumptions behind the rationale of using citations as proxies of "impact" and/or "influence" were addressed. These chapters inform the characteristics of ALI and the extent ALI are appropriate measures in the evaluation of research performance. Are indicators using sub-standard methods to evaluate researchers that they in turn would condemn the researchers under evaluation for using? Yet these chapters have focused solely on the semantics of indicators. Semantics studies are relevant as semantics is basically about concepts, the meaning we give to various elements of ALI. The previous chapters introduced issues which we can use to supplement our judgements of the success or otherwise of an indicator, i.e. whether the indicator can be relied upon to accurately embody the knowledge of human experts it purports to measure. Simply put, if we can't tell how key concepts are operationalized we cannot trust the results. In a thesis about the appropriateness of indicators, to be able to defend the value the indicator and our trust in its results, empirical studies of indicator performance are necessary. The present chapter presents the Ph.D's 7 research papers that empirically explore the indicator characteristics and performance, Section 4.1. The full papers can be found at the end of this thesis, after the references. The papers use a variety of methods to enable us to learn more about the challenges and limitations of disciplinary and seniority appropriate ALI as well as addressing the research questions. The description of the papers is followed by a summary of the main recommendations of the ACUMEN WP5 Section 4.2.

4.1 The research papers

The papers are presented in chronological order and consist of 4 journal papers and 3 conference papers. Paper 1 was co-authored with Jesper W. Schneider and Birger Larsen. The final report from the ACUMEN Work Package 5, Novel Bibliometric Indicators, is included as Appendix A. The report is included as an appendix, as ACUMEN was the premise for this PhD and it contextualizes the papers. I was the first author and main investigator in the sub-reports presented in this report, apart from the study in part 2 "Star Researchers", in which I had no involvement.

Table 3. The research contributions included in the PhD work

No.	Year	Title	Reference	Type
1	2014	A review of the characteristics of 108 author-level bibliometric indicators	Scientometrics, Doi: 10.1007/s11192-014-1423-3	Literature review
2	2014	Table of author-level bibliometric indicators:	e-material to Paper 1. Scientometrics, 11192_2014_1423_MOESM1_ESM.docx	Methodological analysis of indicator composition
3	2014	Just pimping the CV? The feasibility of ready-to-use bibliometric indicators to enrich curriculum vitae	Proceedings of the Iconference, Breaking down the Walls, Berlin http://hdl.handle.net/2142/47339	Descriptive analysis of dataset Predictive analysis Case study
4	2014	The effects of "ready-to-use" bibliometric indicators	Proceedings of the STI, pp:687-691 sti2014.cwts.nl/download/f-y2w2.pdf	Literature review
5	2014	Scaling analysis of author level bibliometric indicators	Proceedings of the STI, pp.692-701 sti2014.cwts.nl/download/f-y2w2.pdf	MDS maps, Performance mapping Scaling analysis
6	2015	A comparison of 17 author-level bibliometric indicators for researchers in Astronomy, Environmental Science, Philosophy and Public Health in Web of Science and Google Scholar	Scientometrics, Doi: 10.1007/s11192-015-1608-4	Concept of average. Rank correlation and standard difference in rank position. Empirical validation.
7	2015	A critical cluster analysis of 44 indicators of author-level performance	Preprint available at: http://arxiv.org/abs/1505.04565	Two step cluster. Ordinal Regression. Odds analysis. Correlation analysis.

Table 4. Appendices

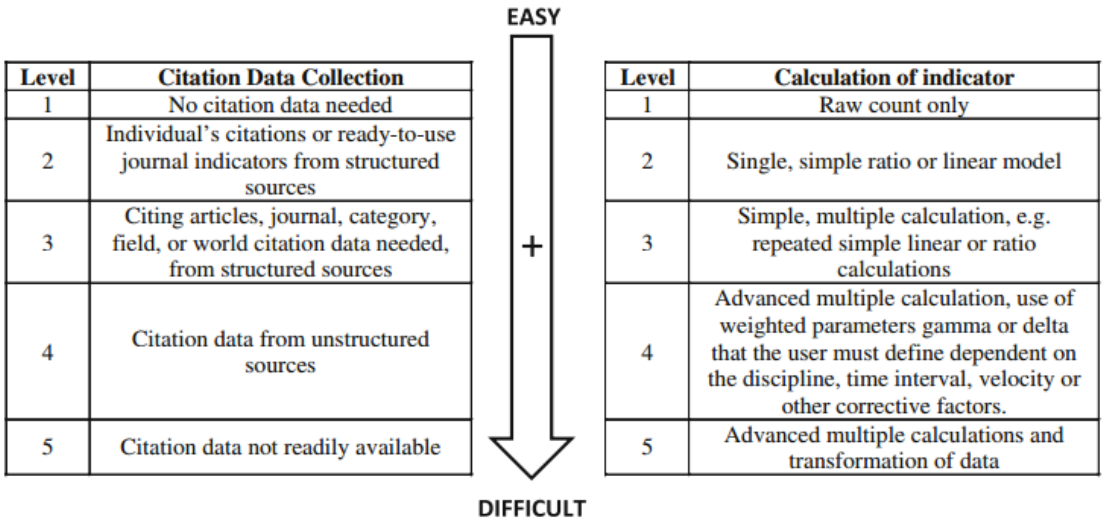
ID.	Title	Type	Attached
A	Deliverable D5.8: Novel bibliometric indicators	Final report of ACUMEN Work Package nr.5	Print appendix
B	Validation chart of bibliometric indicators	Methodological and theoretical evaluation of 68 ALI	e-material, link: http://tinyurl.com/nj4mvca

The disciplinary and seniority appropriateness of ALI (RQ2) was explored using a dataset based on CV data and bibliometric data from researchers in Astronomy, Environmental Science, Philosophy and Public Health; Seniorities were classified as PhD students, Post Docs, Assistant Professors, Associate Professors and Full Professors. These grouping provided analysis to a set of researchers active in the social sciences, natural sciences or humanities at different stages of their academic careers. One issue is what is it exactly the indicators are measuring – is it a concept of academic performance, the success of the mathematical foundations of the indicator to "fit" the bibliometric data or vice versa, or are the measures arbitrary as their value relates to database performance rather

than various aspects of researcher performance? (RQ1 and RQ3) Another issue is identifying appropriate indicators at the individual level that successfully capture the nuances of disciplinary publishing and citing traditions, notwithstanding the fact that the small amounts of data and skewness of citation distributions used as input in these indicator models may steer the researcher's resulting "score" (RQ1 and RQ2).

Paper 1 and 2 consider RQ1, which is concerned with the characteristics of ALI. The two papers together present a detailed analysis of the aims and computation of 108 ALI. The indicators are broadly categorized into indicators of publication count, indicators that qualify output (on the level of the researcher and journal), indicators of the effect of output (effect as citations, citations normalized to field or the researcher's body of work), indicators that rank the individual's work and indicators of impact over time. Each indicator was rated on a five point scale from simple to very complex, in both data collection and mathematical computation, Figure 2. Three important issues were highlighted: 1) the availability and accessibility of publication and citation data does not support the practical application of indicators; 2) Indicators lack appropriate validation and recognition by both the bibliometric and academic community, and 3) the assessment of publication performance cannot be represented by a singleindicator. The aim of the papers was to describe the indicator model, create an overview of which effects of publication activities the indicators are designed to measure, and how complex the indicators are to calculate.

Figure 2. Five point scale assessing two aspects of the complexity of ALI



Level	Citation Data Collection
1	No citation data needed
2	Individual's citations or ready-to-use journal indicators from structured sources
3	Citing articles, journal, category, field, or world citation data needed, from structured sources
4	Citation data from unstructured sources
5	Citation data not readily available

Level	Calculation of indicator
1	Raw count only
2	Single, simple ratio or linear model
3	Simple, multiple calculation, e.g. repeated simple linear or ratio calculations
4	Advanced multiple calculation, use of weighted parameters gamma or delta that the user must define dependent on the discipline, time interval, velocity or other corrective factors.
5	Advanced multiple calculations and transformation of data

Paper 2 presents a schematic overview of the indicators presenting: the definition of the indicator by its creator, the objective of the indicator, its advantages and disadvantages, comments and references to relevant literature, and the complexity rating. Seventy-nine of the indicators were rated simple in data collection and calculation and included clear definitions of how they measure or interpret certain aspects of performance amongst others, *h, g, AR, AWCR, alternative h, f, t, CPP, IQP*. The remaining 29 indicators were rated as being complicated in data collection and calculation. They demand access to complete publication and citation data often from sources not included in generic citation databases as in *knowledge use,* are mathematically difficult, e.g. *generalized h* and *adapted pure h,* require establishing parameters that represent publication and citation practices in specific fields, hα, or require specialist software to compute the indicator, *tapered h*, and *h-sequence and matrices*. Even though these same indicators in tests show their superiority over simple indicators by correcting for mathematical consistencies, providing granular comparisons between researchers and embodying the inertia of the objects they are designed to measure, their application in practice is severely limited by their complexity. The simple indicators, of which there are many, may be coarser or some even mathematically flawed, but as a set they offer great potential for well-rounded author-level bibliometric assessments, especially when used to supplement each other. This is why the technical issues with their mechanisms are further explored in Papers 3, 5, 6 and 7.

Simple indicators are the focus of analysis in the remaining original studies as the objective of this thesis is to determine a set of ALI appropriate for application by end-users. Research question 2 explores the extent author-level indicators are appropriate in the evaluation of researchers from different disciplines and different academic seniorities? Accordingly, in Paper 3, I begin to explore the feasibility of indicators to provide value-added information to curriculum vitae and discuss the dependency between citation indices, the characteristics of the researcher and indicator scores. In Paper 3 the 750 researchers in the dataset, described in Chapter 5, were ranked using 10 simple ALI that can be automatically calculated in citation indices or using free software. The h, g, e, AW indicators showed a predictive relationship, i.e. if you score high on one, you will score high on the others; low on one, you'll score low on the others and these indicators reward researchers with the mathematical ratio "short career length to many papers to high citation count" with the highest scores (Paper 3). In this Paper the possibility of disciplinary specific indicators shows potential, whereas seniority specific indicators already here seem unrealistic as the performance of the researchers in my dataset at the individual level was very different and highly affected by their

visibility in the citation index. Knowing that indicators have characteristics that favour the aforementioned ratio, means that simple indicators can easily be manipulated by for example administrators to promote or demote researchers in scholar rankings or by researchers themselves to increase indicator scores. Hence it is vital that ethical and sociological issues of author-level bibliometric assessment are not being ignored when considering the appropriate use of ALI (Paper 4). Paper 4 is a literature review of the psychological effects of quantitative evaluation undertaken in preparation for the ACUMEN Behavioural Codex for researchers and consumers using bibliometric self-evaluation, Appendix A, pp. 196-203. By linking empirical and conceptual personality traits commonly appraised in evaluations to author-level bibliometric evaluation, we become aware of the role bibliometric indicators can play in strengthening or weakening a researcher's self-esteem, self-efficacy and uncertainty. Likewise, in bibliometric evaluation the motive of the evaluation, evaluator or evaluand[18] is instrumental in the appropriate choice of indicator. Gender, age and culture differences and stereotypes are also reported to affect the appropriateness of author-level bibliometric evaluation.

Continuing the empirical analyses started in Paper 3, Paper 5 continues to study the overlap and redundancy between indicators, investigating methodological challenges in analyzing and interpreting trends in the data. Fifty-two indicators were calculated for each researcher in the dataset. No association between indicator score and seniority was found. On a disciplinary level, the performance of each indicator as a ranking parameter was explored, and I discovered that indicators have a characteristic behaviour – some have a central, controlling quality that determine rank placement of researchers in rankings, e.g in Astronomy the *hg*-index, in Environmental Science the *h*-index, in Philosophy the *IQP*-index and in Public Health the *g*-index, while others are isolated and produce random, indiscriminate rankings. These isolated indicators, e.g. *%not cited*, *Price Index*, or *%self-citations* are interesting as they measure *something* different than the central indicators. Consequently not all indicators are equally appropriate as ranking parameters and continuing the exploration of RQ2 Papers 6 and 7 analyze further the strengths and weakness of indicators in disciplinary researcher rankings. Paper 6 compares researcher rankings using 17 ALI in two very different citation indices, WoS and Google Scholar (GS) and questions if the different counting methods indicators employ affect *our* concept of the "average" researcher. In the first part of the paper disciplinary coverage in WoS and GS was explored and the very different picture the same indicators in these two databases give of researcher performance was discussed. In the second

[18]The evaluand is the object of evaluation (Dahler-Larsen, 2012, p.6)

part, disciplinary averages used to benchmark expected performance were studies and they varied greatly dependent on the type of mean used (harmonic or arithmetic). In a retrospective study lack of fit (summed) between predicted harmonic and arithmetic h-index scores and previously published empirical data was observed. Despite of these differences, indicators that provide cross-database stability in rankings were identified, primarily the *hg* index, and the inherent mathematical characteristics that enabled this stability were studied. Finally, Paper 7 contributes to previous research by investigating the appropriateness of hierarchical clustering as a method to identify groups of similar researchers and analyzes the disciplinary and seniority appropriateness of 44 ALI. This analysis combined a two-step cluster analysis, ordinal regression, odds analysis and correlation analysis to explore the validity of researcher performance measured statistically through indicator scores to researcher performance documented on the individual's CV. The statistical analyses were supplemented with a discussion of disciplinary publishing and citing preferences, and the relationship between the clustering algorithm and completeness of the bibliometric data on rank position, all of which influence and undermine the use of indicators to rank researcher performance. Particularly visible in this study is the observed disconnection between the prestige of the researcher reported on the CV and the prestige indicated by the calculated indicators. The study also confirmed that recommending seniority-specific indicators would not be possible, but the investigation into disciplinary-specific indicators is worth continuing as different indicators were stronger in different disciplines in ranking authors as low, middle, high and extremely high performers: in Astronomy the *h2* indicator, Environmental Science *sum pp top prop*, Philosophy *Q2* and Public Health *e*. Again the mathematical ratio identified in Paper 3 ("short career length to many papers to high citation count") is suspected to be instrumental in cluster placement.

It is a general note, that I am aware that different methodological and statistical approaches may produce different results. Each of the statistical approaches used in these papers are composed differently and may produce different ordination or clustering results when used on the data. Statistical methods unexplored in this thesis and application of the same statistical methods using a different programme or on a different dataset could produce different results. This is why the sensible interpretation and application of the applied statistical methods are critically rationalized and discussed throughout the papers. The results of the studies cannot be directly generalized outside of this dataset but can inform future directions. This because the object of study is a convenience sample produced in a social system, with a lot of attrition rendering probability statements useless and biases challenging the external validity of the dataset. However, it is

important to do studies like the ones presented in this thesis, where the usefulness of statistical models and the application of bibliometric indicators are critically appraised. This will contribute to openness and discussion about the advantages and disadvantages of bibliometric analysis of researcher performance and perhaps help illuminate the inappropriate application of methods and hopefully stop the creation of superfluous indicators.

4.2 Summary of the ACUMEN Work Package 5: Novel bibliometric indicators

The preliminary work for this thesis was undertaken in ACUMEN Work Package 5 (WP5), which is briefly summarized in this section. The final report from WP5 is included as Appendix A. ACUMEN is the premise for this PhD work, provided access to a joint dataset from which the dataset used in this PhD work is extracted, and the findings and main conclusions of WP5 influenced the direction of the 7 papers included in this thesis. WP5 investigated the extent bibliometric indicators can be used in the evaluation of individual researchers and analyzed a wide range of bibliometric indicators such as indicators of production, citations, production & citations, production adjusted for time, production adjusted for field and several measures that describe different aspects of a researcher's publishing portfolio as a whole. Results are discussed in the perspective of three stakeholder groups: 1) the specific fields for which the object of ACUMEN research is most relevant, 2) policy makers and funding institutions, and 3) the academic community at large. WP5 assessed the *need* for the creation of new bibliometric indicators for the assessment of individuals and discussed ethical aspects of bibliometric assessment. Further WP5 contributed to the design, testing and content of the ACUMEN portfolio[19], provided guidelines for computing and interpreting basic indicators, and provides a behavioural codex aiming to guide informed bibliometric evaluation.

Task 5.1 Literature review

A state of the art report about the general development the field of ALI has achieved at the present time, leading to Articles 1 and 2. WP5 concluded there is no pressing need to develop new indicators for the measurement of the performance of individual researchers. A sufficiently large and diverse set of indicators are in use or have been proposed.

[19] A description of the ACUMEN portfolio is found via this link:
http://cordis.europa.eu/docs/results/266/266632/final1-acumen-final-report-29-april-2014.pdf

Task 5.2 Development of novel indicators

As there is no pressing need to develop new indicators, see Task 5.1, it is important to understand the indicators already in existence as well as their appropriateness for researchers in different disciplines and of different academic seniorities. Hence, Task 5.2 "the development of novel bibliometric indicators" is unnecessary. Instead, efforts are focused on recommending the best selection of current indicators. It required further analysis and a redefinition of the WP5 to understand which indicators are required and how these need to be combined to best express a researcher's performance. Hence, Task 5.4 "recommendation of selected indicators" was extended to include a study of the performance of 108 different indicators identified in the review across different disciplines and career stages. It is clear that using a single indicator (e.g. the *h*-index) and interpreting the results out of context of the researcher's field or seniority will result in distorted and useless information. The study shows that even though researchers prefer to use the indicators that maximize their impact and visibility, by providing a strategy of indicators for self-assessment, as well as locally relevant performance benchmarks, the researcher will reach a better understanding of the achievements of their published works and perhaps be able to identify where this can be improved.

Task 5.3 Selection of samples of researchers

The data collection process is described in detail in Chapter 5 of this thesis and Appendix A, D5.8 Part 3 – Selection of Samples. Observations during the data collection contributed to our understanding of the extent researchers used indicators themselves on their CVs. Finding data on individual researchers was difficult, and it was challenging to gather a complete picture of the researcher, as information was separated between personal homepages, institute homepages, PDFs and various online profile tools each with different "sell by dates". Any guidelines for evaluation practices (GEP) must for example describe basic retrieval problems, especially name ambiguity and data incompleteness, and describe how these affect the appropriateness of citation indicators and author-level metrics. Likewise, we cannot expect end-users are willing to sort systematically through two or more citation indicators and remove both duplicate publications and citations to get a "complete" publication and citation picture of a researcher's work. Further, the data collection showed how important personalization is. ACUMEN must encourage the researcher to explore different databases to understand their coverage in these sources and to be critical of what ALI automatically calculated in these sources represent. This must be made obvious to different types of users of the ACUMEN Portfolio as well.

Main recommendations for the ACUMEN Portfolio and GEP:

1. We cannot expect the researcher to sort through two or more citation indexes and remove duplicate citations to get a complete citation record.
2. Name ambiguity problems need to be described in the portfolio including how these affect the usefulness of citation indicators and ready-to-use metrics.
3. Researchers should be encouraged to have an ORCID id or Google Citation profile to ensure the researcher can easily claim his publications.
4. The ACUMEN Portfolio needs to have easy tools to import publication data.
5. The guidelines need to explain the calculation and interpretation of metrics, for all types of users of the Portfolio.

Task 5.4a. Consequences of the use bibliometric indicators: from the researcher's perspective and from the evaluator's perspective

See Paper 4.

When failures come to light, negativity can make the individual feel inadequate. If the quality of evaluation judgments based on standardized indicators is low, it may lead to assumptions about the productivity and citation impact of a researcher which can be unsubstantiated. Given that the results of bibliometric analyses are of personal significance to the individual, it is vital that the bibliometric community assesses if the appropriateness of these types of author-level bibliometrics is limited by psychological factors, such as the affects the assessment has on the researcher's self-worth or how cultural difference affect the value put on indicators. It is anticipated that the individual will seek and utilize whatever information is available to reduce uncertainty and to increase their subjective validity. If the individual provides substantiating, consistent evidence that informs the CV, the more stable it is and positive social comparisons can be implemented by the evaluator. When, however, an evaluator is met with a sporadic CV that lacks continuity, the researcher will likely receive a poor rating. Likewise, if only partial and unreliable information is used to calculate the indicator, the less valid or more uncertain the self-evaluation is assumed to be. Knowing which data is and is not included in indicators can reduce misinterpretation that could cause fabricated self-images and damaged reputations. Accordingly, self-image is the core concept of a CV as the CV is a proxy document for the researcher and is as such a space for researchers to promote their self-image.

As part of this task WP5 developed a Behavioural Codex for researchers and consumers using bibliometric evaluation, Appendix A, pp. 196-203

Task 5.4b. Consequences of the use bibliometric indicators: from the analysis of data collected in Google Scholar

See Paper 3.

This investigation is extended in the supplement to Task 5.4c. No gender-specific patterns were identified in the data and the women in our sample do not appear to need more years to advance the career ladder. PhD students do have enough citation and publication data or years of experience to use classic bibliometric indicators.

Two groups of indicators were identified. The first group showed predictive relations: *h, g, e, AW, m, mg* where a high, middle or low score on one indicator predicted a high, middle or low score on another. The *e, AW, m* supplemented *h* while *mg* supplemented *g*. The top 25%, middle 50% or bottom 25% researchers remained the same but ranked in a different order. The second indicator group was "unpredictive" indicators: *PY, m, P, C, CPP, CPAY*. For example, a low *P* did not result in a high *C* - likewise a high *PY* did not predict a high *P*. No individual or seniority patterns were found across this sub-group of indicators, and ranking resulted in different researchers appearing in the top, middle or bottom quartiles. No difference was observed between *CPAY* and *m*, resulting in redundant information. When we compared *CPP* to their rank position, we found the ratios within seniorities fit for the whole group, which in our dataset is a proxy for the disciplinary level. The expected performance of researchers according to their seniority varies by discipline.

Task 5.4c. Consequences of the use bibliometric indicators: from the analysis of data collected in Web of Science

See Paper 5.

Building on Task 5.4b 52 of the 108 indicators identified in task 5.1 were investigated to learn more about how they perform on data from WoS across four disciplines and five career stages. Using a hierarchical clustering model that illustrated how closely related the indicators are to each other, Task 5.4c discovered that indicators group together in descriptors of production, citations, production & citations, production adjusted for time, production adjusted for field and miscellaneous measures that describe the more subjective aspects of a researcher's publishing portfolio. The clustering of indicators is different from discipline to discipline, as is the strength of their relation. Before it is possible to recommend performance indicators for each discipline, the

role of the indicators within their cluster needs to be investigated: what they measure, if they overlap, how complicated they are, and which of them are redundant.

The 7 papers and the ACUMEN WP5 report presented in this chapter explore what the characteristics and mathematical construction of indicators mean for the performance of indicators and researchers in ranking. Further the possible effects of individual bibliometric evaluation drawing on lessons learnt through studies published in evaluation literature were explored. In the next chapter, the research approach and data collection is described before the empirical analysis of the validity of indicators that follows in Chapter 6. The quality of data collection has a particularly influential part in the extent conclusions of appropriateness of indicators can be drawn.

Chapter 5: Research Approach

5.1 Data collection techniques

The data set was drawn from the ACUMEN shared data set of 2154 researchers identified in a survey by Wp2. Briefly, WP2 conducted a large scale survey in 2011, resulting in information on online presence from 2,154 respondents, a response rate of 7.9%, see Table 5. WP2 extracted automatically a list of emails from published research papers indexed in the Thomsen Reuters Web of Science (WOS) during 2005-2011 in the four studied fields Astronomy, Environmental Science, Philosophy and Public Health, based on WOS subject categories, limited to European countries. Because of the low coverage of Philosophy in WOS the Scopus citation index was also sourced to get sufficient email addresses for this field.

Table 5. General statistics for online survey invitations and response rates

Disciplines	**Total email invitations sent**	**Total (%) responses**	**Response rate**
Astronomy	6,635	528 (24.51)	7.9%
Environmental Science	8,686	573 (26.60)	6.6%
Philosophy	4,591	519 (24.09)	11.4%
Public Health	7,277	534 (25.79)	7.3%
Total	27,189	2,154 (100%)	7.9%

Table 5 reports statistics for sent survey invitations and response rates across the four selected fields. Table 5 shows that the response rate is higher for Philosophy (11.4%) and lower for Environmental Science (6.6%).One explanation for this low response rate might be associated with the limitations of conducting online surveys based on email invitations on a large geographical scale (15 EU countries). Many respondents may consider email invitations as spam, although WP2 used a valid academic email for correspondence (@wlv.ac.uk) and an appropriate subject for the email invitations (e.g., "Philosophy web presence survey"). Moreover, WP2 embedded links in the invitation message to the key related information about the project including ethical clearance, privacy policy and project contacts. Another reason might be the changing of emails over the time due to mobility of researchers (e.g., postdoctoral research fellows) or graduation of students (e.g., PhD. or masters).

5.2 Sampling strategy

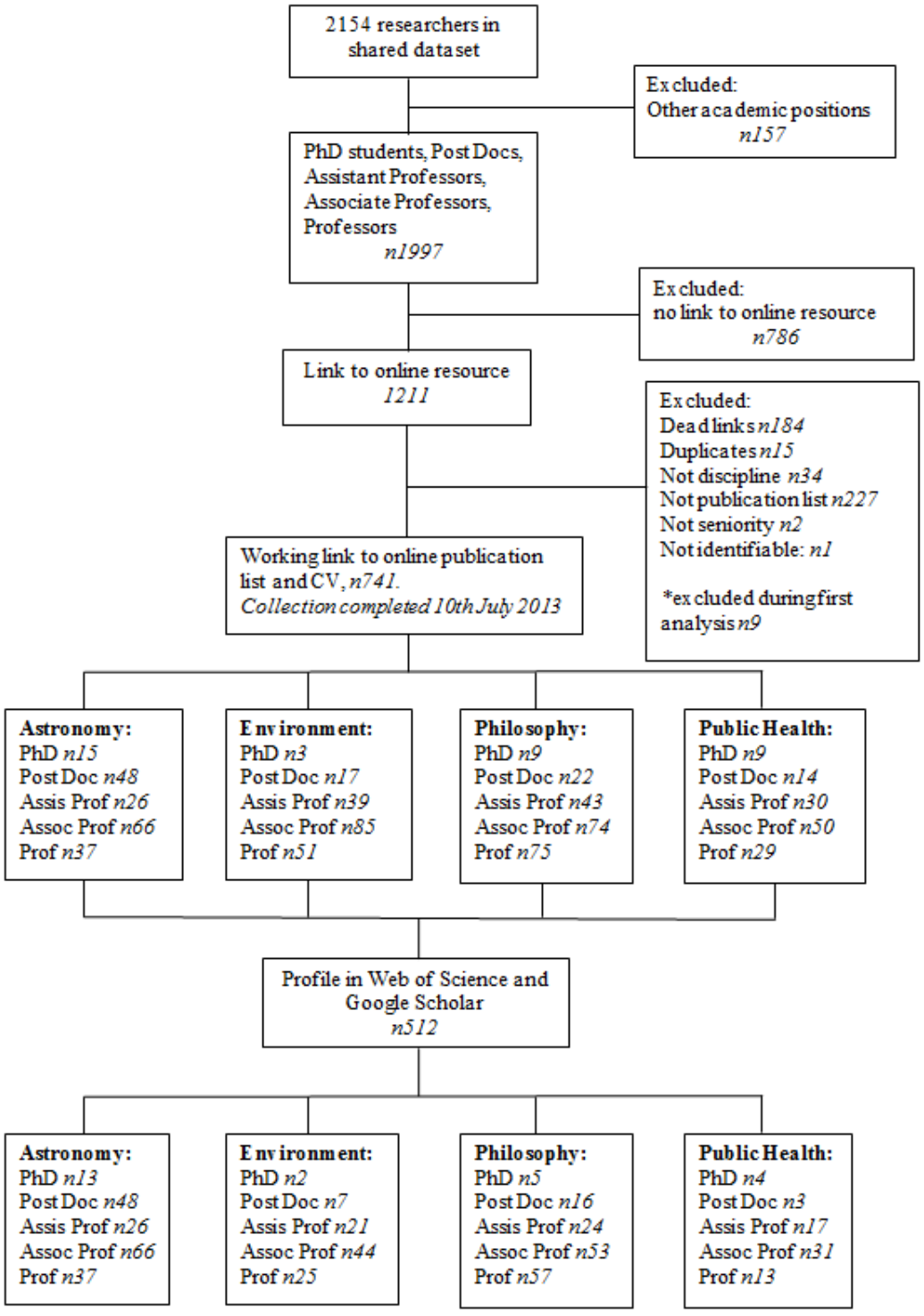
The overall aim of the PhD is to gain knowledge of the extent ALI indicators are appropriate measures of scholarly performance, and recommend a pallet of seniority and disciplinary appropriate ALI. Therefore the premise of analyzing the variance and relationship between indicator values, the researcher's career and research activities steered requirements to the data. Access to CV and publication lists was essential, and not all 2,154 researchers in the shared dataset provided this information. Therefore the shared dataset was reduced in the spring of 2013 to include only researchers', that provided a link to a CV and publication list, see the flowchart Figure 2. Online sources were prioritized as I wished to use dynamic CV and publication data to best represent where the researcher is at the present day rather than where the researcher was two years previously when the online survey was conducted. This meant that each link or links the researcher provided was manually searched and verified. CVs and publication lists were downloaded and stored in a closed database. The flowchart summarizes the data selection process, Figure 3.

Dataset 1

This dataset was used in Papers 3 and 5. To enable investigation of seniority performance and the possibility to recommend seniority-dependent indicators, only researchers who had defined their academic status as PhD Student, Post Doc, Assistant Professor, Associate Professor or Professor were extracted, resulting in a set of 1,211 researchers. The professional titles were limited to these five seniorities to ensure we could investigate potential correlations or trends in academic life cycles and bibliometrics. The titles were updated using information on the researcher's CV, university profile and publication list. All links were followed to verify if they actually led to a CV and publication list. This led to a further reduction of the dataset as the following were excluded: dead links, duplicates, links to materials that were not an individuals' publication list or CV including a list of publications, not one of our identified 5 academic status' or research areas that fell outside our four disciplines. The characteristics of the resulting 750[20] researchers are described in detail in Appendix A, part B pp.155-183, pp.205-211. Each researcher's publications and citations were sourced in WoS and GS to enable comparisons of indicator values computed with data from a structured citation database (Web of Science) with citation data retrieved from a web-crawler based index (Google Scholar). Google Scholar was searched using Harzing's Publish or Perish.

[20] During the first analysis, 9 further scholars who were duplicates or whose work could not be classified under our 4 disciplines were also excluded, resulting in a dataset of 741 scholars.

Figure 3. Flowchart over reduction and specification of the shared dataset



The set resulted in 34,637 citeable publications from WoS and 72,557 citeable publications from GS, Table 8. The methodology used to search and extract the publication and citation data is documented in Appendix A, pp.124-134. Additional publication and citation information on articles and reviews in this data set was kindly provided for the purposes of this study by the Centre for Science and Technology Studies (CWTS) at Leiden University, the Netherlands from their custom version of the WoS. This custom database contains records from the Science Citation Index Expanded, Social Sciences Citation Index and Arts & Humanities Citation Index portions of WoS, and has been specially prepared for bibliometric analysis. The data delivered by CWTS contained a wide range of bibliometric indicators for each paper including field normalised indicators using CWTS standard procedures.

Dataset2

Dataset 1 was reduced further in Papers 6 and 7. I wished to compare the indicator values and scholar rankings of researchers with a profile in *both* WoS and GS, resulting in the exclusion of 237 researchers. This final dataset consists of 512 CVs and publication lists as well as demographic data (gender, affiliation, discipline/specialty, and academic status): 190 from Astronomy, 99 from Environmental Science, 155 from Philosophy and 68 from Public Health. Our sample produced 22,143 journal papers and received in total 423,371 citations from other journal papers in Web of Science. In Google Scholar it was possible to identify 52,227 publications and overall 746,985 citations, Table 9.

5.2.1. Sampling bias

The collected PhD work is best viewed as an extensive case study of indicator epistemology and validity. The results presented in the thesis body together with the statistical analyses of the indicators investigated in the papers, cannot establish any general claims about how all developers of indicators work or the extent every facet of indicator construction is professionally (scientifically) done. I used the ACUMEN shared dataset as it has been an aim of ACUMEN since the kick off meeting in 2011 to connect the work packages through analyses of the same set of data. In this way the findings of the work packages compliment and supplement each other in a way that the respondents and their bibliographic data are investigated through interviews, surveys, institutional documents, altmetrics and bibliometrics, and the effect of gender, discipline, age, and availability of data could be studied. For my work package, WP5, this meant that a sample was

drawn from the shared dataset and is as such defined as “convenience” sampling, i.e. a type of non-probability sampling which involves the sample being drawn from that part of the population which is close to hand rather than a probability sample, in which each researcher in the population has a known nonzero chance of being selected through the use of a random selection procedure. In a convenience sample one cannot control how well the characteristics of the sample (gender, age, race, education, etc.) match the characteristics of the larger population it is intended to represent. However, although convenience samples are not scientific samples, they have value if you recognize their limitations and are open about these limitations when reporting the analyses. Convenience sampling proved useful in documenting that a particular quality of a bibliometric indicator occurs within a given sample and detecting relationships among different phenomena without the complications of collecting a randomized sample. Therefore the papers in this thesis should be regarded as case-studies that test certain questions and explore the data to understand relationships and shortcomings; perception of researchers, indicators and trends associated with bibliometric analyses.

Sampling bias can influence the results in important ways. It is evident that even though the researcher publication and citation matching processes used to generate the dataset is incredibly thorough, meticulous and exhaustive, it is not based on a set of researchers sampled using probability procedures from a known finite population as is required in science to make generalizations, which is essential for inferential statistics e.g.,(Freedman et al., 2007) . Yet this is the same premise for the literature presenting indicators referenced in the process of this PhD work. Indicators in these highly technical papers are proposed using non-experimental, social science data sets and not a probability sample as required in inferential statistical analysis that could be used to generalize the superiority or inferiority of the investigated indicators to the general population. Which is why, papers proposing indicators nearly always call for “further tests” and advise caution in transferring the applicability of the indicator to other datasets. Because the majority of the indicators are validated in the natural sciences, the same appropriateness in the social sciences and humanities cannot be assumed. Throughout the PhD work, researcher CVs and previous empirical findings are used to argue for and against the results presented in this thesis, see Paper 6 for example where the investigations of the fit of the harmonic or arithmetic estimations of “average” researcher performance in Astronomy are compared to similar studies.

Equally I recognize the absence of a random data-generation mechanism in my study and do not submit the data to analyses using probability theory, significance tests or confidence intervals to

qualify the results as these are consequently meaningless (Schneider, 2015; Schneider, 2013b; Berk & Freedman, 2003). Also in the data collection process I have clearly documented that some publications important in the four disciplines were not available in the version of the WoS I had access to and how seriously misrepresented some researchers are in this version of the citation index. Especially Astronomy was affected by missing conference papers and in general the coverage of Philosophy was very poor. So like so many other social science data sets, the dataset used in this PhD work is a convenience sample and the representativeness of the data is affected by the structure of the citation index, e.g. version, scope, indexing; software issues, e.g. citation matching algorithm, method used to collect the data, syntactic matching of author names; computational issues e.g. inclusion of self-citations, time period covered by the citation index (Jasco, 2008); and the dependency of within database references on disciplinary citation habits (and vice versa) (Lancho-Barrentes & Guerrero, 2010). The value of results must then be seen in relation to previous comparable findings and the results as an effort to inform the practical application of ALI.

5.2.2. Challenges in the composition of the dataset

The convenience sample affects the types of analysis I can implement, the statistics I can use and the strength of the conclusions I can draw. These three issues are further affected by the extent the data is representative of the researchers in the sample. The limitations in the composition of the dataset are these:

1. The sample is weighted in favour of senior researchers.
2. The academic seniorities are unevenly distributed across the disciplines, Figure 3.
3. The disciplines are unequally represented, Tables 7, 8 and 9.
4. Harzing's Publish or Perish (POP) software was used to identify and export references from GS. The collection of data from GS was restricted by embargoes enforced by the Google group. This meant that the search for publications authored by a researcher was blocked when one thousands references was reached, meaning that not all possible publications by a researcher were found. Known publications reported on the researchers CV but missing from the list of retrieved documents, were verified one by one.
5. GS has removed the option to select specific subject areas, and therefore filtering the results in POP is no longer possible and had to be done manually through sorting titles, snippets

and publishers. The amount of retrieved documents was thus increased and the limits of 1000 documents regularly reached. A complete list of GS limitations can be found in Appendix A, pp. 105-106, 129-131.

6. The WoS data collection was severely affected by WoS indexing policies. There is a database bias towards international English language journals, and certain document types, primarily articles and the citation culture in article-based disciplines, Appendix A, pp107-109.
7. There is a strong bias in favour of long established publishers against recently started publications, independent journals and conferences and back issues of indexed journals are not accepted in the database (Clarke & Pucihar, 2012; TS, 2012). This means that for some senior researchers, the proportion of their publications that are indexed by WoS is as low. A further consideration is that journals are deleted from Web of Science throughout the year (TS, 2012). This represents historical revisionism, with publications and citations being effectively cleansed from the record (Clarke, 2008).
8. Publications and citation-counts are not cumulative, because they change not only upwards, as new documents are published, but also downwards, as venues in WoS and GS are deleted.
9. The supplementary data provided by CWTS, does not contain data from the Conference Proceedings Citation Indexes. We do not have additional data on 3,693 citable papers and these are subsequently excluded from the present analysis reducing the dataset. The exclusion of the 3,693 records that were mainly in conference proceedings had a great effect on the Astronomy sample. Some researchers lost up to 80% of their publications. Appendix A, p.219, presents a detailed overview.

6.3.1 Overview of indicators and benchmarks investigated in the theoretical and empirical analyses

Legend:
*included in the analyses in the PhD work. 51 ALI (hybrid indicators), 10 publication- and 8 citation-counting indicators were investigated, Appendix B.
no asterisk = the indicators are included in ACUMEN analyses only and discussed in Appendix A.

ID	Type	Abbr.	Indicator	Intention
Productivity metrics				
1*	Publication	**P**	Publication count	Total count of production used in formal communication
2*	Publication	**Fp**	Fractionalized publication count	Contribution
3*	Publication	**App, arithmetic**	Average number of authors per paper over all papers	Indicates average amount of collaboration per paper
4*	Publication	**App, geometric**	Average number of authors per paper over all papers	Indicates average amount of collaboration per paper
5*	Publication	**App, harmonic**	Average number of authors per paper over all papers	Indicates average amount of collaboration per paper
6*	Publication	**Noblesse oblige**	Last author gets 0.5 credit	Indicates importance of last author
7*	Publication	**Fa**	Only first of n authors of a paper receive credit equal to 1	First author credit
8*	Publication	**Pw**	Weighted publication count	Accounts for importance of different publication types for the specialty / discipline
9*	Publication	**Pts**	Publication count in predefined sources	Counts output in sources deemed locally important
10*	Publication	**Cognitive orientation**	Publication count in fields and subfields	Visibility in main fields, subfields and peripheral fields
Impact metrics				
11*	Citation	**C**	Citation count	Use (effect) of all publications
12*	Citation	**Cts**	Citation count in specific database	Indicates database context
13*	Citation	**C-sc**	Citation count minus self-citations.	Use of publications, minus self-use.
14*	Citation	**Sig**	Highest cited paper	Most significant paper in the scholars portfolio
15	Citation	**minC**	Minimum citations	Minimum number of citations
16	Citation	**Sc**	Number of self-citations	Amount scholar builds on own research
17	Citation	**nnc**	Number not cited	Non-effectual papers
18*	Citation/publication	**%sc**	Percent self-citations	Disambiguate self-citations from external citations
19*	Citation/publication	**%nc**	Percent uncited papers	Percentage work not cited
20*	Citation/author	**Fc**	Fractional citation count	Share of citations on multi-authored papers. Aims to remove dependence of co-authorship, all authors receive equal share of citations
21	Citation/time	**C<5**	Citations less than 5 years old	Currency of citations

Hybrid metrics				
22*	Citation/publication/field	**IQP**	Index of Quality & Productivity	Number of citations a scholar's work would receive if it is of average quality in the specialty
23*	Citation/publication/field	**Tc>a**	Times cited more than average (Part of IQP)	Actual times scholar's core papers are cited more than average quality of specialty
24*	Citation/publication/field	**NprodP**	Number of productive papers (Part of IQP)	Number of papers cited more frequently than average, in the specialty
25*	Citation/publication/field	**hn**	Normalized h	Normalizes h-index (to compare scientists across fields).
26*	Citation/publication	**C(t)**	Age of citation	If citations are due to recent or past articles
27*	Citation/publication	**Hw**	Quality	Square root of weighted citations to papers in h core
28	Citation/publication	**%PNC**	Percent not cited	If citations are due to a few or many articles
29*	Citation/publication	**CPP (CPAY)**	Citations per paper (normalized for age)	Average effect of each paper (normalized for publication history)
30*	Citation/publication	**f**	Harmonic mean citations to papers	Average effect of each paper
31*	Citation/publication	**t**	Geometric mean citations to papers	Average effect of each paper
32*	Citation/publication	**h**	h index	Cumulative achievement (productivity and impact)
33*	Citation/publication	**π**	π is 100^{th} the number of citations received by top square-root of ranked papers	Production and impact of researcher
34*	Citation/publication	**%HCP**	Publications cited above the 80 percentile in research area	Indicates papers in top 20% of research area
35*	Citation/publication	**b**	Author citation rate, minus sc, to the power of three quarters multiplied by h	The effect of self-citations on the h index
36*	Citation/publication	**hα**	Granular comparison of scientists with same h	Cumulative achievement
37*	Citation/publication	**hT**	Complete production of researcher evaluated using Ferrers graph	Production and impact
38*	Citation/publication	**hrat**	Rational h indicates the citations needed to increment h one unit	Provides greater distinction in scholar rankings
39*	Citation/publication	**grat**	Rational g; indicates distance to higher g index	Provides greater distinction in scholar rankings
40*	Citation/publication	**gα**	G using fractional papers and citations	Adds a quality aspect of cumulative achievement
41*	Citation/publication	**g**	g index	Cumulative achievement, inc. highly cited papers to distinguish between and rank scientists
42*	Citation/publication	**m**	m index	Median citations to publications included in h to reduce impact of highly cited papers
43*	Citation/publication	**e**	e index	Production and effect of highly cited papers, supplements h
44*	Citation/publication	**wu**	wu index	Impact of researcher's most excellent papers
45*	Citation/publication	**hg**	Hg index	Enables comparison of scholars with similar h and g indices
46*	Citation/publication	**H^2**	Kosmulski index, cube root of C.	Cumulative achievement. Weights most productive papers, improvement of h
47*	Citation/publication	**A**	A index	Magnitude of scholars citations to publications, supplement to h.
48*	Citation/publication	**R**	R index	Magnitude of scholars citations to publications, improvement of A-index
49*	Citation/publication	**ħ**	Miller's h	Overall structure of citations to papers, to enable comparison across field and seniority
50*	Citation/publication	**Q^2**	Quantitative & Quality index	Relates the number and impact of papers in h to the sources the scholar publishes in

51*	Citation/publication/author	**hA**	Alternative h:h/mean number of authors in the h publications	Indicates number of papers a researcher would have written if he/she had worked alone.
52*	Citation/publication/author	**Pure h**	Square root of h divided by authors and credit to their relative rank on the by line of h core articles	Corrects h for number of co-authors
53*	Citation/publication/author	**Adapted Pure h**	H computed using fractionally counted papers and citations counted as square root of number of authors	Finer granularity of individual h-scores normalized for number of co-authors
54	Citation/publication/author	**hi**	individual h	H index divided median number of authors on papers included in h
55*	Citation/publication/author	**hm**	Uses fractional paper counts to compute h	Softens influence of authors in multi-authored papers
56*	Citation/publication/author	**POP h**	Harzing's publish or perish h index	H index normalized for co-authorship effects
57	Citation/publication/author	**FracCPP**	Fractional citations per paper	Average effect of each paper, normalized for multi-authorship
58*	Citation/publication/time	**AWCR**	age weighted citation rate	Number of citations to all publications adjusted for age of each paper
59*	Citation/publication/time	**A(t)**	Difference between c(t) and c(t+1)	Aging rate of citation
60*	Citation/publication/time	**Price**	Currency of work	Percentage citations to papers not older than 5 years at the time of publication of the citing sources
61*	Citation/publication/time	**DCI**	Currency of work	Devalues old citations in parameterizable way using logarithm of the impact in past time intervals
62*	Citation/publication/time	**Hpd**	Papers that have at least hpd citations per decade	Compare output of scholars of different ages (seniority dependent h type index)
63*	Citation/publication/time	**Hc**	Parametrized weighting measured on 4 year citation cycle	Currency of articles in h core to account for active versus inactive researchers
64*	Citation/publication/time	**Ht**	Citations in h given an exponentially decaying weight	Age of article and age of citations to enable field normalization
65*	Citation/publication/time	**Dynamic h**	Indicates the size and contents of h core, number of citations and h-velocity	Detects where 2 scientists have same h and same number of citations but one has change in h and one does not.
66*	Citation/publication/time	**Index age & prod**	Mean n documents by age and CPP (3yr citation window) in 4 yr age brackets	Effects of academic age on productivity and impact
67*	Citation/publication/time	**Class Dur**	Percentile distribution of citations per year, normalized for document type and field	Durability of scientific literature
68*	Citation/publication/time	**H seq & matrices**	H sequences and matrices	Identifies variations in single scientists citation patterns, making scientists comparable
69*	Citation/publication/author/time	**AW**	Age weighted h	Cumulative impact of scholar, normalized for scholar's academic age
70*	Citation/publication/author/time	**AWCRpa**	Per-author AWCR	Number of citations to all publications adjusted for age of each paper and number of authors
71*	Citation/publication /time	**m quotient**	m-quotient	H normalized for academic age
72*	Citation/publication/time	**mg**	Mg-quotient	G normalized for academic age
73*	Citation/publication/time	**AR**	AR-index	Citation intensity and age of articles in the h core, supplement to h.
74*	Citation/publication/field	**n**	Comparison within field/specialty	H index divided by highest h index of journals in his/her field
75*	Citation/publication/field	**hf**	Comparison within field/specialty	Corrects individual citation rates for field variation

76*	Citation/publication/field	**x**	Comparison across fields	Quantity and quality normalized for field (5 yr impact factor)
77*	Citation/publication/database	**hmx**	H in context of citation index	Rank academics by their maximum h measured across WoS, Scopus and Google Scholar
Journal/article-field benchmarks, calculated by CWTS				
78*	Journal/publication/citation	**mcs**	Mean citation score	Prestige of the journal the scholar publishes in
79*	Journal/publication/citation	**mncs**	Mean normalized citation score.	mcs normalized for field, article type and publication year
80*	publication/citation	**pp top n cites**	Pp top number citations	Productivity and impact of the scholar
81*	publication/citation/WoS category	**pp top prop**	Pp top proportion	Identify scholars papers rated at top of their field
82*	Publication/citation/WoS index	**pp uncited**	Pp uncited	Percentage of scholars papers indexed in WoS that are not cited
83*	Journal/citations	**mjs mcs**	Mean journal score:mean citation sccore	Benchmark of prestige, based on expected citations of articles in journals the scholar has published in.
84	Journal/citations	**max mjs mcs**	Maximum mjs mcs	Benchmark of prestige, based on expected citations of articles in journals the scholar has published in.
85	Journal/citations	**mnjs**	Mean normalized journal score	Prestige of journal scholar has published in, normalized for disciplinary difference
86	Publication/WoS index	**pp collaboration**	Percentage collaboration	Percentage inter-institutional collaboration
87	Publication/authior/WoS index	**pp int collab**	Pp internal collaboration	Cognitive orientation

Chapter 6: Reflexive Analysis

Concept operationalization motivates the third research question into the extent concepts being measured by the indicators are defined by indicator developers. The review in Papers 1 and analysis in Paper 2 broadly describe the conceptual and mathematical character of ALI (RQ1). Details of how to calculate the indicators is presented in Appendix B, http://tinyurl.com/nj4mvca. In reflection the empirical analyses in Papers 3, 5, 6 and 7 focused on the arithmetic functions of ALI to capture research performance, directed at answering RQ2 and not the conceptual functions, addressed in RQ3. Chapter 2 presented a discussion of how the three major concepts author, publication and citation are defined in the literature, supported by examples from the 51 ALI used in the empirical analyses. Chapter 3 meanwhile addressed how citations are rationalized theoretically in ALI as proxies for aspects of the effect(s) of research performance. Therefore to complete this PhD work and fully address the extent the concepts being measured are defined in ALI (RQ3) and inform the appropriateness of the indicators, the properties a well-constructed indicator should contain to be valid are evaluated. These are important considerations somewhat overlooked in the 7 papers. Consequently, this chapter presents an empirical study in three parts followed by a presentation of the set of recommended ALI, which is the objective of this PhD work. The conclusions and ultimately implications for user and developers of ALI are presented in Chapter 7.

In the first part of this chapter, Section 6.1, the logic of operationalizing author, publication and citation in 51 ALI is assessed. Second, in Section 6.2, the theoretical and methodological orientations of indicator developers are mapped to explore subject orientations of the developers and third, Section 6.3, Gingras' validation criteria are used to assess the appropriateness of the 51 ALI and 18 publication and citation indicators (Gingras, 2014). The results of the analyses in Sections 6.1 6.2 and 6.3 are supported by Appendix B. Together with results from the 7 papers, this will provide a common understanding for recommending appropriate ALI, Section 6.5.

6.1 The logic of author, publication and citations in 51 ALI

In Chapter 2 the operational definitions of author, publication and citation were explored, however in this section the logic behind the composition of the operationalized variables is analysed so that we can learn more about ALI model performance. Figure 4 presents a logic

grid of the 51 ALI, which draws on the idea of a project planning matrix. The grid is based on Table 6 and Paper 2, where each paper presenting the indicator was read to identify how the developer(s) of the indicator conceptualized, defined, and operationalized author, publication and citation and combined them through different arithmetic functions to measure a specific aspect of research performance. The logic grid is read horizontally and vertically. The horizontal logic, in grey, shows the reasoning which connects the objective of the indicator, which is to link the performance of the person the indicator is designed to evaluate (the author), with the variables used to compute the performance (publication and the citation) and the goal (the aspect of performance the indicator is designed to measure). The definition and operationalization of the variables should lead to achieving the purpose of the indicator. Each of the links between the person, publications, citations and goal may be connected by a hypothesis. For example, the developers of the *%HCP* index believe that taking authors from specific academic seniorities (actively productive tenured scientists, research scientists, research professors), using papers indexed in WoS and conceptualizing citations knowledge transfer and thus impact, will support an indicator that measures excellence. The vertical logic is similar to the horizontal, but in this case we question whether the links between the objective of the indicator (the aspect of performance it measures) and the type of author, production and citation the indicator uses in its model are affected by assumptions that are outside the control of the indicator but that must remain favourable for the indicator to achieve its objective. The implication is we must consider what might cause the indicator to fail to meet its objective and what we could do to reduce that risk of failure. For example, if the *%HCP* is to be successful in producing measures of the excellence of authors from specific academic seniorities with papers in WoS, we can ask ourselves for example what could prevent or limit citations reflecting the transfer of knowledge in publications in WoS? Or, which consequences could there be if we changed the author profile?

The grid describes the logical structure of each ALI, and is useful to check the plausibility of the design of the indicator and helps us to decide how the indicator works. For the sake of simplicity, the grid is not exhaustive in that it does not explain all the nuances of the concept definitions, e.g it does not distinguish between researchers with publications in WoS, as in *alternative h* and researchers with at least 15 publications in WoS, *x-index*. Both these indices are grouped as author type: researcher with publications, publication type: papers in WoS. Appendix B contains detailed information on each indicator.

Figure 4. Logic grid of 51 ALI

		Publication					Citation										Measure										
		Paper	Papers in WoS	Papers in other index	Object with citation	Expression	Effect	Hirsch definition	Influence	Impact	Performance	Popularity	Quality	Reward	Transfer of ideas	Use over time	No definition	Career	Comparison	Currency	Distribution	Durability	Effect	Excellence	Growth	Independence	Quality	Quality & Quantity	Rank
Author	Author	Adapted pureH	fc; h;Q2;b;hT;Q2	Hc;hn;ht									h			hT	fc;hc;hn;ht; Q2;b; adapted pureh		hn	hc	h		Q2;hT	ht;b		Fc; adapted pureh			
	Award Winner		g;AR;mg;R; hα; gα	h				h;R			AR		hα; gα				g;mg		hα						AR;R		gα	h	g;mg
	Published	m; pureH	IQP;x;Alt h;	mquot; ct;at; POP h; AW;AWCR;AWCRpa			Alt h			m; POP h; AW;AWCR; AWCRpa		mquot	IQP;x			ct;at	pureh			ct;at			mquot;AW AWCR	IQP;m		Alt h; POP h; AWCRpa pureh		x	mg
	Published & cited		f;t;	hmx	DCI	Price				DCI						Price	f;t;hmx			Price			f;t				DCI		hmx
	Scientist	e;hg;A;	h2;hpd;wu;hm; index seq & mat;	n;π; dynamic h				hg; A; dynamic h	π		n					index seq & mat	e; hpd;h2; wu;hm		n; index seq & mat; π		A			e; h2; wu	hpd	hm			hg; dynamic h
	Seniority		%HCP; Index age & prod											Index age & Prod	% HCP			Index age & prod						%HCP					
	No definition		hw;Class Dur;hrat;grat;hf					hw					grat		Class Dur		hrat;hf		hf			Class Dur		grat			hw		hrat
Measure	Career		Index Age & prod											Index age & prod															
	Comparison		Hα; index seq & mat; hf	hn;n, π					π		n		hα			index seq & mat	hn;hf												
	Currency			ct;at;hc		Price										ct;at;Price	hc												
	Distribution	A	h					A					h																
	Durability		Class Dur												Class Dur														
	Effect	m	Q2;f;t;hT	mquot; AW;AWCR; hT						m;AW;AWCR		mquot				hT	Q2; f;t												
	Excellence	e	%HCP;IQP;h2;b;grat;wu;	ht									IQP;grat		% HCP		ht;e;h2;b;wu												
	Growth		AR;hpd;R					R			AR						hpd												
	Independence	Pure h; Adapted pureH	fc;hm;Alt h;	POP h;AWCRpa			Alt h	fc		POPh; AWCRpa							fc;hm; pureh; Adapted pureh												
	Quality		hw; gα		DCI			hw		DCI			gα																
	Quality & Quantity		x	h				h					x																
	Rank	hg	g;mg;hrat	dynamic h;hmx				hg; dynamic h									g;mg;hrat;hmx												
Citation	Effect	m	Alt h																										
	Hirsch definition	hg;A	hw;R	h; dynamic h																									
	Influence			π																									
	Impact			POP h; AW;AWCR;AWCRpa	DCI																								
	Performance		AR	n																									
	Popularity			mquot																									
	Quality		IQP; h; hα; gα;grat; x																										
	Reward		Index Age & Prod																										
	Idea transfer		Class Dur; %HCP																										
	Use over time		hT; index seq & mat ;	ct;at		Price																							
	No definition	e; pure h; adapted pureH	g;fc;mg;;h2; hpd;Q2;b;hrat;f;t; wu;hm;hf;fc	hc;hn;ht ;hmx																									

6.1.1 Results

The aim of the process of examining the logical combinations of concepts of author, publication, citation and the objective of the ALI is to cause them to be questioned more closely and supplements the mathematical analysis of the indicators, Appendix B. These results contribute to recommending a set of ALI, as the examination may reveal that the assumptions and "hypotheses" are in fact logically sound or questionable. Referring to Figure 4 we can see that the author concept, in grey, is used to inform us for whom the indicator model is designed, meaning that using other types of author may affect the performance of the indicator for e.g. Price Award Winners (*h, m-quotient, g, AR, R, hα, gα)*; researchers of a specific academic seniority (*Index of age and Productivity, %HCP*; scientists only *(c(t), a(t), e, hg,h2, hpd, A, Wu, hm, n, π, dynamic h)* or general application for anyone who has published a paper *(Hc, Hn, Ht, IQP, m, ħ,* Q2, *b, alternative h, POP h, AW, AWCR, Pure h, DCI, hmx)* but we do not know for whom the, i.a. *hw, rational h, rational g* and *hf* indices are appropriate because the objective of these indicators is to explore the sensitivity of the mathematical model to capture granular differences in ranked data sets and encourage future enquiry.

The concept of publication, in pink, is similar, but less varied, generally defined as papers in WoS or papers in other citation indices. This preferred definition illustrates how developers differentiate between a truthful representation of a researcher's publication output, which could include all the different types of output on his or her publication list, and the practical usefulness of using papers in a citation index to compute ALI. Papers in citation indices have an implicit aspect of quality that can be utilized in indicators of researcher performance to say something about "quality" and make the output of similar researchers comparable. The papers have passed peer-review and are published in core disciplinary journals that must have a certain level of citation to be included in the index and the papers are of course represented by a biographical record that makes the paper searchable and verifiable. Importantly for ALI, the number of works citing a paper is registered and details of these are indexed as well, enabling quantitative studies of scientific communication, more details in (Moed, 2005, p.35-50). Only *Price* and *DCI* suggest calculation using other forms of output, as long as the output has been published and cited.

In combination, the author and publication concepts clearly show that the ALI in Figure 4 and Table 6 are designed for a specific type of author with a specific profile of publications.

Therefore application of these ALI using types of authors and papers not identified in the indicator model, cannot guarantee results similar to those demonstrated in the ALI by the developer(s).

The horizontal rows of Figure 4 present the hypotheses about how the link between the author and publication with the citation and lead to the objective the ALI is designed to measure. In total eleven concepts of citation are operationalized in the 51 indicators, 20 ALI choosing not to define citations. When citations are not defined there is a gap in the logic of the hypothesis, meaning in extension that the decisions and actions based on mathematically well-defined ALI may turn out to be less effective than expected, e.g. the *Q2* index (Figure 4, author row, second cell from the left). The *Q2* index is designed for any type of author with papers in WoS. Following the row along to the definition of citations, we find that *Q2* is in the citation cell labelled "no definition" therefore what citations measure is not defined in the model, but continue across the row and we see that citations when aggregated with publications, are hypothesized to result in a measure of *effect*. Mathematically, the *Q2* index is the square root of "the geometric mean of *h* multiplied by median number of citations to papers in *h* index" (Cabrerizo et al, 2012) and Appendix B[21]. The mathematical functions attempt a robust average based model that combines the *h*-index' measurement of productive papers with the *m*-index (the median number of citations to the papers included in the calculation of *h*) to correct the effect of very highly cited papers and accordingly the skewed distribution of citations to papers. Yet following the logic of the vertical column, where we identify the assumptions external to the indicator that need to be favourable for the indicator to reach its objective, we can see the logic is broken. Reading down from *Q2* in the cell under "Papers in WoS", the aim is to measure the "effect" of these papers. Because as the citation is undefined, we do not know how to identify the citations necessary to prove if the objective of the indicator has been reached and how this measure of "effect" is not a measure of for example "distribution". According to this logic, the *e, pure h, adapted h, g, fc, mg, hc, hn, ht, h2, hpd, b, hrat, f, t, wu, hm, hf, hmx* also present flawed hypotheses.

Only the developers of the *Index of Age and Productivity*, *Classification of Durability* and *% HCP* indices fully define citations as a concept within the constraints of the indicator (Costas et al., 2011; Costas et al., 2010a; Costas et al., 2010c). Costas et al accommodate the citation as a variable that is different depending on what aspect of performance the indicator is designed to capture. In the *Classification of Durability* and *%HCP* citations represent the

[21] E-material: http://tinyurl.com/nj4mvca

impact and transfer of knowledge of documents beyond their original producers; in *Index of Age and Productivity* citations are defined as one form of reward for academic dividend that can be counted, Appendix B.

Six ALI accept Hirsch's definition of citation as "a broad definition of overall scientific impact", (Hirsch, 2005), though they do not necessarily define what "broad impact" means, Appendix B, and impact is, as Martin & Irvine (1983) discuss, very complicated. The *h, R, hg, A, dynamic h,* and *hw* indicators refer to Hirsch and utilize the definition "broad impact" with different combinations of author and paper to achieve different objectives: to measure distribution of a scientist's papers to citations as in the *A* index; the growth citations to papers in WoS authored by Price award winners, *R*; the number of citations an author's papers in WoS would have received if the author had worked independently of his/her co-authors, *fc*; the square root of a set of papers in WoS, weighted for highly cited papers, as an indication of quality, *hw;* and the rank position of a scientist using papers in citation indices as the ranking factor, *hg* and *dynamic h.* Meanwhile, the *ht, ct, at, Price* and *h sequences and matrices* define citation as an indication of "use over time" to respectively measure effect, distribution of citations to papers and to compare similar authors. Otherwise in the remaining 17 ALI the concept of a citation is defined very differently.

The developers of the h^c, *hn* and h^t indicators choose not to define citations at all and include a waiver in their indicator proposal, Appendix B, stating:

> "conducting theoretical analysis of the properties of the proposed indexes is the next step in this work, but it is beyond the scope of this paper" (Sidiropoulos et al., 2007).

This approach suggests they designed the indicator and proposed its mathematical logic before investigating the logic of the hypotheses and the variables they were working with. As the logic grid shows h^c, *hn* and h^t are intended for authors with papers indexed and cited in WoS and/or other citation indices, and these indicators hypothesize that when aggregated citations (undefined) to these papers can be used to 1) *compare* researchers when the papers in the *h* core are normalized for age (*hn*), 2) evaluate the *currency* by normalizing the citation to these papers for age (*hc*) and 3) the *effect* of the author's papers by normalizing for the number of papers in the *h* index (*ht*). These models investigate structural obstacles of the *h* index. Using the vertical axis of the logic grid the flaw in the hypotheses of these ALI is exposed, and as the developers clearly write, "the *meaning* of the numbers produced by these

indicators have yet to be investigated". Therefore the indicators are still at the experimental stage and not intended for implementation in evaluation. Exploring the mathematical model rather than the evaluative dynamic of an indicator appears to be a common tactic in indicator development. As a researcher's citations and publications increase and decrease over time they are by no means static. Fixed indicators do not capture this evolution, therefore testing the stability and robustness of mathematically complex indicators such as the *rational h, rational g, hT, hα* and *gα* can be the objective of indicators rather than in an evaluative perspective defining how publications and citations together represent measures of performance at the individual level. The *hw* and *dynamic h*-indicators exemplify Lotkanian informetrics in that they practice as what Kuhn would refer to as "pre normal" science, i.e. their contributions are firmly based on past achievements particular to bibliometrics, that supply a foundation for further practice and more open ended questions for further research (Kuhn, 1970, p.15). This is a theoretical approach to bibliometric indicators, building on the Lotka power law that uses mechanisms of size and frequency. Other power-models are the Zipf-type laws, used in rank approaches and time-type distributions constructed using Price and Brookes type laws of growth and ageing that can be used individually or combined in developing especially ranking indicators, (Ye, 2011). In the pre normal science perspective the *hw* and *dynamic h* indicators attempt to understand how the distribution between citations and publications can provide different views and raise new phenomena in the interpretation of physical referencing behavior and bibliometric distributions, and hence question how they should be studied (Kuhn, 1970) rather than how to evaluate researchers. They are thus explorative indicators that raise questions and are not designed to produce absolute answers. Hence the focus is not on conceptualizing citations, but studying equations as theoretical interpretations of bibliometric laws and stimulating further studies (Ye, 2011). In the example of the *hw* and *Dynmanic h* indicators, the Lotkanian approach uses the concepts of size and rank frequency to provide steady patterns of the evolution of papers in the h-core over time, i.e. quality papers, indicating that no publication can instantly become a highly cited one, thus implying that in dynamic indicators the *h*-inconsistency cannot occur, (Ye, 2012), discussed in Section 2.2.3. Further, the Loktanian approach is used to develop indicators that use citations as a constant rank-frequency function in the mathematical model to increase the granular precision of scholar rankings, see amongst others the proposal of the *tapered h-index* by

(Anderson et al., 2008).

Plotting the ALI in the logic grid identified how authors, publications and citations are linked to the objective of the ALI, and is a simple way to question if the indicator makes sense and which external assumptions may affect the objective. It became clear that the ALI are designed for 1) specific researcher profiles, 2) they can be logically flawed, 3) only some ALI are actually proposed for the evaluation of researchers, while 4) other ALI are proposed as experimental models of the relationships between publications and citations with the objective to encourage further bibliometric research and at this stage, unconcerned with researcher evaluation.

6.2 Theoretical and methodological orientation of indicator developers

In this section I explore the extent ALI research is interdisciplinary. This section explores if the design of ALI in the data set involve collaboration with people with different theoretical and methodological backgrounds. That is, if the indicators are developed with a broad or narrow perspective on the sociology and communication of science and particularly if developers with different disciplinary orientations design indicators to fit a specific user profile or indicators with a specific objective. This knowledge will help identify disciplinary appropriate indicators.

6.2.1 Results

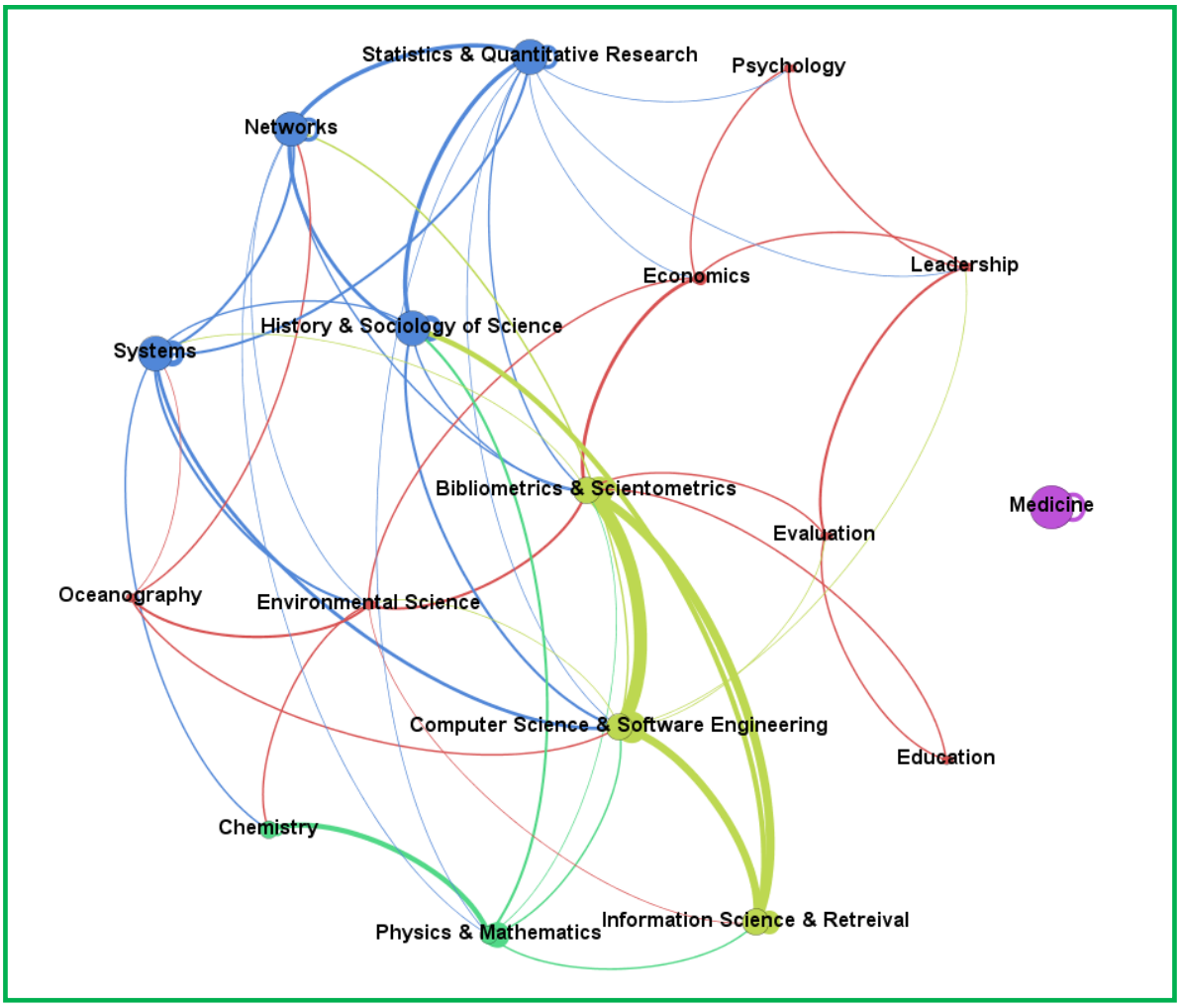
A network map was created by collecting each developer's field of expertise, as described in their own words on their online profile, either in Google Citations, Microsoft Academic Search or on their online university profile, Figure 5. One hundred and twenty-five research specialties were identified, and for simplification collapsed into 17 broad fields of study, represented by the circular "nodes" on the map. The lines between the nodes, i.e. the "edges", illustrate the extent different field specific expertise are used together in the development of indicators: the thicker the edge, the stronger the collaboration. The map was visualized using Gephi[22], ranking parameter modularity class, Fruchterman Reingold layout algorithm and colour of the lines (the edges) linking the fields is the target.

In the centre of the map, Bibliometrics & Scientometrics are identifiable by a very strong interdisciplinary cooperation with Computer Science & Software Engineering and Information Science & Retrieval, but have a weak connection with specialties outside of this

[22] The map was created with Gephi, a free network visualization tool available at: http://gephi.github.io/

triangle of collaboration. Only developers from Information Science & Retrieval explicitly incorporate expertise from the History & Sociology of Science in indicator construction while Biblio- and Scientometrians choose to draw external knowledge from Network specialists and to some extent experts in Systems and in Evaluation and Leadership. Physics & Mathematics collaborate strongly with Chemistry but also include specialist knowledge from the fields of Information Science & Retrieval, Computer & Software Engineering, the History & Sociology of Science and Bibliometrics & Scientometrics. The greatest multi-disciplinary collaboration is between the members of the blue cluster, who further include expertise in Chemistry, Psychology, Leadership, Economics, the History & Sociology of Science, Computer Science & Software Engineering,

Figure 5. Map of inter-disciplinary collaboration between indicator developers



Physics & Mathematics as well as Bibliometrics &Scientometrics in the development of indicators. The red cluster, primarily Economics, Education, Psychology and Leadership has a strong science policy character to it. The ALI from Environmental Science and Oceanography are developed in collaboration with Network and Computer Science specialists, and this appears to be a strategic collaboration that has been established for the purpose of designing the indicator. As in the papers presenting the ALI it is clear the developers are scientists acutely aware that they are being quantitatively evaluated and wish to improve the current state of evaluation.

Table 10, p.124, presents Cluster membership, the objective of the indicator in evaluation, the profile of the researcher the indicator is designed to measure and the amount of indicators developed in collaboration. In indicator construction it appears developers from different research specialties are collaborating to produce indicators within and across disciplinary boundaries. Five distinct collaboration groups were identified. The Green Cluster collaborated in the development of 38 out of 51 indicators, the Blue 23 out of 51 indicators, the Red 15, Dark Green 19 and Purple 1 indicator, Table 10. In the red, blue, dark green and purple cluster the identity of the researcher the indicator is designed for is very concrete. They are scientists or authors with published papers indexed in citation indices. In the light green cluster, Scientometrics and Bibliometrics, Information Retrieval and Computer Science, the profile can be somewhat abstract, using concepts such as “expression” and “object” but primarily this cluster illustrates a non-uniformity in approaches to defining the profile of the subject under evaluation which resulting in many profiles or scenarios being explored. What is interesting is the relation between the objective of the indicator and the disciplinary orientation of the developers, Table 10, second column. The indicators developed by the Red Cluster (Economics, Leadership, Evaluation, Psychology) are classic ALI of performance that enable ranking of researchers across similar fields and indication of the effect of all, selected or excellent research papers. Meanwhile the Blue Cluster (Sociology, Networks, Statistics) develop indicators that enable granular comparison between researchers within speciality, to peers, across fields and across domains. The Green Cluster (Scientometrics, Bibliometrics, Computer Science, Information Science) develop indicators of all round researcher performance, the majority of indicators enable comparison between researchers, researcher rankings, average impact normalizing for distribution of citations, the currency of the researcher and the indicators that normalize for collaborative works to indicate independence. Meanwhile, the Dark Green Cluster, (Chemistry, Physics,

Mathematics) focuses on identifying pioneer research and the actuality of a researchers work, while the Purple Cluster (Medicine) design one indicator to compare scientists across medical fields. Across all clusters, the common interest is to identify excellent research and the currency of the work under evaluation.

6.3 Properties a well-constructed indicator should possess in order to be valid

In the previous sections the logic of the indicators was explored and the collaboration and disciplinary orientations of developers of ALI mapped. The first study highlighted indicator developers' ambiguity towards author and citation, leaving the logical conclusion that an indicator that cannot isolate the variables of measurement, itself has a certain level defect. Yet this does not mean that the indicator should be dismissed as irrelevant or misleading. But before I can recommend appropriate indicators the validity of the ALI has to be assessed. This study is relevant because a given indicator could still measure what it is supposed to measure even though the definitions of the components of the indicator are poorly articulated. This is because the indicator is not the concept itself but a proxy, "used as a way of measuring how the reality behind the concept changes over time or place," (Gingras, 2014). Gingras (2014) suggested that developers, administrators and researchers need to learn how to evaluate indicators before using them since indicator values effect both decisions, careers and self-worth (Paper 4). To complete the study of ALI, this final analysis evaluates the validity of the ALI. Each indicator was assessed using Gingras' three evaluation criteria. The three criteria that define the essential properties a well-constructed indicator should possess in order to be considered valid are these:

1) The adequacy of the indicator for the object/concept it measures

The properties of the indicator should be checked against the properties the concept is assumed to have. The indicator should be strongly correlated with what we presume to be the inherent characteristics of the concept we want to measure using that specific indicator.

1a. Does the indicator correspond to the object or concept being evaluated?

1b. Are the results produced by the indicator of the correct order of magnitude, given what we know about the object/concept?

2) Sensitivity to the intrinsic inertia of the object/concept measured

Inertia is resistance to change, thus a good indicator varies in a manner consistent with the inertia of the object being measured since different objects change with more or less rapidity or difficulty, (e.g. a university cannot dramatically raise or lower the rating within 2-3 years).

The indicator value increases in a manner consistent with an increase in the concept that the indicator measures

2a. Does the timescale of the indicator make sense for the rate of change in the object/concept?

3) The homogeneity of the dimensions of the indicator

An indicator should be homogeneous in its composition. Homogeneous indicators of research output can be constructed using for example the number of articles published in scientific journals. However, summing indicators that have different measures makes an indicator heterogeneous, for example combining publication numbers with a citation measure produces a composite heterogeneous indicator. The fundamental problem with composite heterogeneous indicators is that when they vary, it is impossible to have a clear idea of what that change really means, since it could be due to different factors related to each of the composite heterogeneous parts of the indicator.

3a) Is the indicator homogeneous?

6.3.1 Results of the evaluation study

Table 11 summarizes the evaluation of the validity of indicators of publication count, Table 12 summarizes the evaluation of indicators of citation count and Table 13 summarizes the evaluation of the hybrid ALI. Gingras' criteria three criteria for validity were used to assess the properties of each indicator. The first column of each table is the abbreviation of the indicator, the second the concept the indicator approximates as defined in papers presenting the indicator; Gingras's criteria are presented in the next three columns and the rationale behind the evaluation is presented in the final column. The green rows are the indicators that fulfil all three of the criteria.

Gingras claims that his three criteria are sufficient for detecting any valid indicator. However, even the indicators that fulfil all three criteria do not necessarily produce reliable results and have caveats limiting their application. For indicators that count the number of publications and citations, Tables 11 and 12, such caveats could be the scope and precision of the data collection process, the variation of publication and citation rates between fields and what constitutes an internal or external citation further affects the validity and reliability of the indicator value. Five of the seven citation indicators fulfilled all three criteria, and four of the ten publication indicators fulfilled all three criteria. For these simple counting indicators Gingras' criteria made us question if there were disconnections between the concepts the

indicator is designed to measure. However for the ALI (hybrid) indicators, the criteria were more difficult to apply and evaluation could not be done without also analysing the mathematical model of the ALI, Table 13 and Appendix B. Only 11 out of 51 ALI indicators had the properties to fulfil all three criteria, 4 of the 10 publication indicators and 5 out of the 8 generic citation indicators.

The criteria of "inertia" and "homogeneity" were particularly difficult to apply, not at all as simple as Gingras claims. To confidently evaluate these two properties, the equation used in the indicator composition to argue the mathematical relationship between author, publications, citations, and the other variables operationalized in the indicator were deconstructed and literature reporting tests of indicator performance consulted. Thirty-four out of 51 ALI failed to meet the property of homogeneity, while none of the 18 publication and citation indicators failed. In indicators "homogeneity" means that the only entities present in the equation are the unknown *function* and its *derivatives* (possibly with some coefficients), so $y''=xy$ is homogeneous, where $''$ indicates the derivatives. $y''=xy+x+1$ is not homogeneous, since x+1 doesn't "involve" y or its derivatives. The *function* is a mathematical relationship in which the values of a single dependent variable are determined by the values of one or more independent variables. The *derivative* tells us the slope of a function at any point on a curve and the rate of change. Taking the *h*-index as an example, *h* aims to counteract emphasis on sheer volume of papers by combining with the citation count of productive papers, however *h* correlates with the number of published papers and is determined by the quantity of publications rather than the notion of quality or impact, which means the value of the *h* does not necessarily go up when the notion of quality or scientific impact goes up. Thus *h* fails to meet all three properties of concept, inertia and homogeneity. The size dependence of *h* on publications is explored in Paper 7 were the ratio *publications* to *h* (total publications divided by *h* value) was shown to determine the position of the researcher in rankings. It was observed that if a researcher had a ratio *P:h* of ≥ 3 they would fall in rank position, and if the ratio is <3 they gain in rank position. This size dependency was found in changes in rank position using the *A, m* and similarly the *m-quotient* with a dependency on number of years, causing vast jumps in rank position. This finding about the size dependence of *h*, in which larger numbers of publications generally command higher *h*-indices has previously been discussed in i.a. (Costas et al., 2010a; Vinkler, 2007; van Raan, 2006). The problems with *h* cannot be corrected by inventing even less intuitive and more complicated ideas like the *ht*, *hα* or *adapted pure h*. This just makes the indicators even less

transparent and when they vary, it is impossible to have a clear idea of what that change really means, since as Gingras writes "it could be due to different factors related to each of the composite heterogeneous parts of the indicator" (Gingras, 2014). Therefore in the present study the *h*-index is evaluated as invalid, a decision based on Gingras' criteria, published studies in the literature and my own investigations presented in this PhD work. On the basis of Gingras' criteria the *CPP*, *AWCR, f* or *t* indicators are evaluated as appropriate indicators of average number of citations per paper as they retain their intuitive relation to the concept of "average" they seek to measure. However, according to statistical theory such average-based indicators are problematic, as a single statistic of centrality may not adequately summarize the asymmetries of skewed citation distribution. In practice this means that two researchers could have the same average indicator values, but the upper and lower parts of the distribution of citations to publications could be very different. Therefore to contribute to an appropriate evaluation of the researcher, these average-based indicators should be supplemented with other indicators that capture what takes place at the tail ends of the citation distribution (Albarrán, 2011; Moed, 2005).

6.5 The road to recommending indicators

In bibliometrics we rely less on anecdotes and more in the favour of data, the idea being that numbers tend to lie less badly than people do. But the analyses in the previous sections and the papers included in this PhD work show that identifying the indicators that do not exactly tell the truth but are "appropriate" is complicated. This section summarizes the methodological framework and major findings that support the research questions and objective of the PhD, i.e. to recommend appropriate ALI.

The first step in recommending appropriate indicators was to identify ALI that use simple, and transparent models with realistic demands to data collection. In Papers 1 and Paper 2 the methodological characteristics of ALI were rated on a 5 point scale that graded the complexity of the data-collection and the complexity of the computation of indicator values, Figure 2. Published documentation of indicator development and application, as well as informal documentation such as discussion of indicators in scholarly blogs by non-bibliometricians were collected and used to support argumentation for or against the scoring and application of specific indicators (Paper 1). The aim, definition and mathematical equation of each indicator were systematically noted along with criticisms, advantages and disadvantages (Paper 2) and Appendix B. Only indicators rated simple enough for end-user

application, scores ≤ 3, were further analyzed throughout the PhD work and considered in recommending a set of appropriate indicators. The seniority and disciplinary appropriateness of these ALI were explored in Papers 3, 5, 6 and 7. In Paper 3 the potentials of ALI to increase the value of, i.e. enrich, the publication information on a researcher's CV was explored, questioning if certain indicators were seniority and disciplinary appropriate and if the indicator score had a negative or positive effect on the profile of the researcher. No seniority trend between the amount of years active as a researcher, number of papers and number of citation was found, making benchmarking at this level inappropriate. The simple calculation of *number of citations per publication per years-since-first-publication* (*CPAY)* proved a more appropriate and transparent computation for setting expected disciplinary performance benchmarks for comparing researchers to their peers than the tested *ALI*. The results show *CPAY* scores below a disciplinary-specific critical value will determine a low ALI score and rank researchers in the bottom 25% of their discipline, meaning that ALI are not appropriate for these researchers and will not enrich positively the information on their CV. Using *CPAY* to determine benchmarks for middle and high scoring researchers proved unstable, but identified clear disciplinary differences in indicator scores and a suspected predictive relationship between ALI that was further investigated in Paper 5. Before summarizing the findings of Paper 5, Paper 4 will be addressed. Drawing on literature from evaluation studies, Paper 4 explored aspects of responsible metrics in evaluation that should be considered in compiling a set of recommended ALI. The aim was to learn more about potential positive and negative effects a bibliometric evaluation may have on a researcher, a theme continued from Paper 3 where low or high indicator scores were discovered to be disjunct and sometimes contradictory from the researcher profile documented on the CVs of the researchers in the dataset. Evaluations based on ALI were thus hypothesized to lead to assumptions about the productivity and citation impact of a researcher, these assumptions may be unsubstantiated, and affect the psychological character of the individual. Considerations in recommending ALI can be understood in terms of:

1) *Transparency*: computing ALI on the best possible data in terms of completeness and accuracy. In the computation of ALI we should be aware that end-users may utilize whatever information is available in the computation of ALI, perhaps self-regulating the information so that ALI scores will increase their subjective-validity and self-worth. It is important the results of the ALI can be verified. Knowing what data is and is not included and understanding how the arithmetic in the indicator model works can reduce misinterpretation that could cause fabricated self-images and damaged reputations.

2) *Demographics:* ALI have the potential to compensate for some of the gender and cultural differences and stereotypes that was evidenced in the literature on expert assessment. Evaluators' rate CVs and journal articles lower for women than men and ALI can supplement peer assessment by objectively documenting e.g. the citation impact, and use or currency of the said articles. However stereotypical character traits of the evaluand may still affect the ALI scores. For example being competitive or assertive is commonly attributed to male researchers and these traits are favoured in indicators that document winning awards or ALI that are designed for award winners. Females are on the other hand associated with communal qualities such as being nice or compassionate, qualities that do not lead to awards, meaning they are likely to perform poorly on award-based indicators. ALI users should be aware of potential demographic differences that may bias the indicator scores.

3) *Motive*: the ALI have to fit the motive of the evaluation and therefore the both the objectives of the evaluation and the ALI have to be clear. Thus the results of the bibliometric evaluation may promote self-improvement rather than self-protection. This point led to the development of the logic grid, Section 6.1.

4) *Diversity*: the ALI should account for disciplinary and specialty variation. Clear criteria should be used to account for variance in measures across disciplines and in specialty publishing practices and communication of research.

5) *Openness:* In a set of recommended indicators it is important to capture different aspects of researcher performance. The flaws and effects, advantages and disadvantages of ALI must be explicit: the database used to compute them and other external factors that can affect the evaluation should be reported. ALI evaluate only one aspect of researcher performance and ALI should be used in combination with each other to provide a bibliometric profile of the researcher. This final point led to the development of Appendix B.

The psychological effects of ALI can be addressed, though not solved, by promoting knowledge and understanding of the challenges and limitations associated with them. Yet to recommend a set of ALI and promote indicators with valid methodologies, the mathematical qualities of the indicators in capturing research performance had to be tested and redundancy and dependency of indicators on one another explored. Investigating which factors determine a high or low indicator score and determine position in researcher rankings motivated Paper 5. By computing the ALI and publication and citation indicators for each researcher in the dataset, ranking the scores and mapping the change in rank position, it was possible to

identify central and isolated indicators. The central indicators had strong links to the other indicators in the set and a high score on these central indicators predicted high scores on the related indicators. Each discipline had its own central indicator: In Astronomy the *hg*-index made rankings with over 25 other hybrid indicators redundant; in Environmental Science the *h*-index, in Philosophy the *IQP*-index and in Public Health the *g*-index made between 22 and 28 other indicators redundant. Across all disciplines the following trend was observed: If a researcher was ranked in top 10% of the sample by the central indicator, the researcher is placed in the top 10% using the other indicators that the central indicator has strong links to. Likewise, for researchers in the top 25%, middle 50% and bottom 25%.
Indicators that count either publications *or* citations or combine *all* publications and citations were identified as "isolated" indicators and these indicators produce haphazard and uninformative rankings, e.g. *number of authors per paper*, *age of citations*, and *% not cited*. Isolated indicators are interesting because they measure different aspects of research performance than the central indicators. Combining the disciplinary-specific central indicator with isolated indicators was recommended in Paper 5 to provide well rounded bibliometric profile of a researcher.

The appropriateness of ALI as indicators to rank researchers in social comparisons was further explored in Paper 6, addressing 1) the extent rankings with author-level indicators produce similar ranks between researchers in GS and WoS, 2) if different counting methods affect our concept of the ''average'' scholar, and 3) the effect of discipline and seniority on scholar rankings and our concept of the average scholar. The researchers in the dataset were ranked using ALI computed in WoS and GS. The arithmetic, harmonic and geometric mean of the ALI were used to investigate disciplinary averages and our expectations to average performance. When compared to previous empirical studies using similar data from authority disciplinary-specific citation indices, the indicator averages produced in WoS and GS were very different from authority sources and produced distorted expectations to average researcher performance. The paper concluded that the type of mean matters in benchmarking the average performance of ranked researchers as it creates expectations related to performance, particularly with respect to our concept of a below or above average researcher. The implementation of inappropriate averages in researcher assessments can result in inaccurate assessments, disillusioned researchers and misinformed assessors. Therefore to be able to recommend one indicator that produces a stable indication of rank position across databases could be advantageous and was explored in Paper 6. The ALI that build on the

geometric mean proved superior in providing stable cross-database rankings, thus extending the work of (Panaretos & Malesios, 2014). The standard deviation of the differences between researcher rankings in WoS and GS show that the *hg* indicator produced rankings with less variance than the other indicators, although the agreement appeared to be highly influenced by the amount of missing data. Nevertheless, this paper makes no assumptions about what the indicator says about a researcher's excellence and recommends only the potential of *hg* as a cross-database visibility ranking parameter. In the paper I recommend end-users treat ALI as statistical analysis tools, in which one first needs to know the distribution of the data before applying the statistical model. In this way ALI that use e.g. the arithmetic average functions as in *CPP* will not be applied to data that does not follow a normal distribution but is in fact highly skewed and will result in inaccurate results. To avoid this, end-users of ALI need to determine the distribution of their data before applying ALI and supplement fixed-average indicators with models that indicate performance in the tail ends of the distribution.

In the final paper, Paper 7, the extent ALI capture the performance of bottom, middle, top and exceptional researchers was explored to learn more about social comparison and disciplinary specific indicators (RQ 2). The method combined a two-step cluster analysis, ordinal regression, odds analysis and correlation analysis to explore the validity of researcher performance measured statistically through indicator scores. These scores were compared to the researcher's "performance" as documented on the corresponding CV. In Astronomy the *h2* indicator, *sum pp top prop* in Environmental Science, *Q2* in Philosophy and *e*-index in Public Health grouped the researchers in four clearly demarcated clusters of low, middle, high and extremely high scoring researchers. No seniority specific clustering indicators were identified. Academic age, measured as the number of years since the first publication recorded in WoS, was statistically significant for grouping researchers with a substantial increase in the odds of researchers with a longer academic age being placed in higher clusters (15% increase with each unit increase in age). But this did not explain all of the variance in cluster placement. A further analysis confirmed the strong influence of publications and citations, which normalized for academic age, is suspected to steer the placement of researchers in performance groups of researchers who score low, middle, high and extremely high indicator scores, an effect also noticed in Papers 3 and 6. Meanwhile the within-cluster rank position of the researcher was determined by the ratio between the number of publications and the indicator value, previously reported on page 96, which means that

researchers could strategically include or exclude publications from the calculation of the indicators to improve rank position and artificially plump their statistics.

In order to complete investigations into the characteristics of ALI (RQ1) and recommend disciplinary appropriate indicators (RQ2) the extent the concepts being measured are defined in ALI were investigated in the PhD body (RQ3), Chapter 6 and Appendix B. The logic grid and analysis of the developers disciplinary orientation drew our attention to the fact that ALI are designed for specific researcher profiles, they can be logically flawed, and not designed for evaluation purposes but to further experimental explorations. Twenty-four of the 51 ALI presented logically sound theoretical models. Therefore it is vital for end-users to read the original indicator proposal to determine who the ALI is designed for, what type of publication and citation it measures and the objective of the ALI i.e. if it fits the motive and subject of the evaluation. Finally the evaluation of the properties an ALI should possess in order to be considered valid were assessed using Gingras' criteria of correspondence to concept, inertia and homogeneity. Only 5 out of 51 ALI indicators had the properties to fulfil all three criteria, were assessed as logically sound *and* scored 3 or less in complexity of data collection and calculation, Appendix B. The criteria of simplicity excluded valid and theoretically robust indicators such as the *Index of Age and Productivity*, *Classification of Durability,* π and *a(t)* from the final recommended set. Likewise *hg* that performed strongly as a ranking parameter in cross-database rankings, Paper 6, is also excluded the final recommended set. The logic test of *hg* shows that it fulfills its original intention as a ranking indicator of scientific papers and fulfills the criteria of a logical hypothesis. Gingras' validity criteria showed however that *hg* does not contain the properties to be considered valid, failing on the sensitivity of the inertia of the concept being measured (the value of *hg* does not go up when the notion of quality or scientific impact goes up and can disproportionate to average publication rate) and the homogeneity of the dimensions of the indicator (combining *h* and *g* does not improve discriminatory power and *hg* has no direct meaning in terms of papers and citations of a scientist and can lead to hasty judgements). Confirming the conclusion that *hg* has strong properties for ranking information in citation indices, which it is designed to do, but in an evaluative perspective is uninformative about the performance or recognition of a researcher's work because this is not its objective.

When the building blocks of indicators are deconstructed as they are in the chapters and in the papers included in this PhD-work, determining appropriate indicators is very challenging. One also begins to question the value put on the indicators in evaluation as there are so many

caveats to consider. But it is important to remember that the worth of an indicator could be limited to end-user using indicators in situations for which they are not designed or in flaws in the interpretation by the end-user rather than flaws in the construction of the indicator. The analyses in the previous sections and the findings in the Papers have resulted in a set of indicators that has been evaluated for: homogeneity, inertia, concept definition, logic, mathematical complexity, theoretical robustness, data availability, redundancy with other indicators, their stability in rankings, disciplinary appropriateness and database bias. It was not possible to identify seniority appropriate ALI and a large discrepancy between the ALI indications of researcher prestige and the profile on the researcher's CV was observed. On this background a set of recommended ALI are presented in the next section.

6.5.1. The set of recommended ALI

The recommended **publication** indicators, Table 14, are homogeneous, correspond to the concept they are defined to measure and fulfill the criteria to inertia, that is the indicator value increases in a manner consistent with an increase in the concept that the indicator measures. Data collection can be adapted to the type of publication deemed important for the evaluation or discipline without compromising the properties of the specific indicator and they can be adapted to fit the discipline. These indicators are simple to calculate, are logical and what they measure is unambiguously clear. They are isolated indicators with no information redundancy between them.

The recommended **citation** indicators, Table 15, are also homogeneous, correspond to the concept they are defined to measure and fulfill the criteria to inertia. The indicators are simple to calculate in practice and what they measure is unambiguously clear. They can be applied in different disciplines. They are isolated indicators that are not controlled through relationships with other indicators and each indicator informs on a different aspect of a researcher's citation count. Counting whole citations and publications independently does not however inform on the so called "citation impact" of a particular researcher which is traditionally the objective of ALI. Yet average-based indicators that attempt to indicate an overall citation impact are problematic (Albarrán, 2011; Moed, 2005) because a single statistic of centrality may not adequately summarize the asymmetries of skewed citation distribution, as some publications will have scored the average number of citations, some will have scored higher and some lower. Therefore before using the recommended ALI indicators, it is advisable to describe the central position of the frequency distribution and pattern of

citations to publications using mean indicators *CPP (arithmetic mean), f (harmonic mean), t (geometric mean), median* and *mode* supplemented with measures of spread to summarize how spread out the citations to publications are. To describe this spread, describing the range and quartiles will be informative.

The recommended **ALI (hybrid)** indicators, Table 16, are homogeneous, correspond to the concept they are defined to measure and fulfill the criteria to inertia. While the indicators are still rated as simple, data collection and calculation is more demanding than indicators of publication *or* citation count but this increase in computation complexity appears to correlate with an increased validity of the indicator. The indicators have been tested empirically in WoS and other databases and are suitable for implementation in different disciplines. There is no redundancy between indicators. Each ALI indicates different facets of the publication performance of a researcher and can be used together, within the context of the citation index, to indicate:

- the currency and relative currency of the researcher's work (*c(t), Price*),
- the cumulative effect of the researcher's body of work, (*AWCR)*
- the excellence of the researcher's work compared to *subjectively* defined specialty standards that are based on the researcher's publication habits (*IQP).*

6.5.2 Where are the ranking indicators?

The **disciplinary ranking** indicators identified in Paper 6 and 7 do not fulfill the criteria to inertia or homogeneity and further they contain caveats that seriously inhibit their operationalization as measures of the concepts of "excellence" or "effect". *h2* (Kosmulski, 2006) has a consistency problem (Waltman & van Eck, 2012), which means it cannot determine excellence as it does not discriminate between scientists having different number of publications with quite different citation rates for relatively high *h2* indices. *Sum pp top prop* is a variation of the *PPtop 10%* indicator designed by the CWTS group, Leiden University[23]. It determines the amount of publications a researcher has produced belonging to the top 10% of all WoS publications in the same field (i.e., the same WoS subject category) that have the same publication year and are of the same document type. A disadvantage of *sum pp top prop* is the artificial dichotomy it creates between publications that belong to the

[23] The explanation of design, advantages and limitations of CWTS indicators is credited to Dr. Clara Calero-Medina, a researcher at the Centre for Science and Technology Studies (CWTS) of Leiden University. The explanation of the the sum pp top prop indicator in the text is based on numerous email correspondences with her throughout the ACUMEN project.

top 10% and publications that do not belong to the top 10%. A publication whose number of citations is just below the top 10% threshold does not contribute to the indicator, while a publication with one or two additional citations does contribute to the indicator. Because of this arbitrariness the indicator is not designed to stand alone, but in context with other indicators that present a field's mean normalized citation score. An obvious caveat is the field definitions on which these indicators are based on. WoS subject categories are used to define the field, but unlike WoS subject categories, the fields in reality do not have well-defined boundaries; they overlap, consist of multiple specialties and have heterogeneous citation characteristics. *Sum pp top prop* does not correct for this within field heterogeneity and can misrepresent the individual researcher. The *Q2* index (Caberizoa et al., 2012), which combines the h-index' measurement of productive papers with the median number of citations to these papers, i.e. the m-index (Bornmann et al., 2008a) uses the geometric mean of these combined indicators to provide an estimate of the average effect of this set of papers. But the *Q2*-index results are closer to *h* than to *m* and this can be interpreted as a penalization of the *m*-index in the cases of a very low *h*-index. *Q2* also suffers from the same inconsistency problems as *h* (Rubem et al., 2015). Likewise the *e*-index (Zhang, 2009) is another indicator of excellence dependent on the calculation of *h.* The *e*-index is the (square root) of the surplus of citations in the h-set beyond h^2, i.e., beyond the theoretical minimum required to obtain an *h*-index of 'h'. Consequently *e* only makes sense when *h* is given and is inherently flawed. *hg* (Alonso et al., 2010) also adapts *h*, as it calculate the square root of the sum of *h* multiplied by the *g*-index (Egghe, 2006). Consequently, it includes in its calculation an arbitrary threshold of citations creating a fractional size-frequency function and it can be disproportionate to average publication rate (Alonso et al., 2009; Costas & Bordons, 2007b). This means the *hg* index of a scientist with one big hit paper and a mediocre core of papers could grow a lot in comparison with scientists with a higher average of citations.

There is no certainty if the recommended indicators will continue to be dominant on other datasets within the same disciplines. But the recommended indicators will contribute to convincing researchers under assessment and administrators that there are better and more transparent indicators than the famous *h* to apply in author-level assessment. The empirical studies presented in this PhD work and in the papers provide real world examples of the extent the concepts of authors, publications and citations are defined and operationalized in indicator construction and the ambiguities in benchmarking average performance with different ranking indicators.

Chapter 7: Conclusions and concerns

The present chapter aims to answer the research questions posed in Chapter 1. It provides the conclusions and a short discussion addressing concerns and future work. Through the research questions our knowledge of what constitutes the characteristics of ALI was extended (RQ1), the disciplinary and seniority appropriateness of ALI analyzed (RQ2) and the extent the concepts being measured are defined in the indicator model discussed (RQ3). Through the empirical investigations in the Papers and PhD body the theoretical and mathematical characteristics of ALI were analyzed and disciplinary and seniority appropriateness assessed. In the previous chapter, Section 6.5, a summary of the methodology displays the framework and interrelated structure of the entire PhD work and results from the 7 papers that led to the set of recommended indicators. The research questions are summarized below.

7.1 Summary of research questions

RQ1. What are the characteristics of ALI of academic performance?

ALI indicators range from raw counts of publications and citations *P* or *C*, to simple measures of central tendencies, e.g. the *f*, *t* and *CPP*-indices, to differential equations that are used to model the "real-life" effect of publication and citation impact to express publications and citations dynamically as a function of time, *rational h*. In Paper 1 each indicator was broadly categorised into what we at the time thought were characteristics: indicators of publication count, indicators that qualify output (on the level of the researcher and journal), indicators of the effect of output (effect as citations, citation relative to field or the researcher's body of work), indicators that rank the individuals work and indicators of impact over time, indicators that are *h*-dependent or independent. But this categorization proved a grouping based on "species" rather than characteristics. Our data suggest that the characteristics are mediated by other factors than a taxonomic relationship between species. ALI are inherently more complicated.

Factors that can be attributed to forming the characteristics of ALI are their conceptual and mathematical definitions. The concepts ALI operationalize are defined very differently, dependent on the purpose of the indicator. There is continued ambiguity in the scientometric community about what constitutes an author, publication, and citation but there is agreement that these concepts are multidimensional, but the difference between terminologies can be fluid, Chapter 2. Likewise in indicator development the approach to how to define these concepts and use them together to construct measures of "impact" or "effect" or "excellence"

can also be unclear and inconsistent. Yet we refrain from calling for a standardization of terminologies, as it is invalid to assume that a consensus among practitioners about terminology will lead to sound indicators. ALI are after all descriptive models based on qualitative assumptions about its variables, their interrelationships and the effect of assumptions external to model. Implications could be that this lack of definitive definitions is a methodological deficiency and a lack of documentation in the model could induce misuse and lead to false expectations of precision in application.

We were able to study the mathematical construction of indicators easily because in contrast to the conceptual characteristics of ALI, the mathematical characteristics are very clearly defined, Appendix B, Papers 1, 2, 6 and 7. This led to the discovery of a characteristic size-dependency of ALI. The number of publications or years used in the calculation of the indicator when they are used in rankings, links to a fall or rise in rank position, Section 6.3.1 and Paper 7. This characteristic dependency on publications can easily be manipulated in rankings. Further, I found that not all ALI are designed for evaluation purposes, Section 6.1.2, but are in themselves experiments aimed at gaining knowledge of paradigmatic mathematical laws and properties with the objective at moving the science of studying bibliometric distributions forward e.g. *hw*, *dynamic h* and *tapered h*, rather than application ALI in evaluations. The question is if I was right to compare these type of indicators with indicators designed for practical application to draw conclusions regarding the appropriateness of indicators in evaluation. Aware of this possible methodological flaw, the construction and objectives of the indicators studied in this PhD suggest that theoretical bibliometricians rather than "practical" bibliometricans have a dominant influence on the characteristics of indicators.

The characteristics of indicators are complex, perhaps verging on schizophrenic. In conclusion, studying the characteristics of indicators taught us that indicators are designed from very different disciplinary perspectives, with very different objectives, different operationalization of variables, different requirements to data, different mathematical models and very different inherent "personalities" which means they cannot directly be applied in evaluation without considering these characteristics. This knowledge leads us to consider if some models are indeed more appropriate in some disciplines or for some academic seniorities than others, leading to research question 2.

RQ2. To what extent are ALI appropriate in the evaluation of researchers from different disciplines and different academic seniorities?

No seniority appropriate indicators that fulfilled the criteria of simplicity in computation and data-collection were identified. Further, our attempts to group researchers using their academic titles were not informative. Papers 3, 6 and 7 confirmed that indicator values are determined by a ratio between the number of years since the researcher's first publication registered in the citation index and the total number of publications and citations credited to a researcher not academic position.

Different indicators were indeed found to more appropriate in some disciplines than others. Central indicators were identified (*hg* in Astronomy, *IQP* Philosophy, *h* Environmental Science and *g* in Public Health) which made disciplinary rankings using other ALI redundant (Paper 5). Additionally isolated indicators that could be applied across disciplines were also identified, *(%nc, P, C, APP, Age of citations,* etc). These indicators measure independently different aspects of researcher performance and their use is further justified in that the mean-based central indicators do not adequately represent what takes place in highly skewed citation distributions, as indicated in Albarrán et al (2011). The combination of central and isolated indicators will produce informative complements in a bibliometric evaluation and improve the appropriateness of the bibliometric analysis in the way the researcher's performance is described and interpreted, which is an important step as, as this PhD exemplifies, there can be conflicts in the representation between the researcher's CV and the citation index. Further, disciplinary-specific ALI that clustered scholars in groups of low, middle, high and extremely high performers were also suggested in Paper 7, but more research is needed about the robustness of these results and the extent the indicators could inform appropriate disciplinary-specific benchmarks. The suggested indicators are: the *h2* in Astronomy (cumulative achievement), *sum pp top prop in* Environmental Science (papers at the top of the field), *Q2* in Philosophy (effect of all productive papers), and *e* in Public Health (production and effect of highly cited papers). Finally, this PhD contributes with new knowledge on the ability of the *hg* index to improve the agreement in researcher rankings between WoS and GS, Paper 6. This finding is interesting as the discussion of how different databases provide a different picture of the researcher's impact is a matter of concern (Farhadi et al., 2013; Patel et al., 2013; De Battisti. & Salini.S., 2012; Bar-Ilan, 2008).

The above conclusions from the empirical studies are all well and good, but there is a "but". Exploring the extent ALI are disciplinary appropriate also included the necessary evaluation of their methodological construction and this led to significant problems in being able to claim that the indicators suggested above really are "appropriate", in that their homogeneity, inertia and correspondence with the concept they are designed to measure appears to be flawed, Chapter 6 and RQ3. Furthermore, there is the issue of bias and issues of generalizability.

The issue of database bias towards the hard sciences has been well documented in the scientometric literature. As expected calculating bibliometric statistics at the individual level thus means the indicator values are very subjective and heavily influenced by the researcher's specialty within the discipline, publication history, age and language of publication, representation in the citation index, as discussed i.a (Jasco, 2008) but further, the analysis of the characteristics and properties of indicators in Chapter 6 suggest an inherent bias in ALI, towards specific researcher profiles and outputs befitting the hard sciences.

The degree of the issue of generalizability was somewhat more surprising. The disconnection or rather the size of the gap between the performance of the researcher documented on their CV and the performance of the researcher based on bibliometric indicators, even hard scientists, was unanticipated, Paper 7. This has implications for the implementation of any ALI in any discipline as they can severely misrepresent the researcher. The disciplines and specialties studied in this PhD work have different publication to citation curves, and using indicators that are not designed to fit these curves sets up unfair and biased comparison. Consequently our data suggest that indicators should be treated as bias inducing mechanisms through which the mathematical models and database policies may further influence outcomes. Methods for developing, applying, interpreting and assessing indictors and importantly bias must be developed, specifically further methodological research should focus on how indicator bias is handled in evaluation to ensure appropriate use of indicators. Ideally, the representation of each researcher in the citation databases and their resulting indicator values should be judged individually to assess if observations are contrived or of real practical importance and at all comparable to established disciplinary standards.

Therefore it is vital that the bibliometric community clarify what it is that is being counted and how it is being counted, leading to research question three. The third question further addresses the methodological background of bibliometric indicators to assess if they measure what they are designed to measure.

RQ3. To what extent are the concepts being measured defined in indicator construction?

Even though citation and reference theorists have provided many definitions of the theoretical and operational concepts of citation, effect and impact, etc., the majority of the indicators studied in this thesis do not refer to these definitions or even provide their own definitions to improve their conceptual scheme. The concepts operationalized in the indicators can be poorly or only partially articulated, Section 6.1.1 and Appendix B. This does not mean however that the indicator models are unscientific, but on the other hand it does not mean that concept definition is not necessary. While the ALI studied in this PhD are a valuable resource for stimulating future work and attempt to improve the sensitivity and appropriateness of measurements in individual evaluation, they have limitations. It seems that sometimes foundational theoretical information is missing and providing information on how the model is not only mathematically but also theoretically constructed will give us a better chance of providing appropriate effect estimates. As Watt (1995) argues, using theory provides deeper explanation of the constructed phenomena; creates meaning, and links the subjective with the objective realms of science.

The motivation factor for this thesis is that evaluations effect people and a basic ethical principle should be that ALI are theoretically and operationally defined before the ALI is recommended in research assessment. We noted however that it is not the purpose of all the indicators studied in this thesis to be applied in evaluation but rather to explore the fit of bibliometric distributions after mathematical analysis (Ye, 2011). However, be the indicator designed as a study of a distribution function or a study of the impact of a set of researchers, one must consider that if the concepts are not defined or are defined in ways that do not stand up to scrutiny, the decisions and actions based on the ALI could turn out to be less effective than expected. Simply put, defining the concepts in ALI means we can critically examine the indicators and understand how the developer has done the measurement; we can repeat the measurement, reflect on the meaning associated with the concept, and compare the conclusions with previous findings. Finally, thinking methodically about the varieties of real world phenomena which should be encompassed by our concept label will often suggest improvements to the theoretical definition. Any changes in the theoretical definition imply corresponding changes in the operational definition, and vice versa and this will improve the face and measurement validity of the indicator (Watt, 1995; Gingras, 2014). Ultimately, as discussed in Chapters 2, 3 and 6, good theoretical definitions will aid us in selecting valid

operational measurement items and help us learn more about conflicting findings in different studies focusing on the same phenomenon.

7.3 Implications and Epilogue

Providing well-founded conceptualization and operationalization that draws on theories of authorship, publication and citation, and verbalizing ambiguities in the level and interpretation of "effect" will contribute to preventing indicator values being reduced to "meaningless numerology" and a "*practice that belongs to the penumbral world of professional ritual*", (Cronin, 1984). Edge provides a critical review of the implications of quantitative indicators that is still relevant today, (Edge, 1979). He argues that developers of indicators make implicit assumptions about the nature of science and they "*Gloss over as unproblematic precisely the points that we find to be crucially at issue*", (Edge, 1979, p.102). This could explain why concepts of citation or impact, amongst others, are either diffusely defined or ignored in indicator construction, as they are indeed problematic and "messy". Yet in the absence of theoretical foundations, the main characteristic of bibliometric indicators may be interpreted as only indicating the visibility of a researcher's papers in the context of the citation index and say nothing about the researcher's conceptualized "effect", "impact" or "excellence".

It is not my intention in this PhD work to belittle the craftsmanship required to create the robust mathematical models that form the indicator or dismiss the developers work because the ALI are hampered by the absence of theoretical foundation, random sampling or their applicability inhibited by technical issues and data-incompleteness. The papers indicators are presented in address issues such as how the development of the indicator contributes to closing gaps in knowledge about research performance and solutions to novel quality problems which have not been formulated beforehand. This is important work, which I by no means disregard in this thesis. The problem as I see it is a fundamental one - the absence of robust theoretical definitions of the concepts the indicator is designed to measure. Concept definitions drawing on the theoretical heritage of bibliometrics as well as the operational functions the models perform will contribute to *strengthening* the validity of indicators as appropriate measures of research performance. If developers are intent on treating indicators as mathematical models that capture real life phenomenon that can be observed, then they must also explain the philosophical and scientific foundation of the objects being measured.

It is understandable that indicator construction appears to be orientated to the plausible approach rather than the accurate because the theoretical objects that are attempted to be measured, objects such as citations, impact or use, have unobservable qualities (Kitchin, 2014). So based on some mathematical criteria attention is focused on the indicators that seemingly offer the most likely or most valid way forward. This approach means that indicators gain empirical meaning for the observed terms alone such as number of citations to a document registered in a citation index, and only partial meaning is drawn up to the theoretical meaning of a citation, by osmosis as it were (Putnam, 1979).

7.3.1 Implications for the end-user – evaluand and evaluator

ALI feed both the narcissism and insecurities of researchers: indicator values are used to grade good, better, best researchers, they reduce a researcher's profile to a score that can be easily compared, establish the legitimacy of a researcher, and used to justify decisions, investments and promotions. ALI produce seemingly verifiable numbers and give the impression they are easy to understand, valid and representative because they come as numbers: they are marketed in ready-to-use packages, popular, hated, discussed, compared, and developed prolifically. They provide simple ways to objectively measure the creativity, talent and prestige of a researcher across his or her career, which is why bibliometric methods are embraced by an ever-increasing number of users but with ever-decreasing regard for validity and reliability, (Gläser & Laudel, 2007, p.101-123). It is of great concern that ALI will become indispensable for evaluating research. Regardless of the vagueness in concept definition and the ambivalence of the bibliometric community to implement guidelines for indicator development (Hicks et al., 2015), the power of numerical values to communicate strengths and weakness of academic enterprise is headily attractive.

Our results show that only a few indicators are built on reflection of the conceptualization and operationalization of the core concepts of author, publication and citations as well as the mathematical equation, Appendix B Tables 6, 14, 15, and 16. This is problematic as these concepts do matter in researcher assessments. ALI are formalized descriptors of performance, used in steering the direction of research, investment in science, researcher assessment and understanding the sociology of science. ALI make people not publications the objects of study and put a sociocultural value on a person (Day, 2014), and (Paper 4). At the heart of this PhD work are metrics which foundationally are not valid, and politically driven by an evaluation culture and market values rather than professional practice and the "rules of science" (Wouters, 2014b; Dahler-Larsen, 2012). Consequently constructing indicators for

mainstream application remains a challenge, however the discussion of the conflicting agendas between the bibliometric community and the evaluation community falls outside of the scope of this paper, but can be referenced amongst many others in (Dahler-Larsen, 2012; Schneider & Aagaard, 2012) (Bach, 2011; Seglen, 1997).

The appropriateness of indicators in evaluation of the individual researcher is extenuated by the problem of representation, visibility dynamics, misinterpretation of how to apply the indicator, and the issues ambivalence, inconsistency, exogenous variables, and commercialization, raised in Sections 2.2.1 to 2.2.6. A further difficulty is partial interpretation. At the individual level partial representation can have direct implications because indicators are built on theories with false observational consequences, such as the number of citations recorded to a document in a citation index informing on the quality of a document or prestige of the author. This has no interpretation meaning, and the mathematical approach and the existing theories used to interpret the numerical values produced on this premise are not wrong, but can be senseless (Putnam, 1979). How can we accept a conclusion that is based on undefined concepts and apply it to say something about the quality of a researcher?

7.3.2. Implications for developers

If we are to advance indicators and provide appropriate indicators in evaluations, directly interpreting only the observational terms of author, publication and citations as recorded objects in a citation index is not an acceptable model as an applied theory in indicator construction. To a greater or lesser extent the developers are reflecting on the size and frequency of bibliometric distributions to patterns in behaviour around publishing and citing, but still for methodological reasons design the indicator to fit the specifications of a specific citation index, Table 6. They are using a theoretical mathematical ballast to home in the precision and reliability of the mathematical procedures and techniques (Wouters, 1999) yet mathematical equations alone do not say anything about the physical or social causes behind the observed bibliometric distributions. And there is a fundamental problem in that developers *do not* define what it is they are measuring. They do not fully define the concepts operationalized in the indicators nor do they define how the impact or effect of the researcher is connected to the mathematical structure of the indicator, which means that indicators can be robust and valid but can still be criticized for arbitrary cut off values and parameters Appendix B. Complex mathematical expressions make it difficult to determine what a change

in the indicator value means because the objects and concepts measured in the indicator are given meaning through mathematical logic (Putnam, 1979). As the indicator model becomes more sophisticated it becomes just as difficult to understand as the real-world processes it represents. Yet, people have difficulty making sense of results when more and more variables and functions are interacted (Starbuck, 2006, p.101). Are changes due to a change in the combination of data, a change in one of the elements in the mathematical equation, a change in the performance of the researcher, or a change in the relationship between the equation and the completeness of the bibliometric data? Designing mathematically robust indicators is important work but proof theorems and quantifiable certainty are not enough to base the futures of researchers under evaluation on. The best indicators combine robust mathematics with sensitivity to the concepts being measured and fit the purpose of practical application (Hicks et al., 2015). The theoretical discussion in Chapter 3 and following analysis in Chapter 6 suggested that the majority of indicators studied in this PhD work are not as theoretically robust as they were mathematically robust, but are rather as Kahneman describes *"artistically-crafted illusions of validity that continue to be developed because they are supported by a powerful professional culture"* (Kahneman, 2011). The fact that indicators are readily available in citation indices and continuously developed adds to the legitimacy of this culture and the legitimacy of the indicators in the eyes of the bibliometric community (Haustein & Larivière, 2015) and the legitimacy is reinforced by the continued application of ALI by end-users. A major implication for developers is indicators can be conceived as a quest for possible biases, distortions and measurement errors rather than tools for author-level evaluations, as investigated in Papers (5, 6 and 7).

7.3.3 Implications for future research

Consequently, while much research has been done in developing novel indicators, few studies have been done on the validity of these proposals and conflicts among the concept of author, publication and citation and the mathematical operationalization and they have not focused specifically on the appropriateness of indicators in evaluations. Future research should reanalyze ALI using raw data sets from other databases and disciplines and compare these results in study reports and subsequent publications. Future development of indicator crucially depends resolving the culture of ambivalence (Wilsdon, 2015; Hicks et al., 2015; Gläser & Laudel, 2007; Wouters, 1999). Bibliometricians should not only develop indicators but give end-users, evaluand and evaluator, access to protocol and transparency guidelines as

the indicators can easily be manipulated and often biased. Additionally, the advantages and shortcomings of ALI should be disclosed and better managed.

Like most indicators ALI have gone a long way in alienating the researcher group they purport to benefit (Wilsdon, 2015). This is an on-going problem with evaluation studies, not just bibliometrics, see Paper 4. ALI dehumanize researchers by reducing them to a few data-values that in turn are based only on the researcher's works that are registered in the citation index. Accordingly, a future direction steering the requirement to indicator development is to discuss the immediate and long term need to evaluate and highlight the psychological and career-related implications of author-level bibliometric evaluation as well as the practical limitations of ALI, (Wilsdon, 2015). This direction was unfortunately not fully assessed in this thesis because the primary focus was to study the building blocks and characteristics of ALI. We found that ALI are created as statistical exploration of distributions rather than advancing knowledge about the performance of researchers and the future of explorations like these must address how these distributions explain the phenomena of citing and the process of scientific communication. Such an approach is appealing and supports the mathematical strengths of indicator construction, and interpretation of the numerical values, downplaying the role of modelling and generalization as required in theoretical science, and rather describing and simulating real life phenomena as experimental and computational science (Kitchin, 2014).

Data is not generated free from theory, and free of human bias or framing (Gould, 1981). Making sense of bibliometric data is always full of values, context, domain knowledge, technical and human influences and of course disagreements, as Jascó pointed out (Jasco, 2005a) and discussed in Paper 6. By looking into the background of developers of ALI, Section 6.2.1, we discovered that anyone with a reasonable understanding of how science in their subject is produced, published and cited can publish an indicator designed to capture social processes, networks, collaborations. In the dataset there are indicators developed by Horticulturists, *wu* (Wu, 2010), Meteorologists, *b* (Brown, 2009), Oceanographers, *hT* (Andersen, 2008) and Dermatologists, *x* (Namazi & Fallahzadeh, 2010). Subject expertize from both bibliometricians and domain experts is essential to assess the validity and appropriateness of indicators, especially as indicators deal with human behaviour, (Kitchin, 2014). Without subject matter experts to articulate problems in advance, the indicators will undoubtedly produce poor and invalid results (Porway, 2013). Consequently future research

should be done in increased collaboration between bibliometricians, computer scientists and information sciences (the green cluster in Figure 5) and domain experts. The reanalysis of existing ALI in such collaborations will thus be the starting point in revealing what do the distribution patterns the ALI capture mean and what are their social consequences? This requires theory *and* contextual knowledge to make sense of how to construct indicators and also how to employ them and frame their worth (Hicks et al., 2015; Wouters, 2014a; Wouters, 2014b; Wouters, 1999; Watt & van den Berg, 1995; Putnam, 1979)

7.3.4. Epilogue

Until there is an open critical reflection between indicator developers and end-users, such indicators as the *h*-index and its variants will prevail, even though they have long been discussed in the bibliometric literature as inconsistent measures and have been mathematically proven to be flawed and alternatives are available. This PhD work recommends a set of indicators, that are perhaps not as famous or sexy as *h,* but are none the less valid indicators that produce useful information provided the resulting values are interpreted within the limits of the citation index used to source the bibliometric data. Consequently, this thesis highlights both the absence of theoretical ballast and the absence of openness around what it is ALI actually measure

Twenty years after the call for bibliometric standards, (Glänzel, 1996; Glänzel & Schoepflin, 1994), what bibliometric indicators aim to measure is still unclear; ambiguous labels such as "impact" or "excellence" are thrown about like buzz words but are not conceptualized and supported by the wealth of theoretical ballast the field of bibliometrics supplies. But this is not the same as saying that ALI should be discarded, as the deficiencies of the indicators cannot be attributed to the indicator alone. Rather we must readdress the principles that guide research evaluation as well as correct the deficiencies in indicator methodology and the documentation used to defend them. This will contribute to reducing misplaced concreteness and false precision (Hicks et al., 2015). The uninformed use of ALI risk can overrate the quantity of scientific publications which can compromise research quality and integrity of the individual researcher (Aagaard, 2015). Yet still assessment committees and institutional leaders are paying attention to them. ALI are increasing in importance in assessments that influence careers, money, science, jobs and communication (Wilsdon, 2015) and transparent documentation of the construction of indicators and the rationale for indicator validity needs

to be accessible to a wider audience, not just the technical few. With the increased use of bibliometrics in evaluations there is an increased need for guidelines for good, objective evaluation practice. Objectivity is an inherent value of evaluation of scientific performance in researcher assessment and it is important we continue to find ways to incorporate the complexity of the communication and use of science in research evaluation. The fear in the bibliometric community is that ALI are on the brink of being implemented in institutional evaluations placing too much emphasis on narrow or poorly-designed indicators (Wilsdons, 2015; Hicks et al., 2015; Glänzel, 1996). I further suggest that application is problematic because many ALI are published at the prototype stage and remain that way without follow-up reanalyses. M. Zitt stresses:

> "*[…]the contrast between the highly sensitive nature of evaluation issues and the eagerness of scholars or administrators to elicit a number and forget crucial warnings about statistical distributions and methodology artifacts*" (Zitt, 2005).

Therefore to ensure the transparency of evaluation using ALI precautions are sorely needed as author-level bibliometrics may reinforce this detrimental evaluation behavior. Ultimately the following issues need to be addressed:

1. The lack of conceptual clarity, the ambivalence problem, the representation problem and resulting validity problems, discussed in Chapter 2 and 6, which could result in people in the evaluation system (evaluator or evaluand) choosing from a range of indicators without fully identifying and assessing the unique approaches of each.
2. Author-level bibliometric indicators are to some point equally weak and coherence is a major challenge.
3. Monitoring and evaluating indicator production also appears neglected. The lack of conceptual clarity in indicator construction is a warning of structural problems in the focus of the indicator that must not be ignored.

The way to improve this situation is to update understanding of concepts, constraints, technical problems, inadequate sources; address the ad hoc nature of indicator development, give guidance in application, promote cohesive evaluation to prevent conflict and most importantly, correct the weak follow-up of indicator development. Not a short list but this thesis is a practical step in the right direction.

Appendix 1

Table 6. Definition measure, author, publication, citation, developer specialty and cluster membership

Indicator	Designed to Measure	Definition of Author	Definition of Publication	Definition of a Citation	Specialty of Developer(s)	Cluster
H	Quality & quantity	Researcher with scientific output / Price winner	Papers, (implicit in citation index)	Broad indication of overall scientific impact	Physics	
m-quotient	Effect best papers	Researcher with scientific output		Popularity		
G	Rank	Author/Price winner	Papers in WOS	No definition.	Mathematics, Methodology Of Social Science, Information Science and Retrieval, Quantitative Social Research, Scientometrics	
c(t)	Currency	Scientist with a bibliography	Items in sources	Use over time		
a(t)						
fc	Independence	Author	Papers in WOS	No definition		
hw	Quality	No definition	Publications WOS	Use Hirsch definition_ Broad indication of overall scientific impact		
AR	Growth	Price award winner	Articles in WOS	Performance		
mg-quotient	Rank	Author/Price winner	Papers in WOS	No definition inferred citation as a function		
H^c	Currency	Author	Proceedings, articles, conference series, journals in the DBLP digital library	"Conducting theoretical analysis of the properties of the proposed indexes is the next step in this work, but it is beyond the scope of this paper"	Computer Science, Scientometrics, Distributed Systems, Databases	
Hn	Comparison across fields					
H^t	Pioneer research					
Index of age & productivity	Career	Differentiates between author as a tenured scientist, research scientist and research professor that are actively productive	All types of publications in WOS	Rewards for products Academic dividend, others are awards, students, and general visibility within their discipline	Bibliometrics, Scientometrics, Databases, Citations	
Classification of Durability	Durability	No definition	Documents covered by WOS	impact and transfer of knowledge of documents beyond their original producers		
%HCP	Excellence	Differentiates between author as a tenured scientist, research scientist and research professor that are actively productive	All types of publications in WOS	impact and transfer of knowledge of documents beyond their original producers		
IQP	Excellence	A researcher who has published	Papers indexed in WOS	Quality, proxy of influence	Leadership, Psychometrics, Leadership Development, Organizational Behaviour, Research Methods, Labour Economics, Public Economics, Social Economics, Microeconometrics	
m	Effect of best papers	Researchers with publications (implied)	Papers	Impact	Sociology Of Science, Peer Review, Evaluation, Research On Higher Education, Citation	
e	Excellence	Scientists	Papers	No definition	Physics	
hg	Ranking	Scientists	Published papers	Impact (accepts Hirsch definition)	Software Engineering, Information Retrieval, Artificial Intelligence	

h2	Excellence	Scientists	Papers in WOS	No definition	Surface And Colloid Chemistry	
Hpd	Growth	Scientists	Scientific papers in WOS	No definition		
A	Distribution of citations	Scientists	Articles	Indicates acceptance of Hirsch definition	Information Science, Science Culture Communication	
R	Growth	Price award winner	Articles in WOS			
h	Distribution of citations	Author	Article, Letter, Review, Correction, Editorial Material, or Note in WOS	Quality/interest	Physics	
Q2	Effect of all papers	Author	Scientists research output in WOS	No definition	Soft Computing, Computing With Words, Consensus, Fuzzy Logic, Multiple Criteria Decision Making, Fuzzy Sets, Scientometrics, Computer Science, Artificial Intelligence, Algorithms & Theory, Information Retrieval	
Hα	Comparison within specialty	Researchers, Price award winners	Papers in WOS	(inferred) citations are an indication of quality, performance (field dependent)	Bibliometric Mapping, Visualization Tools And Algorithms	
Gα	Quality			Quality (field dependent)		
b index	excellence	Author	Papers in WOS	No definition	Analytical Chemistry, Chemical Metrology, Measurement Science, Air Quality	
hT index	Effect all papers	Author	Papers in WOS	Influence over time	Oceanography, Ecosystem modelling, Computational Statistics, Social Networks, Software Design	
Rational h	Rank	No definition	Published papers in disciplinary/scientific indices: Scopus, WOS, Econlit, IDEAS/REPEC	No definition	International Economic And Industrial Development, Economic And Social Research, Environmental Economics, Energy Economics, Economics, Climate Change, Scientometrics	
f	Average effect	Prolific researcher	Papers indexed in disciplinary/scientific databases (IDEAS/REPEC, WOS)	No definition	Environmental Economics, Energy Economics, Economics, Climate Change, Scientometrics	
t						
Rational g	excellence	No definition	Published papers in disciplinary/scientific indices: Scopus. WOS, Econlit, IDEAS/REPEC	Quality		
Wu index	Broad impact of masterpeices	A scientist with papers	Papers in WOS	No definition	Horticulture And Gardening	
Hm	Independence	A scientist with papers	Papers (implicit WOS)	No definition	Physics	

n index	Comparison within specialty	A scientist with papers	Papers indexed in Scopus	Performance	Dermatology, Molecular Dermatology, Nephrology, Transplantation, Dermatology, Diabetic Nephropathy, Hemodialysis	
H index sequences and matrices	Comparison to peers & domain	Productive scientists	Papers in WOS	Impact over time	Quantitative Social Research, Scientometrics	
hf	Comparison to peers	No definition	Publications classified as "article" or "letter" in WOS	Discusses the caveats of citations, but doesn't come with a definition of how they define a citation. Hints at "importance" normalized for field variation.	Complex Systems, Networks, Science Of Science, Sport Statistics, Statistical Physics Of Social Dynamics, Science Of Science, Complex Networks, Opinion Dynamics	
π index	Comparison across similar fields	Active scientists with papers	Journal Papers with available data in citation indices	Influence	Biocomplex Research, Scientometrics, Computer Science, Information Science	
x index	Quantity & quality	Researchers with at least 15 publications	Papers in journals covered by JCR if co-authorship score of 2/(Np + 1).	(Inferred) Quality, dependent on field variation	Technology And Operations Management	
Alternative h	Independence	Researchers that have authored papers	Research output in WOS	Effect	Nuclear Physics, Computer Science, Statistical Physics, Mathematical Physics, Methodology Of Social Science, Chemical Physics & Material Physics	
POP h	Independence	Person listed on author-byline of a published paper	Papers returned by Google Scholar or Microsoft Academic Search in reply to a query.	Impact	HQ-Subsidiary Relations, International HRM, Language In IB, Quality & Impact Of Academic Research	
AW	Effect of all papers					
AWCR						
AWCRpa	independence					
Pure H	Independence	Person listed on author-byline of a published paper	Publications	No definition	Information Science, Scientomtrics, Bibliometrics	
Adapted Pure H	Independence	Author	Articles			
Dynamic h	Rank	Scientist, first author on papers	Papers (implicit in citation index)	acceptance of Hirsch definition		
Price index	Currency	A person working at the research front who has produced a paper	Papers are an expression of the state of a researcher at a particular time, a concept of a hypothesis	The usefulness of literature as a function of its age	Physics, History Of Science, Information Science, Scientometrics,	
DCI index	Quality	A person that has produced objects that have received citations	Objects that have received citations	Impact that is devalued with the age of the citation	Information Retrieval, Bibliometrics	
Hmx	rank	Active academics with scholarly production	Publications listed in citation indices: WOS. GS, Scopus	No definition	Information Retrieval, Search	

Appendix 2

Table 7. Number of publications reported on publication lists and identified in WoS and GS

	Publications on list	Publications WoS	Publications GS
Astronomy			
PhD	295	151	347
Post Doc	2230	1464	2913
Assis	1878	1543	2979
Assoc	9811	6138	12555
Prof	6955	5483	9333
Environmental Science			
PhD	24	13	42
Post Doc	912	274	901
Assis	1994	802	1857
Assoc	7936	3339	6872
Prof	5854	4228	7780
Philosophy			
PhD	139	15	112
Post Doc	654	160	457
Assis	1709	509	1376
Assoc	3373	757	2978
Prof	8815	2359	9668
Public Health			
PhD	114	106	153
Post Doc	365	177	276
Assis	1554	1096	1977
Assoc	3720	2763	5033
Prof	3314	3260	4948

Appendix 3

Table 8. Dataset 1. Publications and citations to 741 researchers in Web of Science

Publications					Citations		
Discipline	**Sample**	**Range**	**Median**	**Mean**	**Range**	**Median**	**Mean**
***Astrology*, 192 researchers**							
Ph.D	15	2-36	7	10.8	8-529	150	149.4
Post Doc	48	3-103	19.5	26	3-3177	201.5	561.1
Assis Prof	26	10-142	39.5	51	69-4009	702	1118,6
Assoc Prof	66	7-292	61.5	77.7	19-9083	1214	1981.1
Professor	37	34-327	90	121.3	177-16481	1889	3579.1
***Environmental Science*, 195 researchers**							
Ph.D,	3	3-5	4	4	16-60	34	36
Post Doc	17	2-59	9	12.8	10-642	41	91.7
Assis Prof	39	2-46	18	19	0-573	148	185.4
Assoc Prof	85	1-103	29	36.8	2-2519	326	520.1
Professor	51	1-425	51.5	59.7	6-14141	435	998.1
***Philosophy*, 222 researchers**							
Ph.D	8	1-5	1	2	1-33	0.5	6.2
Post Doc	22	1-31	4	7	0-235	8	21.4
Assis Prof	43	1-106	6.5	10.8	0-1829	6.5	74.3
Assoc Prof	74	1-45	7	10	0-565	8	50.7
Professor	75	1-140	18	28.1	0-3495	29	157
***Public Health*, 132 researchers**							
Ph.D	9	4-27	8	12.2	7-253	60	82.2
Post Doc	14	1-23	11	12	0-353	80.5	113.6
Assis Prof	30	3-288	22	36.2	10-3796	167	417.4
Assoc Prof	50	4-220	43	54.6	4-3649	518	778.5
Professor	29	5-661	76	110.2	13-13520	954	2104

Appendix 4

Table 9. Dataset 2. WoS and GS combined

Discipline	Publications WoS	Publications GS	Citations WoS	Citations GS
Astronomy, n190	***12318***	***28044***	***320971***	***513611***
PHD, *n13*	121	285	1769	2487
Post Doc, *n48*	1249	2920	26933	40764
Assis Prof, *n26*	1328	2978	29084	41923
Assoc Prof, *n66*	5130	12548	130756	206576
Full Professor, *n37*	4490	9313	132429	221861
Environmental Science, n99	***3228***	***7425***	***34851***	***62351***
PHD, *n2*	7	34	76	141
Post Doc, *n7*	58	322	235	1086
Assis Prof, *n21*	337	777	2569	4281
Assoc Prof, *n44*	1413	3424	13996	26045
Full Professor, *n25*	1413	2868	17975	30798
Philosophy, n155	***2223***	***9538***	***7108***	***66386***
PHD, *n5*	6	19	9	48
Post Doc, *n16*	91	323	182	865
Assis Prof, *n24*	192	526	353	2753
Assoc Prof, *n53*	454	1856	1121	6671
Full Professor, *n57*	1480	6814	5443	56049
Public Health, n68	***4374***	***7220***	***60441***	***104637***
PHD, *n4*	52	76	172	468
Post Doc, *n3*	25	53	164	526
Assis Prof, *n17*	621	1024	6224	11269
Assoc Prof, *n31*	1743	3236	22197	41129
Full Professor, *n13*	1933	2831	31684	51245
Total	***22143***	***52227***	***423371***	***746985***

Appendix 5

Table 10. Developer specialty and collaboration in indicator production and indicator typology

Colour	Cluster Members	Objective of the indicator in evaluation	Definition Author: Publication	collaboration
Red	Economics, Evaluation and Education, Environmental Science, Leadership, Psychology, Oceanography	effect of work, excellence, influence, success	No definition: Paper in WoS Scientist:WoS Scientist :Paper in other index Published/cited:WoS Author :WoS Published:Paper Published :WoS Published:Paper in other index	15/51
Blue	History and Sociology of Science, Networks, Statistics and Quantitative Research, Systems	authority, currency, comparison, performance changes & growth,	No definition: Paper in WoS Award winner:WoS Scientist:Paper Scientist:WoS Published/cited:Expression Author:WoS Author:Papers in other index Published:WoS Published:Paper in other index Published:Paper No definition: Paper in WoS	23/51
Green	Bibliometrics and Scientometrics, Computer Science and Software Engineering, Information Science and Retrieval	ranking, comparison, independence, average performance, currency, quality defined as excellence & pioneer research, performance changes & growth	No definition: Paper in WoS Seniority:Paper in WoS Award winner:WoS Scientist:Paper Scientist:WoS Scientist:Paper in other Published/cited:Expresssion Published/cited:WoS Published/cited:Object Published/cited:Paper in other index Author:Papers Author:WoS Author:Papers in other index Published:Paper Published:WoS Published:Paper in other index	38/51
Dark Green	Chemistry, Physics and Mathematics	authority, excellence, independence, currency	No definition: Paper in WoS Award winner:WoS Scientist:Paper Scientist:WoS Published/cited:Expression Author:WoS Published:WoS Published:Paper in other index	19/51
Purple	Medicine	comparison across specialties	Scientist:Paper in other index	1/51

Appendix 6

Table 11. Validation of indicators of publication count

Indicator	Concept being evaluated	Corresponds to concept	Corresponds to inertia	Homogeneous	Rationale
P	Production of papers published in journals and by academic book publishers	Yes	Yes	Yes	Counts papers published in journals and by publishers
Fp (APP arithmetic)	Independence	No	Yes	Yes	Gives an equal fraction of a publication to each author. Indicates nothing about how much work a scholar would produce independently
APP proportional	Contribution	No	No	Yes	Weights the fraction of a publication to each author, according to their place in the author by-line. Reveals nothing about actual contribution or if contribution increase with position on by-line.
APP geometric	Contribution	No.	No	Yes	As above
APP harmonic	Contribution	No	No	Yes	As above
FA	Contribution	No	No	Yes	As above
Noblesse Oblige	Contribution	No	No	Yes	As above
Weighted Publication count	Production of specific types of publication	Yes	Yes	Yes	Counts and weights different types of publication separately according to the assessment or discipline
Publications in selected sources	Production in selected sources	Yes	Yes	Yes	Counts publications in sources defined as important by scholars, their institute, discipline or assessment
Cognitive orientation	Cognitive orientation	Yes	Yes	Yes	Counts and aggregates papers according to scientific subfields the individual publishes or is cited in.

Appendix 7

Table 12. Validation of indicators of citation count.

Indicator	Concept being evaluated	Corresponds to concept	Corresponds to inertia	Homogeneous	Rationale
C	Times cited	Yes	Yes	Yes	Counts citations, including self-citations
Database dependent citation count	Times cited in specific sources	Yes	Yes	Yes	Counts citations recorded in a specific database
C-sc	External citations	Yes	Yes	Yes	Sum of citations, excluding self-citations
Sig	Most significant work	No	No	Yes	Indicates paper with the highest number of citations. This is not necessarily the most significant work.
Sum pp top prop	Identify scholar's papers that are rated top of their field	No	No	Yes	Proportion of papers in the top 10% of the world.
%nc	Work that has not been cited	Yes	Yes	Yes	Computes the share of papers that have not received citations
%sc	Self-use	Yes	Yes	Yes	Computes the share of citations a scholar gives to his or her own work.
Fc	Number of citations scholar would have received if worked alone	No	No	Yes	Gives an equal fraction of a citation to each author of a paper. Indicates nothing about how many citations a scholar would receive of worked independently

Appendix 8

Table 13. Validation of ALI (Hybrid) indicators

Indicator	Concept being evaluated	Corresponds to concept	Corresponds to inertia	Homogeneous	Rationale
h	Quality and quantity	no	no	no	The value of h does not go up when the notion of quality goes up, as the value of h correlates with number of published papers and is determined by it
CPP	Average number of citations per paper	yes	yes	yes	Computes the average number of citations per paper
m-quotient	Effect of best papers	no	no	no	m-quotient is the h index divided by the number of years since the scholars first publication or PhD defence. Suffers from same deficiencies as h and unclear the extent a change in the m-quot value is due to number of publications, citations or years used in the computation
g	Rank of scholar	no	no	no	G is disproportionate to average publication rate. The G-index of a scientist with one big hit paper and a mediocre core of papers could grow in a lot comparison with scientists with a higher average of citations
c(t)	Currency	yes	yes	yes	Computes the age of the citations referring to a scholars work, providing insight into the sustainability or obsolescence of a scholars work
a(t)	Currency	yes	yes	yes	Computes the age distribution of citations to a set of documents
hw	Quality	no	yes	no	Computes the square root of the total weighted citations received by the highest number of articles in the h-core, using their rank position to indicate the number of citations articles in the h-core have received over time. Assumes correlation with continuously cited articles and quality.
AR	Growth	no	no	no	Deficiencies as h-index. Divides the citation counts of articles in h-core by the raw age of the publication. Thus the decay of a publication is very steep
mg-quotient	Rank of scholar	no	no	no	Mg-quotient is the g-index divided by the number of years since the scholar's first publication or PhD defence. Deficiencies as m-quotient
Hc	Currency	yes	yes	no	Gives an indication of the age of the articles in the h-core and the 4 year evaluation window is appropriate at the individual level, but suffers from the same deficiencies as h and the weighted parametric is unclear.

Hn	Comparison across fields	no	no	no	Hn Define how many articles are included in the h-index and subtracts this number from the total number of publications. Can only be used as a supplement to h, has the same deficiencies as h and can produce paradoxical results for scholars with only few publications
Ht	Pioneer research	yes	yes	no	Ht computes if articles still get citations by looking at the age of the citations. Each citation is assigned an exponentially decaying weight to estimate the impact of a scholar's work in a particular time instance. Suffers from the same deficiencies as h.
Index of Age & Productivity	Career	yes	yes	yes	Computes the mean number of publications by age and CPP in 4 year age brackets, adjusted to research fields as defined in Web of Science (WoS)
Classification of Durability	Durability	yes	yes	yes	Computes the percentile distribution of citations that a publication receives each year, accounting for all document types and research fields, as defined in WoS.
%HCP	Excellence	no	yes	yes	Indicates highly cited articles. High citation count may or may not be a facet of excellence
IQP	Excellence	yes	yes	yes	Computes the number of citations a scholar's work would receive if it is of average quality in the specialty. The specialty is defined by the top publications the author publishes in.
m	Effect of best papers	no	no	no	Median number of citations to publications in the h-core. Suffers the same deficiencies as h.
e	Excellence	no	no	no	Has to be combined with the h-index to give useful information.
hg	Rank of scholar	yes	no	no	Hg combines the h and the g index. Combining h and g does not improve discriminatory power and no direct meaning in terms of papers and citations of a scholar
H2	Excellence	no	no	no	Only a small subset of a scholars papers is used to compute the H2 index. Scholars with high H2 values can differ greatly in the number of papers and citation rates.
Hpd	Growth	yes	no	no	Hpd uses a scaling factor of 10 to improve granularity between researchers but is as an arbitrary number, which randomly favours or disfavors individuals. 10 years is a very long publication window at the individual level.
A	Distribution of citations	yes	no	no	A is the arithmetic average of the amount of citations a publication in the h-core has. Same deficiencies as h.
R	Growth	no	no	no	R is the square root of the A index and computes the magnitude of citations to publications.

ħ	Distribution of citations	yes	no	yes	ħ is the square root of half of CPP. The value of ħ increases with the notion of quality articles receiving more citations and is not dependent on the h-index.
Q2	Effect of all papers	no	no	no	Q2 is the square root of the geometric mean of h-index multiplied by median number of citations to papers in h index. Suffers same deficiencies as h.
Hα	Quantity OR quality	yes	no	no	Computation is the same as h, and there is no agreement on the value of α and it can be manipulated
Gα	Rank AND quality	yes	no	no	No agreement on the value of α, based on same ideas as g-index and more tests are needed to understand its performance
b index	Excellence	no	no	no	Demands computation of the h-index, identification of self-citations and field specific reference standards
hT index	Effect all papers	yes	yes	no	Conceptually complex, evaluates the complete production of the researcher, all citations giving to each of paper receives a value equal to the inverse of the increment that is supposed to increase the h-index one unit.
Rational h	Rank of scholar	yes	yes	no	The rank of the scholar increases in smaller steps providing greater distinction between individuals but still dependent on computation of h-index
f	Average number of citations per paper	yes	yes	yes	Computes average number of citations per paper
t	Average number of citations per paper	yes	yes	yes	Computes average number of citations per paper
Rational g	Excellence	no	yes	no	The rank of the scholar increases in smaller steps providing greater distinction between individuals but still dependent on computation of g-index
Wu index	Broad impact of masterpieces	no	yes	yes	Describes the quantity of a scholars productive core of papers.
Hm	Independence	no	no	yes	Uses fractional citation counts to compute h-index: counteracts the h value being determined by the number of publications, but combing citations, papers and authors is not informative about independence and suffers same deficiencies as h.
n index	Comparison within specialty	yes	no	no	n is the h-index divided by the highest h-index of the journals of a scholars major field of study. Scholar h is computed in Scopus, journal h in SCImago
H index sequences and matrices	Comparison to peers and domain	yes	no	no	Calculates h-sequences by continually changing the time spans of the data. The value depends on the citation and publication window. Requires specialist software to construct a h-matrix based on a group of correlative h-sequences

hf	Comparison to peers	yes	no	yes	Each paper is normalized by the average number of citations per paper in the subject category of the paper under observation. Value is dependent on the publication and citation window used in the construction of the matrix
π index	Comparison across similar fields	yes	yes	yes	π can be calculated on a small number of papers and is unique because it is defined in terms of the summed number of citations rather than the square root of the sum or the average
x index	Quantity and quality in cross disciplinary comparisons	yes	yes	no	Combines publication count, counting authors per paper, 5 year impact score of journals the scholar published in, a weighted average of the absolute scores for scholars with more publications in each journal where the author has published and a co-authorship coefficient. The magnitude of x depends on the specific magnitudes of the differences and weights of the publications.
Alternative h	Independence	no	no	no	Alternative h, is the h-index divided by the mean number of authors in the h publications. Suffers same deficiencies as h.
POP h	Independence	no	no	no	Normalising by mean number of authors leads to a reduction of the h-value that does not increase with the expectation of independence. Calculates h using fractionalized publication and citation counts
AWCR	Effect of all papers	yes	yes	yes	Computes average CPP by dividing the citations to a paper by the age of the paper, summing over all papers and dividing by total number of papers.
AW	Effect of all papers	yes	yes	no	AW is the square root of AWCR to reduce effect of a few highly cited papers on the average score. It is more rigorous to assign weights to each of the publications, calculate the average weighted citations (using the arithmetic mean), and then normalize that result to one of the publications
AWCRpa	Independence	no	no	yes	Citations to a given paper divided by age of that paper and number of authors, summed over all papers is the value indicating independence and effect the scholar would have had working alone.
Pure H	Independence	no	no	no	Hp is the square root of h divided by normalised number of authors and credit to their relative rank on the by-line of the h-core articles. The value varies depending on the method used to share credit between authors and suffers same deficiencies as h.
Adapted pure h	Independence	no	no	no	As Pure h, but uses the square root of author count instead of full author count.
Dynamic h	Rank	yes	yes	no	A dynamic h value that solves the inconsistency of h, but is still a heterogeneous composite indicator.

Price index	Currency	yes	yes	yes	Computes the percentage references to documents, not older than 5 years, at the time of publication of the citing sources
DCI index	Quality	no	no	no	Weights citation count over time, but the weighting parameter causes some authors to arbitrarily gain and others to loose citations. Divides using the logarithm of past time intervals to place value on recently cited papers.
hmx	Rank	yes	no	no	Hmx is the maximum h score in Scopus, WoS and GS. Suffers same deficiencies as h.

Appendix 9

Table 14. Recommended publication indicators

Table 14: Recommended publication indicators								
Indicator	**Calculation**	**Definition**	**Advantages**	**Disadvantages**	**Concept definition**	**Complexity**		**Comments**
						Col.	**Cal**	
P	Sum of publications	Count of production used in formal communication	Potentially, all types of output can be included or selected in regards to theme of evaluation	Does not measure importance, impact of papers, duration or volume of research work	Production of papers published in journals and by academic book publishers	1	1	Counts vary across disciplines due to nature of work and conventions for research communication
Weighted publication count	Weighted score applied to the type of output.	Distinction between different document types	Accounts for importance of different publication types for communication within a field	Has to be designed individual to field as no gold standard.	Production of specific types of publication	1	1	Enables of comparison of like with like
Publications in selected sources	Count of publications in predefined sources	Count of publications in sources defined as important by scholars affiliated institute, field or evaluation committee	Reflects output in sources deemed locally important	Provides a distorted or incomplete picture of production	Production in selected sources	1	2	Provides only a snap shot of productivity
Cognitive orientation	Papers aggregated according to scientific subfields the individual publishes or is cited in	Identify how frequently a scientist publishes or is cited in various fields; indicates visibility/usage in the main subfields and peripheral subfields	Can easily be related to the position a researcher holds in the community	More applicable in some fields than others as often journal based and limited to the database used to source publications definition of scientific fields	Cognitive orientation	3	1	Useful to identify future areas for collaboration and production.

Appendix 10

Table 15. Recommended citation indicators

Table 15 Recommended citation indicators								
Indicator	**Calculation**	**Definition**	**Advantages**	**Disadvantages**	**Concept definition**	**Complexity**		**Comments**
						Col.	**Cal.**	
C	Sum of citations, including self-citations	Indication of number of citations for whole period of analysis	Reflects social side of research and the cumulative development of knowledge	Quality and timeliness of citation not considered	Times cited	3	1	Self-citations affect the reliability & validity of the measure on small amounts of data in assessments
Database dependent citations counting	Number of citations recorded in a specific database	Citation number is dependent on the database used to collect citation information	Indicates how coverage of researcher in database can effect calculation of bibliometric indicators and performance of researcher	Many of the more sophisticated indicators and field benchmarks are reliant on wos and as such cannot be compared with data from other sources	Times cited in specific sources	2	1	Scope, validity, reliability and cost of the citation collection is dependent on choice of citation index
C-sc	Sum of citations, minus self-citations	Measure of external citations	Reflects social side of research and the cumulative development of knowledge	Quality and timeliness of citation not considered; unclear what to exclude: cites of oneself, a co-author or institutional colleague	External citations	3	2	Does not account for older articles being more cited and variation of citation rates between document types and fields
%nc	The number of uncited papers divided by sum of citations/100	Percentage of papers that have not been cited	Can contextualize the number of papers not cited to academic age or used to explain performance on other such as CPP	Does not indicate lack of citation means lack of quality or usefulness. Papers may not be recorded as cited due to citation database indexing policy	Work that has not been cited	1	1	Illustrates the types of publications although important that do not receive citations, i.e. technical reports, guidelines etc.
%sc	The number of self-citations divided by sum of citations/100	Percentage of self-citations	Illustrates how work builds on previous findings. Advertises the work and the author	Unclear what a self-citation is: cites of oneself, a co-author or institutional colleague.	Self-use	2	1	Self-citation is highly variable among individuals and its contribution highly variable. Self-citations are not dismissible when calculating citation statistics

Appendix 11

Table 16: Recommended ALI (hybrid)

Recommended ALI (hybrid indicators)										
Indicator	**Calculation**	**Definition**	**Designed for**	**Citation**	**Measure**	**Advantages**	**Disadvantages**	**Complex**		**Comments**
								Col.	**Cal.**	
c(t) Egghe & (Rousseau 2000)	c(t) is the difference between the date of publication of a researcher's work and the age of citations referring to it	The age of citations referring to a scholars work	Authors who have published and indexed in a citation index	Use over time	Currency	The entire distribution of the citation ages of a set of citing publications provides insight into the level of obsolescence or sustainability	Possibility of measuring aging in a meaningful way is questionable by means of citation counting as this doesn't account for role of literature growth, availability of literature and disciplinary variety	3	3	Usage and validity are not necessarily related
IQP (Antonakis & Lalive 2008)	Calculated as a) A= (mnjs*Pyrs*p+1)/2. (number of citations if author was of average quality for field), b) A/number of papers (to give estimated performance per paper, c) define actual number of citations, d) IQP=actual citations/b+number of papers, e) calculate field impact per papers *number of papers	IQP is the expected average performance of a scholar in the field, amount of papers cited more frequently than average and how much more than average these papers are cited.	Authors with publications in WoS	Quality	Excellence	Corrects citation count for scholarly productivity, author's academic age, and field-specific citation habits with reference to estimated citation rate. Online calculator: http://www.hec.unil.ch/jantonakis/iqp%20calculator%20version%202008.xls	Tested in natural sciences, medicine and psychology and dependent on WOS field specific journal impact factors	**3**	**3**	Correlates better with expert ratings of greatness than h index. Allows comparison as brings papers in low cited fields on same scale as papers in highly cited fields
AWCR (Harzing 2012b)	(Citations to all papers, divided by age of paper)/number of publications	AWCR measures the number of citations to an entire body of work, adjusted for the age of each individual paper	Authors with publications in citation indices	impact	Effect of all papers	Using the sum over all papers instead, represents the impact of the total body of work allowing younger, less cited papers to contribute to the AWCR	Field norm has to be decided to account for field characteristics such as expected age of citations, "sleeping beauties", and delayed recognition	2	3	AWCR offers by default empirical insight into research seniority and career age
Price index (Price 1970)	PI = (n1/n2)*100, where n1 is the number of cited references with a relative age of less than 5 years old, n2 is the total number of references	Percentage references to documents, not older than 5 years, at the time of publication of the citing sources	Author who has published and is cited some form of expression	Relative use over time	Currency	Accounts for the differing levels of immediacy characteristic of the structurally diverse modes of knowledge production occurring in the different sciences	Does not reflect the age structure in slowly ageing literature	**3**	2	In the calculation of PI it is unclear whether the year of publication, is year zero or year one. Moreover, it is unclear whether or not this year is included

References

Aagaard, K. (2015). How incentives trickle down: Local use of a national bibliometric indicator system. *Science and Public Policy*, doi: 10.1093/scipol/scu087.

Abbasi, A., Altmann, J., and Hwang, J. (2010). Evaluating scholars based on their academic collaboration activities: two indices, the RC-index and the CC-index, for quantifying collaboration activities of researchers and scientific communities. *Scientometrics*, 83(1), 1-10.

Abramo, G., D'Angelo, C. A., and Murgia, G. (2014). Variation in research collaboration patterns across academic ranks. *Scientometrics*, 98 (3), 2275-2294.

Aksnes, D. W. (2009). Researchers' perceptions of citations. *Research Policy*, doi:10.1016/j.respol.2009.02.001.

Aksnes, D. W. (2005). *Citations and their use as indicators in science policy. studies of validity and applicability issues with a particular focus on highly cited papers*. Doctoral Thesis. University of Twente.

Aksnes, D. W., Olsen, T. B., and Seglen, P. O. (2000). Validation of Bibliometric Indicators in the Field of Microbiology: A Norwegian Case Study. *Scientometrics*, 49(1), 7-22.

Albarrán, P., Ortunno, I., Ruiz-Castillo, J. (2011) Average-based versus high- and low-impact indicators for the evaluation of scientific distributions. *Research Evaluation,* 20(4), 325-339.

Alonso, S., Caberizoa, F. J., Herrera-Viedmac, E., and Herrerac, F. (2010). Hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. *Scientometrics,* 82(2), 391-400.

Alonso, S., Cabreriazo, F., Herrera-Viedma, E., and Herra, F. (2009). H-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273-289.

Anderson, T. R., Hankin, R. K. S., and Killworth, P. D. (2008). Beyond the durfee square: Enhancing the h-index to score total publication output. *Scientometrics*, 76(3), 577-588.

Antonakis, J., and Lalive, R. (2008). Quantifying scholarly impact: IQP versus the hirsch h. *Journal of the American Society for Information Science and Technology*, 59(6), 956-969.

Archambault, È. & Larivière, V. (2010). The limits of bibliometrics for the analysis of the social sciences and humanities literature. In F.Caillods (Ed.) *World Social Science Report 2010* (pp. 251-254). UNESCO Publishing.

Bach, J. F. (2011). *On the proper use of bibliometrics to evaluate individual researchers*. Académie des sciences. Retrieved 23-6-2015 from: http://www.academie-sciences.fr/activite/rapport/avis170111gb.pdf

Ball, P. (2005). Index aims for fair ranking of scientists. *Nature*, doi:10.1038/436900a

Banerjee, P. (1998). Indicators of "innovation as process". *Scientometrics*, 43(3), 331-357.

Bar-Ilan, J. (2008). Which h-index? — A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2), 257-271.

Barthes, R. (1977). The death of the author (S.Health, Trans.). In R.Barthes (Ed.), *Image, music, text* (pp. 142-148). New York: Hill & Wang.

Bartoli, A. and Medveta, E. (2014). Bibliometric Evaluation of Researchers in the Internet Age. *The Information Society: An International Journal*, 30(5), 349-354.

Batista, P., Campiteli, M., Kinouchi, O., and Martinez, A. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics*, 68(1), 179-189.

Beaver, D. Deb. and Rosen, R. (1978). Studies in scientific collaboration. Part 1. The professional origins of scientific co-authorship. *Scientometrics*, 1(1), 65-84.

Berk, R. A. & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T.G.Blomberg & S. Cohen (Eds.), *Punishment and social control* (pp. 235-254). New York: Walter de Gruyter.

Bertocchi, G., Gambardella, A., Jappelli, T., Nappi, C. A., and Peracchi, F. (2013). *Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy*. CEPR Discussion Paper No.DP9724, November 2013. Retrieved 15-4-2015, from: http://ssrn.com/abstract=2353861

Birnholtz, J. P. (2006). What does it mean to be an author? The intersection of credit, contribution, and collaboration in science. *Journal of the American Society for Information Science and Technology*, 57(13), 1758-1770.

Bishop, D. (2014). *Metricophobia among academics*. BishopBlog. Retrieved 23-6-0015, from: http://deevybee.blogspot.dk/2014_11_01_archive.html

Bishop, M. (2015). *Scholarly publication in Astronomy: evolution or revolution?* International Astronomical Union Focus Meetings (GA). Retrieved 18-5-0015, from: http://www.iau.org/science/events/1130/

Bollen, J., Van de Sompel, H., Hagberg, A., and Chute, R. (2009). A Principal Component Analysis of 39 Scientific Impact Measures. *PLoS ONE*, doi: 10.1371/journal.pone.0006022

Bordons, M. and Gomez, I. (2003). One step further in the production of bibliometric indicators at the micro level: Differences by gender and professional category of scientists. *Scientometrics*, 57(2), 159-173.

Bordons, M., Zulueta, M. A., Cabrero, A., and Barrigon, S. (1995). Research performance at the micro level: Analysis of structure and dynamics of pharmacological research teams. *Research Evaluation*, 5(2), 137-142.

Bornmann, L. (2012). Evaluations by peer review in science. *Science Reviews*, doi: 10.1007/s40362-012-0002-3

Bornmann L. Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. Journal of Informetrics. 2010;4(3):441–443. doi: 10.1016/j.joi.2010.04.004

Bornmann, L. and Daniel, H-D. (2007). Convergent validation of peer review decisions using the h index: Extent of and reasons for type I and type II errors. *Journal of Informetrics*, 1(3), 204-213.

Bornmann, L., Mutz, R., and Daniel, H-D. (2008a). Are there better indices for evaluation purposes than the h-index? A comparison of nine differnt variants of the h-index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830-837.

Bornmann, L., Mutz, R., Neuhaus, C., and Daniel, H-D. (2008b). Citation counts for research evaluation: standards of good practice for analyzing bibliometric data and presenting and interpreting results. *Ethics in Science and Environmental Politics*, 8(1), 93-102.

Bornmann, L. and Mutz, R. (2009). Do We Need thehIndex and Its Variants in Addition to Standard Bibliometric Measures? *Journal of the American Society for Information Science and Technology*, 60(6), 1286-1289.

Bornmann, L. and Werner, M. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 98(1), 487-509.

Bošnjak, L. and Marušić, A. (2012). Prescribed practices of authorship: a review of codes of ethics from professional bodies and journal guidelines across disciplines. *Scientometrics*, 93(3), 751-763.

Boyack, K. W. and Börner, K. (2003). Indicator-assisted evaluation and funding of research: visualizing the influence of grants on the number and citation counts of research papers. *Journal of the American Society for Information Science and Technology*, 54(5), 447-461.

Browman, H. and Stergiou, K (2014). Factors and indices are one thing, deciding who is scholarly, why they are scholarly, and the relative value of their scholarship is something else entirely. *Ethics In Science And Environmental Politics*, 8, 1-3.

Brown, R. (2009). A simple method for excluding self-citations from the h-index: The b-index. *Online Information Review*, 33(6), 1129-1136.

Burrell, Q. L. (2001). Ambiguity and scientometric measurement: a dissenty view. *Journal of the American Society for Information Science and Technology*, 52(12), 1075-1080.

Cabrerizo, F. J., Alonso, S., Herrera-Viedmac, E., & Herrera, F. (2012). Q2-index: Quantitative and qualitative evaluation based on the number and impact of papers in the Hirsch core. *Journal of Informetrics*, 4(1), 23-28.

Cameron, B. D. (2005). Trends in the usage of ISI bibliometric data: Uses, Abuses and Implications. *Libraries and the Academy*, 5(1), 105-125.

Carabone, V. (2011). Fractional counting of Authorship to Quantify Scientific Research Output. *arXiv*:1106.0114 [physics.soc-ph].

Castellani, T. (2014). Epistemological Consequences of Bibliometrics: Insights from the Scientific Community. *Social Epistemology Review and Reply Collective*, 3(11). 1-20.

Cawkell, A. E. (1976). Understanding science by analysing its literature. *The Information Scientist*, 10(1), 3-10.

Chen, D-Z, Lin C-P, Huang, M-H, and Huang, C-Y (2014). Constructing a new patent bibliometric performance measure by using modified citation rate analyses with dynamic backward citation windows. *Scientometrics*, 82(1), 149-163.

Clarke, R. and Pucihar, A. (2012). *The Web of Science Revisited: Is it a Tenable Source for the Information Systems Discipline or for eCommerce Researchers?* Retrieved 11-2-2015, from: http://www.rogerclarke.com/SOS/WoSRev.html

Clarke, R. (2008). *An Exploratory Study of Information Systems Researcher Impact*. Communications of the Association for Information Systems, 22(1): Retrieved 11-2-2015, from: http://www.rogerclarke.com/SOS/Cit-CAIS.html

Claro, J. and Costa, C. A. V. (2011). A made-to-measure indicator for cross-disciplinary bibliometric ranking of researchers performance. *Scientometrics*, 86(1), 113-123.

Cole, J. R. and Cole, S. (1973). Citation Analysis. *Science*, 4120, 28-33.

Cole, S. and Cole, J. R. (1967). Scientific output and recognition: a study in the operation of the reward system in science. *American Sociological Review,* 32(3), 377-390.

Colledge, L. (2014). *Snowball Metrics Recipe Book*. Snowballmetrics.com. Retrieved 11-2-2015, from: http://www.snowballmetrics.com/wp-content/uploads/snowball-metrics-recipe-book-upd.pdf

Collini, S. (2012). Bibliometry. In *What are universities for?* (pp.120-131) London: Penguin.

Costas, R. and Bordons, M. (2007a). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics,* 1(3), 193-203.

Costas, R. and Bordons, M. (2007b). The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. *Journal of Informetrics,* 1(3), 193-203.

Costas, R. and Bordons, M. (2005). Bibliometric indicators at the micro-level: some results in the area of natural resources at the Spanish CSIC. *Research Evaluation*, 14 (2), 110-120.

Costas, R., Bordons, M., van Leeuwen, T. N., and van Raan A.F.J. (2009). Scaling rules in the science system: influence of field-specific citation characteristics on the impact of individual researchers. *Journal of the American Society for Information Science and Technology*, 60(4), 740-775.

Costas, R., van Leeuwen, T. N., and Bordons, M. (2010a). A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level: The Effects of Age on Productivity and Impact. *Journal of the American Society for Information Science and Technology*, 61(8), 1564-1581.

Costas, R., van Leeuwen, T. N., and Bordons, M. (2010b). Self-citations at the meso and individual levels: effects of different calculation methods. *Scientometrics*, 82(3), 517-537.

Costas, R., van Leeuwen, T. N., and van Raan, A. F. J. (2011). The "Mendel Syndrome" in science: Durability of scientific literature and its effects on bibliometric analysis of individual scientists. *Scientometrics*, 89(1), 177-205.

Costas, R., van Leeuwen, T. N., and van Raan, A. F. J. (2010c). Is scientific literature subject to a sell by date? A general methodology to analyze the durability of scientific documents. *Journal of the American Society for Information Science and Technology*, 61(2), 329-339.

Cozzens, S.E. (1989). What do citations count? The Rhetoric First model. *Scientometrics,* 15(5-6), 437-447.

Cozzens, S. E. (1982). Split citation identity. Journal of the American Society for Information Science, 33(4) 233-236.

Cozzens, S. E. (1981). Taking the measure of Science: A review of Citation Theories. *ISSK Newsletter on New Directions in the Sociology of Science,* 7(May 1-2), 16-21.

Cronin, B. (2014). *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact*. Boston: MIT Press Ltd.

Cronin, B. (2000). Semiotics and Evaluative Bibliometrics. *Journal of Documentation,* 56(4) 440-53.

Cronin, B. (1984). *The citation process: The role and significance of citations in scientific communication*. London: Taylor Graham.

Cronin, B. (1981). The Need for a Theory of Citing. *Journal of Documentation*, 37(1), 16-24.

Dahler-Larsen, P. (2012). *The Evaluation Society.* California: Stanford University Press.

Day, R. E. (2014). The data - it's me! ("Les données - cést moi!). In B.Cronin & C. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing multidimensional indicators of scholarly impact* (pp. 67-84). Massachusetts: The MIT Press.

De Battisti, F. and Salini, S. (2012). Robust analysis of bibliometric data. *Statistical Methods and Applications*, 22(2), 269-283.

De Bellis, N. (2014). History and Evolution of (Biblio)metrics. In *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 23-44). Massachusetts: MIT Press.

De Bellis, N. (2009). *Bibliometrics and citation analysis: From the science citation index to cybermetrics*. Lanham, Md: Scarecrow Press.

de Neufville, R. (2010). Instrumentalism. In M.Bevir (Ed.), Encyclopedia of Political Theory: A-E (pp. 702-703). California: Sage Reference.

de Vries, J. (2010). Is New Public Management Really Dead? *OECD Journal on Budgeting*, 2010 (1), 1-5.

Dzombak, R. (2013). Scholarly advances in Humanitarian Engineering and Social Entrepreneurship: a typology of Research Publications. *International Journal for Service Learning in Engineering*, Special Edition, 98-116.

Edge, D. (1979). Quantitative measures of communication in science: a critical review. *History of Science*, 17(36 Pt 2), 102-134.

Egghe, L. (2013). On the correction of the h-index for career length. *Scientometrics*, 96(2) 563-571.

Egghe, L. (2006). Theories and practise of the g-index. *Scientometrics*, 69(1), 131-152.

Egghe, L. & Rousseau, R. (1990). *Introduction to Informetrics: Quantitative Methods in Library, Documentation and Information Science*. Amsterdam: Elsevier.

Egghe, L. and Rousseau, R. (2000). Aging, obsolescence, impact, growth and utilization: Definitions and relations. *Journal of the American Society for Information Science and Technology*, 51(11), 1004-1017.

Egghe, L., Rousseau, R., and Hooydonk, G. (2000). Methods for accrediting publications to authors or countries: Consequences for evaluation studies. *Journal of the American Society for Information Science and Technology*, 51(2), 145-157.

Eloy, J. A., Svider, P., Chandrasekhar, S. S., Husain, Q., Mauro, K. M., Setzen, M. et al. (2013). Gender disparities in scholarly productivity within academic otolaryngology departments. *Otolaryngology - Head and Neck Surgery* (United States), 148(2), 215-222.

Emmeche, C. (2014). *Den Bibliometriske Forskningsindikator - fordele og ulemper*. Faggruppe68. Retrieved 23-6-2015, from: http://faggruppe68.pbworks.com/w/page/6015700/Den%20Bibliometriske%20Forskningsindikator%20-%20fordele%20og%20ulemper

Erikson, M. and Erlandson, P. (2014). Taxonomy of motives to cite. *Social Studies of Science*, 44(4), 625-637.

Fanelli, D. and Ioannidis, J. P. A. (2013). US studies may overestimate effect sizes in softer research. *Proceedings of the National Academy of Sciences of the United States of America*, 110(37), 15031-15036.

Farhadi, F., Salehi, H., Yunus, M. M., Chadegani, A. A., Farhabi, M., and Fooladi, M. (2013). Does it matter which citation tool is used to compare the h-index of a group of highly cited researchers? *Australian Journal of Basic and Applied Sciences*, 7(4), 198-202.

Ferrara, A. and Salini, S. (2012). The challenges in modelling bibliographic data for bibliometric analysis. *Scientometrics*, 93(3), 765-785.

Foucault, M. (1979). What is an Author? In J.V.Harari (Ed.), *Textual strategies: Perspectives in post-structuralist criticism* (pp. 141-160). Ithaca, NY: Cornell University Press.

Franceschet, M. (2009). A cluster analysis of scholar and journal bibliometric indicators. *Journal of the American Society for Information Science and Technology*, 60(10), 1950-1964.

Franceshini, F., Maisano, D., and Mastrogiacomo, L. (2013). The effect of database dirty data on h-index calculation. *Scientometrics*, 95(3), 1179-1188.

Frandsen, T. F. and Nicolaisen, J (2008). Intradisciplinary differences in database coverage and the consequences for bibliometric research. *Journal of the American Society for Information Science*, 59(10), 1570-1581.

Freedman, D. A., Pisani, R., & Purves, R. (2007). *Statistics*. (4 ed.) New York: W.W.Norton & Company.

Furner, J. (2014). The Ethics of Evaluative Bibliometrics. In B.Cronin & C. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 85-107). Massachusetts: MIT Press.

Galam, S. (2011). Tailor based allocations for multiple authorship: a fractional gh-index. *Scientometrics*, 89(1), 365-379.

Garfield, E. (1998). From citation indexes to informetrics: Is the tail now wagging the dog? *Libri*, 48(2), 67-80.

Garfield, E. (1994). The concept of citation indexing: A unique and innovative tool for navigating the research literature. Current Contents. http://wokinfo.com/essays/concept-of-citation-indexing

Garfield, E. (1985). Uses and misuses of citation frequency. *Current Contents*, 43(October 28), 3-9.

Garfield, E. (1979). *Citation Indexing: Its theory and Applicationin Science, Technology and Humanities*. New York: Wiley.

Garfield, E. (1975). The Obliteration Phenomenon in science - and the advantage of being obliterated. *Essays of an Information Scientist*, 2, 396-398.

Garfield, E. (1970). Citation indexing for studying science. *Nature*, 227(5259), 669-671.

Garfield, E. (1964). Science Citation Index - A New Diemnsion in Indexing. *Science*, 144(3619), 649-654.

Gaster, N and Gaster, M (2012). A critical assessment of the h-index. *Bio Essays*, 34(10), 830-832.

Gilbert, G. N. (1977). Referencing as persuasion. *Social Studies of Science*, 7(1), 113-122.

Gingras, Y. (2014). Criteria for evaluating indicators. In B.Cronin & C. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact* (pp. 109-125). Cambridge, Massachusetts: The MIT Press.

Gingras Y, Larivière V. There are neither "king" nor "crown" in scientometrics: Comments on a supposed "alternative" method of normalization. Journal of Informetrics. 2011;5(1):226–227. doi: 10.1016/j.joi.2010.10.005

Glänzel, W. (2003). *Bibliometrics as a research field. A course on theory and applications of bibliometric indicators*. Retrieved 21-5-2015, from: https://www.researchgate.net/publication/242406991_BIBLIOMETRICS_AS_A_RESEARCH_FIELD_A_course_on_theory_and_application_of_bibliometric_indicators

Glänzel, W. (1996). The need for standards in bibliometric research and technology. *Scientometrics*, 35(2), 167-176.

Glänzel, W. and Schoepflin, U. (1999). A Bibliometric Study of Reference Literature in the Sciences and Social Sciences. *Information Processing and Management*, 35(1), 31-44.

Glänzel, W. and Schoepflin, U. (1994). Little Scientometrics, big Scientometrics - and beyond? *Space Med Med Eng* (Beijing), 30(2-3), 375-384.

Glänzel, W. and Schubert, A. (1992). A characterization of scientometric distributions based on harmonic means. *Scientometrics*, 26(1) 81-96.

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. Scientometrics, 87(2), 415–424.

Gläser, J. & Laudel, G. (2007). The Social construction of bibliometric evaluations. In R.Whitely & J. Gläser (Eds.), *The Changing Governence of the Sciences*, (pp.101-123) Netherlands: Springer Science-Business Media B.V.

Goodhart, C. A. E. (1975). Problems of Monetary Management: The U.K. Experience. *Papers in Monetary Economics*, 1975(1).

Gorard, S. (2006). Towards judgement-based statistical analysis. *British Journal of Sociology of Education*, 27(1), 67-80.

Gould, P. (1981). Letting the data speak for themselves. *Annals of the Association of American Geographers*, 71(2), 166-176.

Grant, M. and Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91-108.

Hagen, N. T. (2010). Harmonic publication and citation counting: sharing authorship credit equitably – not equally, geometrically or arithmetically. *Scientometrics*, 84(3), 785-793.

Hagen, N. T. (2008). Harmonic allocation of authorship credit: Source-level correction of bibliometric bias assures accurate publication and citation analysis. *PLoS ONE*, doi: 10.1371/journal.pone.0004021

Harzing, A. W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences? *Scientometrics*, 94(1), 23-34.

Harzing, A. W., Alakangas, S., and Adams, D. (2014). hIa: An individual annual h-index to accommodate disciplinary and career length differences. *Scientometrics*, 99(3), 811-821.

Haustein, S. & Larivière, V. (2015). The Use of Bibliometrics for Assessing Research: Possibilities, Limitations and Adverse Effects. In I.M.Welpe, J. Wollersheim, S. Ringelhan, & M. Osterloh (Eds.), *Incentives and Performance: Governance of reserach organizations* (pp. 121-139). New York: Springer International Publishing.

Herbertz, H (1995). Does it pay to cooperate? A bibliometric case study in molecular biology. *Scientometrics*, 33(1), 117-122.

Hicks, D. (2012). One size doesn't fit all: on the co-evolution of national evaluation systems and social science publishing. *Confero: Essays on Education, Philosophy and Politics*,1(1), 1-21.

Hicks, D. (2004). The Four Literatures of Social Science. In H. Moed, W, Glänzel and U. Schmoch., *Handbook of Quantitative Science and Technology Research.* (pp. 473-496). Netherlands: Springer.

Hicks, D. Wouters, P. Waltman, Ludo. de Rijcke, S., and Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature,* 520(7548), 429-431.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46), 16569-16572.

Hjørland, B. (2006). *Document typology. Core Concepts in Library and Information Science (LIS).* Retrieved 28-5-0015, from: http://www.iva.dk/bh/core%20concepts%20in%20lis/articles%20a-z/document_typology.htm

Holden, G., Rosenberg, G., and Barker, K. (2005). Bibliometrics. *Social Work in Health Care*, 41(3), 4-67.

Hooydonk, G. (1997). Fractional counting of multiauthored publications: Consequences for the impact of authors. *Journal of the American Society for Information Science and Technology*, 48(10), 944-945.

Iivari, J. (2008). Expert evaluation vs bibliometric evaluation: experiences from Finland. *European Journal of Information Systems*, doi:10.1057/ejis.2008.10.

Jasco, P. (2008). The plausibility of computing the h-index of scholary productivity and impact using reference enhanced databases. *Online Information Review,* 32,(2), 266-283.

Jasco, P. (2005a). As we may search: Comparison of major features of the Web of Science, Scopus, and Google Scholar citation-based and citation-enhanced databases. *Current Science*, 89(9), 1537-1547.

Jasco, P. (2005b). Google Scholar: the pros and the cons. *Online Information Review,* 29(2), 208-214.

Jin, B.H., Liang, L.L., Egghe, L. (2007) The R and AR indices: Complementing the h-index. *Chinese Science Bulletin,* 52(6), 855-863.

Južnic, P., Peclin, S., Žaucer, M., Mandelj, T., Pušnik, M., and Demšar, F. (2010). Scientometric indicators: peer-review, bibliometric methods and conflict of interests. *Scientometrics,* 85,(2), 429-441.

Kahneman, D. (2011). The illusion of validity. In *Thinking, fast and slow* (209-222). New York: Farrar, Straus and Giroux.

Katz, J. S. (1996). Bibliometric standards: Personal experience and lessons learned. *Scientometrics,* 35(2), 193-197.

Katz, J. S. & Hicks, D. (1997). How much is a collaboration worth? A calibrated bibliometric model, *Scientometrics*, 40(3), 542-554.

Kermarrec, A. M., Faou, E., Merlet, J. P., Robert, P., & Segoufin, L. (2007). *What do Bibliometric indicators measure?* Analysis document INRIA evaluation committee (Report, Institut National de Recherche en Informatique et en Automatique). Versailles: INRIA Evaluation Committee.

Kitchin, R. (2014). Big Data, new epistemologies and paradigm shifts. *Big Data and Society*, 1(1), 1-12.

Koeing, E. D. (1983). Bibliometric indicators versus expert opinion in assessing research performance. *Journal of the American Society for Information Science*, 34(2), 136-145.

Kosmulski, M. (2009). New seniority-independent Hirsch-type index. *Journal of Informetrics*, 3(4), 341-347.

Kosmulski, M. (2006). A new type Hirsch-index saves time and works equally well as the original h-index. *ISSI Newsletter,* 2(3), 4-6.

Kosmulski, M. (2013). Family-tree of bibliometric indices. *Journal of Informetrics*, 7(2), 313-317.

Kuhn, T. (1970). The Structure of Scientific Revolutions. 2nd Edition. University of Chicago Press: Chicago, p.15.

Lancho-Barrentes, B. and Guerrero, V. P. (2010). What lies behind the averages and significance of citation indicators in different disciplines? *Journal of Information Science*, 36(3), 371-382.

Latour, B. (1987). *Science in Action*. Cambridge, MA: Harvard University Press.

Lazarev, V. S. (1996). On chaos in bibliometric terminology. *Scientometrics*, 35(2), 271-277.

Le Pair, C. (1995). Formal Evaluation Methods: Their Utility and Limitations. *International Forum of Information and Documentation*, 20(2), 16-24.

Leimu, R. and Koricheva, J. (2005). What determines the citation frequency of ecological papers? *TRENDS in Ecology and Evolution*, 20(1), 28-32.

Lepori, B., Reale, E., and Tijssen, R. (2011). Designing indicators for policy decisions: challenges, tensions and good practices: introduction to a special issue. *Research Evaluation*, 20(1), 3-5.

Leydesdorff, L (1998). Theories of citation? *Scientometrics*, 43(1), 5-25.

Leydesdorff, L. (1987). Towards a theory of citation. *Scientometrics*, 12(5-6), 287-291.

Leydesdorff, L. and Bornmann, L. (2014) The Operationalization of "Fields" as WoS Subject Categories (WCs) in Evaluative Bibliometrics: The cases of "Library and Information Science" and "Science & Technology Studies" *Journal of the Association for Information Science and Technology*, DOI: 10.1002/asi.23408

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. Journal of the American Society for Information Science and Technology, 62(2), 217–229

Leydesdorff, L and Opthof, T. (2010a). Normalization, CWTS indicators, and the Leiden Rankings: Differences in citation behavior at the level of fields. *arXiv*:1003.3977 [physics.soc-ph]. Retrieved 24-6-2015, from: http://arxiv.org/abs/1003.3977.

Leydesdorff, L and Opthof, T. (2010b). Scopus' SNIP Indicator. *arXiv*:1006.2895 [cs.DL]. Retrieved 24-6-2015, from: http://arxiv.org/abs/1006.2895.

Leydesdorff, L and Van den Besselaar, P (1997). Scientometrics and communication theory: Towards theoretically informed indicators. *Scientometrics*, 38(1), 155-174.

Lindsey, D. (1978). *The scientific publication system in social science.* San Francisco: Jossey-Bass.

Liu, X. Z. and Fang, H. (2012). Fairly sharing the credit of multi-authored papers and its application in the modification of h-index and g-index. *Scientometrics*, 91(1), 37-49.

Logan, E. (1991). A bibliometric analysis of collaboration in a medical specialty. *Scientometrics*, 20(3), 417-426.

Lopez-Cozar, E. D., Rbinson-Garcia, N., and Torres-Salinas, D. (2014). The Google Scholar experiment: How to index false papers and manipulate bibliometric indicators. *Journal of the Association for Information Science and Technology*, 65(3), 446-454.

Lundberg, J. (2007) Lifting the crown-citation score. *Journal of Informetrics*, 1(2), 145-154.

Luukkonen, T (1997). Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics*, 38(1), 27-37.

MacRoberts, M. H. and MacRoberts, B. R. (1996). Problems of citation analysis. *Scientometrics*, 36(3), 435-444.

Marchant, T. (2009). Score-based bibliometric rankings of authors. *Journal of the American Society for Information Science and Technology*, 60(6), 1132-1137.

Martin, B.R. (2013). Whither research integrity? Plagiarism, self-plagiarism and coercive citation in an age of research assessment. *Research Policy*, 42(5), 1005-1014.

Martin, B. R. (1996). The use of multiple indicators in the assessment of basic research. *Scientometrics*, 36(3), 343-362.

Martin, B. R. and Irvine, J. (1983). Assessing Basic Research: Some partial indicators of scientific progress in Radio Astronomy. *Research Policy*, 12(2), 61-90.

Martyn, J. (1964). Bibliographic Coupling. *Journal of Documentation*, 20(4), 236.

Meho, L. and Rogers, Y. (2008). Citation Counting, Citation Ranking, and h-Index of Human-Computer Interaction Researchers: A Comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, 59(11), 1711-1726.

Mendez, A., Gomez, I., and Bordons, M. (1993). Some indicators for assessing research performance without citations. *Scientometrics*, 26(1), 157-167.

Merton, R.K. (1977) The sociology of science: an episodic memoir. In: Merton RK, Gaston J (eds). *The sociology if science in Europe*. Southern Illinois University Press, Carbonale, pp 3-141.

Merton, R. K. (1973). *The normative structure of science*. Chicago: University of Chicago Press.

Miller, C.W. (2006). Superiority of the h-index over the impact factor in physics. arXiv:physics/0608183 [physics.soc-ph]

Minasny, B., Hartemink, A. E., McBratney, A., and Jang, H-J. (2013). Citations and the h index of soil researchers and journals in the Web of Science, Scopus, and Google Scholar. *PeerJ* 1:e183 doi: https://dx.doi.org/10.7717/peerj.183.

Miquel, J. F. (1994). Little Scientometrics and big Scientometrics…… and beyond Scientometrics. *Scientometrics*, 30(2-3), 443-445.

Moed, H. (2010) Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, 4(3):265–277. doi: 10.1016/j.joi.2010.01.002

Moed, H. F. (2005). *Citation analysis in research evaluation*. New York: Springer

Moed, H. (1989). Bibliometric measurement of research performance and Price's theory of differences among the sciences. *Scientometrics*, 15(5), 473-483.

Moed, H. (1985a). A comparative study of bibliometric past performance analysis and peer judgement. *Scientometrics*, 8(3-4), 149-159.

Moed, H. (1985b). The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics*, 8(3), 177-203.

Namazi, M. R. and Fallahzadeh, M. K. (2010). n-index: a novel and easily-calculable parameter for comparison of researchers working in different scientific fields. *Indian Journal of Dermatology, Venereology and Leprology*, 76(3), 229-230.

Nederhof, A. J. and van Raan, A. F. J (1987a). Peer review and bibliometric indicators of scientific performance: A comparison of cum laude doctorates with ordinary doctorates in physics. *Scientometrics*, 11(5), 6-333.

Nederhof, A.J., Zwaan, R. A., De Bruin, R. E., and Dekker, P. J. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: A comparative study. *Scientometrics*, 15(5-6), 423-435.

Nicolaisen, J. (2004). *Social behavior and scientific practice - missing pieces of the citation puzzle.* Doctoral Thesis. Department of Information Studies, Royal School of Library and Information Science.

Nørretranders, T. (2007). *Civilisation 2.0: Miljø, fællesskab og verdensbillede i linkenes tidsalder.* Frederiksberg: Thaning & Appel.

Noruzi, A. (2005). Google Scholar : the new generation of citation indexes. *Libri*, 55(4), 170-180.

Ortega, J. L. (2015). Relationship between altmetric and bibliometric indicators across academic social sites: The case of CSIC's members. *Journal of Informetrics*, 9(1), 39-49.

Panaretos, J and Malesios, C. C. (2014). Assessing scientific research performance and impact with single indices. *Scientometrics*, 81(3), 635-670.

Patel, V. M., Ashrafian, H., Almoudaris, A., Makanjuola, H., Bucciarelli-Ducci, C., and Darzi, A. (2013). Measuring academic research performance for healthcare researchers with the H index: Which search tool should be used? *Medical Principles and Practice*, 22(2), 178-183.

Pendlebury, D. A. (2008). *Using Bibliometrics in Evaluating Research.* Thomsen Reuters, Retrieved 24-6-2015, from: http://wokinfo.com/media/mtrp/UsingBibliometricsinEval_WP.pdf

Peters, H. P. F. and van Raan A.F.J. (1994). A bibliometric profile of top-scientists - a case-study in chemical-engineering. *Scientometrics*, 29(1), 115-136.

Pillay, A. (2013). Academic promotion and the h-index. *Journal of the American Society for Information Science*, 64(12), 2598-2599.

Plomp, R. (1994). The highly cited papers of professors as an indicator of a research groups scientific performance. *Scientometrics*, 29(3), 377-393.

Plume, A. and van Weijen, D. (2014). Publish or Perish? The rise of the fractional author. *Research Trends*, Retrieved 21-5-2015, from: http://www.researchtrends.com/issue-38-september-2014/publish-or-perish-the-rise-of-the-fractional-author/

Porter, A. L., Chubin.D.E., and Jin, X-Y. (1988). Citations and scientific progress: Comparing bibliometric measures with scientist judgments. *Scientometrics*, 13(3-4), 103-104.

Porway, J. (2013). *You can't just hack your way to social change*. Harvard Business Review Blog. Retrieved 16-6-2015, from: https://hbr.org/2013/03/you-cant-just-hack-your-way-to/

Prathap, G. (2012). The Inconsistency of the H-Index. *Journal of the American Society for Information Science and Technology*, (63)7, 1480-1481.

Preece, A. (2011). Evaluation verification and validation methods in Knowedge Engineering. In Rajkumar Roy (Ed.), *Industrial Knowledge Mangaement* (pp. 91-104). London: Springer.

Price, D. d. S. (1981). Multiple authorship. *Science,* 212 (4498), 986.

Price, D. d. S. (1970). Citation measures of hard science, soft science, technology and non-science. In C.E.Nelson & D.K.Pollack (Eds.), *Communication among scientists and engineers* (pp. 3-22). Lexington: Heath Lexington Books.

Pritchard, A. (1969). Statistical bibliography or bibliometrics? *Journal of Documentation*, 25(4), 348-349.

Putnam, H. (1979). What theories are not. In H.Putnam (Ed.), *Mathematics, Matter and Method* (2 ed.), England: Cambridge University Press.

Radicchi, F., Fortunato, S., and Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, 105(45), 17268-17272.

Ravetz, J. R. (1971). *Scientific knowledge and its social problems.* Harmondsworth.

Ravichandra Rao, I. K. (1996). Methodological and conceptual questions of bibliometric standards. *Scientometrics*, 35(2), 265-270.

Reed, K. L (1995). Citation analysis of faculty publication: beyond Science Citation Index and Social Science Citation Index. *Bulletin of the Medical Library Association,* 83 (4), 503-508.

Riviera, E. (2012) Mapping scientific literature: structuring scientific communities through Scientometrics. Doctoral thesis. Università degli Studi di Milano Biocca. Milano.

Rousseau, R (1998). Citation analysis as a theory of friction of polluted air? *Scientometrics*, 43(1), 63-67.

Rowlands, I. (2003). Knowledge production, consumption and impact: policy indicators for a changing world. *ASLIB proceedings*, 55(1/2), 5-12.

Rubem, A., de Moura, A., and Soares de Mello, J. (2015). Comparative analysis of some individual bibliometric indices when applied to groups of researchers. *Scientometrics*, 102(1), 1091-1035.

Russell, J. M. (1994). Back to the future for informetrics . *Scientometrics*, 30(2-3), 407-410.

Russell, J. M. & Rousseau, R. (2002). Bibliometrics and institutional evaluation. In *Arvantis (Ed.), Encyclopedia of Life Support Systems (EOLSS) Part 19.3 Science and Technology Policy*, Oxford: Eolss Publishers.

Sandström, E. & Sandström, U. (2009). Meeting the micro-level challenges: bibliometrics at the individual level. In B. Larsen & J. Leta (Eds.), *12th Conference on Scientometrics and Informetrics*, July 14-17, 2009, Rio de Janeiro, Brazil, BIEREME/PAHO/WHO (pp. 846-856),

Sandström, U. & Hällsten, M. (2007). Gender; funding diversity and quality of research. In Torres-Salinas, D. & Moed, H. Proceedings of ISSI 2007: *11th International Conference of the International Society for Scientometrics and Informetrics*, 2009, (pp. 685-690).

Schmoch, J. (1997). Indicators and the relations between science and technology. *Scientometrics*, 38(1), 103-116.

Schmoch, U., Schubert, T., Jansen, D., Heidler, R., and von Görtz, R. (2010). How to use indicators to measure scientific performance: a balanced approach. *Research Evaluation*, 19,1 2-18.

Schneider, J. W. (2015). Null hypothesis significance tests. A mix-up of two different theories: The basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1), 411-432.

Schneider, J. W. (2014). Null hypothesis significance tests: a mix up of two different theories, the basis for widespread confusion and numerous misinterpretations. *Scientometrics*, 102(1), 411-432.

Schneider, J. W. (2013a). Caveats for using statistical significance tests in ersearch assessments. *Journal of Informetrics*, 7(1), 50-62.

Schneider, J. W. (2013b). Caveats for using statistical significance tests in ersearch assessments. *Journal of Informetrics*, 7(1), 50-62.

Schneider, J. W. & Aagaard, K. (2012). Stor ståhej for ingenting - den danske bibliometriske indikator. In K.Aagaard & N. Mejlgaard (Eds.), *Dansk Forskningspolitik efter årstusindsskiftet* (pp. 187-213). Århus: Århus Universitetsforlag.

Schoepflin, U. and Glänzel, Wolfgang (2001). Two decades of Scientometrics: An interdisciplinary field represented by its leading journal Scientometrics. *Scientometrics*, 50(2), 301-312.

Schreiber, M. (2013). Do we need the g-index? *Journal of the American Society for Information Science and Technology*, 64(11), 2396-2399.

Schreiber, M., Malesios, C. C., and Psarakis, S. (2012). Exploratory factor analysis for the Hirsch index, 17 h-type variants, and some traditional bibliometric indicators. *Journal of Informetrics*, 6(3), 347-358.

Schreiber, M. (2008). An empirical investigation of the g-index for 26 physicists in comparison with the h-index, the A-index, and the R-index. *Journal of the American Society for Information Science and Technology*, 59(9), 1513-1522.

Seglen, P. O. (1997). Why the Impact Factor of Journals Should not be Used for Evaluating Research. *British Medical Journal*, 314(7079), 498-502.

Seglen, P. O. (1996). Bruk av siteringer og tidsskriftimpaktfaktor til forskningsevaluering. *Biblioteksarbejde*, 48 (årgang 17), 27-34.

Sen, B. K. (1997). Mega-authorship from a bibliometric point of view. *Malaysian Journal of Library and Information Science*, 2(2), 9-18.

Sidiropoulos, A., Katsaros, D., and Manopoulos, Y. (2007). Generalized hirsh h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2), 253-280.

Singleton, A. (1976). Journal ranking and selection: a review in physics. *Journal of Documentation*, 32(4), 258-289.

Sivertsen, G. (2009). *A Bibliometric Funding Model based on a National Research Information System*. Norwegian Institute for Studies in Innovation, Research and Education. Retrieved 23-6-0015, from: http://www.issi2009.org/agendas/issiprogram/public/documents/ISSI%202009%20Sivertsen%20Vista-094456.pdf

Skupin, A. (2009). Discrete and continuous conceptualizations of science: Implications for knowledge domain visualization. *Journal of Informetrics*, 3(3), 233-245.

Small, H. G. (1987). The significance of bibliographic references. *Scientometrics*, 12(5-6), 339-341.

Small, H. G. (1978). Cited documents as concept symbols. *Social Studies of Science*, 8(3), 327-340.

Starbuck, W. H. (2006). *The production of knowledge: The challenge of social science research.* England: Oxford University Press.

Suppe, F. (2015). The structure of a scientific paper. *Philosophy of Science*, 65(3), 381-405.

Tinkler, J. (2011). *Maximizing the impacts of your research: a handbook for social scientists*. LSE Public Policy Group. Retrieved 21-4-2015, from: http://www.lse.ac.uk/government/research/resgroups/LSEPublicPolicy/Docs/LSE_Impact_Handbook_April_2011.pdf

Tol, R. S. J. (2009). Of the H-index and its alternatives: An application to the 100 most prolific economists. *Scientometrics*, 80(2), 317-324.

TS (2012). *The Thomson Reuters Journal Selection Process*. Thomson Reuters. Retrieved 11-2-2015, from: http://wokinfo.com/essays/journal-selection-process/

UFM (2015). *Basismidler efter kvalitet*. Retrieved 22-6-2015, from: http://ufm.dk/uddannelse-og-institutioner/videregaende-uddannelse/universiteter/okonomi/basismidler-efter-kvalitet

van Eck, N. J. and Waltman, Ludo (2008). Generalizing the g- and h-indices. *ECON Papers*. Retrieved 12-4-2015, from: http://EconPapers.repec.org/RePEc:ems:eureri:13210

van Leeuwen, T. N. (2014). Testing the validity of the Hirsch-index for research assessment purposes. *Research Evaluation,* 17(2), 157-160.

van Leeuwen, T. N. (2006). The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. *Scientometrics*, 66(1), 133-154.

van Leeuwen, T. N. (2005). Descriptive versus evaluative bibliometrics. In H.Moed, W. Glänzel, & U. Schmoch (Eds.), *The Handbook of Quantitative Science and Technology Research: the use of publication and patent statistics in studies on S&T systems* (pp. 373-389). Dordrecht: Kluwer Academic Publishers.

van Leeuwen, T. N., van der Wurff, L. J., and van Raan A.F.J. (2001). The use of combined bibliometric methods in research funding policy. *Research Evaluation,* 10(3), 195-201.

van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment of 147 chemistry research groups. *Scientometrics*, 67(3), 491-502.

van Raan A.F.J (1998). The influence of international collaboration on the impact of research results. *Scientometrics*, 42(3), 423-428.

van Raan, A. F. J. (1997). For Your Citations Only? Hot Topics in Bibliometric Analysis. *Measurement*, 3(1), 50-62.

van Raan, A. F. J. (1996). Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises. *Scientometrics*, 36(3), 397-420.

Vanclay, J. K. (2015). On the robustness of the h-index. *Journal of the American Society for Information Science and Technology*, 58(10), 1547-1550.

Vieira, E. S., Cabral, J. A. S., and Gomes, J. A. N. F (2014). How good is a model based on bibliometric indicators in predicting the final decisions made by peers? *Journal of Informetrics*, 8(2), 390-405.

Vieira, E. S. and Gomes, J. A. N. F (2011). An impact indicator for researchers. *Scientometrics*, 89(2), 607-629.

Vinkler, P. (2007). Eminence of scientists in the light of the h-index and other scientometric indicators. *Journal of Information Science*, 33(4), 481-491.

Vinkler, P. (1996). Some practical aspects of the standardization of scientometric indicators. *Scientometrics*, 35(2), 235-245.

Waltman, Ludo and van Eck, N. J. (2013). Source normalized indicators of citation impact: an overview of different approaches and an empirical comparison. *Scientometrics*, 96(3), 699-716.

Waltman, L. and van Eck, N. J. (2012). The inconsistency of the h-index. *Journal of the American Society for Information Science and Technology*, 63(2), 406-415.

Waltman, L. and van Eck, N. J. (2009). *A Taxonomy of Bibliometric Performance Indicators Based on the Property of Consistency*. ERIM Report. Retrieved 20-1-2015, from: repub.eur.nl/res/pub/15182/ERS-2009-014-LIS.pdf

Wan, J., Hua, P., and Rousseau, R. (2007). Calculating an author's h- index by taking co-authors into account. *ELIS*. Retrieved 20-4-2015, from: http://eprints.rclis.org/10376/

Watt, J. H. & van den Berg, S. A. (1995). Elements of scientific theories: concepts and definitions. In *Research methods for communication science* (pp. 11-22). England: Allyn and Bacon.

Weingart, P. (2005). Impact of bibliometrics on the science system: Inadvertent consequences? *Scientometrics*, 62(1), 117-131.

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.

White, H. D. (2004). Reward, persuasion, and the Sokal hoax: A study in citation identities. *Scientometrics*, 60(1), 93-120.

White, H. D. (2001). Authors as citers over time. *Journal of the American Society for Information Science and Technology*, 52(2), 87-108.

White, H.D. (1990) Author co-citation analysis: Overview and defense. In: Borgman, C.L. (ed). *Scholarly Communication and Bibliometrics,* 84-106. Newbury Park: Sage.

Whitely, R. (2007). Changing Governance of the Public Science: The Consequences of Establishing Research Evaluation Systems for Knowledge Production in Different Countries and Scientific Fields. In R.Whitely & J. Gläser (Eds.), *The Changing Governance of the Sciences: The Advent of Research Evaluation Systems, Sociology of Sciences* (pp. 3-27). The Netherlands: Springer.

Wiberley Jr, S. E. (2003). A methodological approach to developing bibliometric models of types of humanities scholarship. *The Library Quarterly*, 73(2), 121-159.

Wildgaard, L (2015). A critical cluster analysis of 44 indicators of author-level performance. *arXiv.* Retrieved 18-5-0015, from: http://arxiv.org/abs/1505.04565

Wildgaard, L, Schneider, J. W, and Larsen, B (2014). A review of the characteristics of 108 ALI. *Scientometrics*, 101(1), 125-158.

Wilsdon, J. (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

Wouters, P. (2014a). *Semiotics and Citations*. Unpublished manuscript.

Wouters, P. (2014b). *A key challenge: the evaluation gap.* Retrieved 29-1-2015b, from: https://citationculture.wordpress.com/2014/08/28/a-key-challenge-the-evaluation-gap/

Wouters, P (1999). Beyond the holy grail: from citation theory to indicator theories. *Scientometrics*, 44(3), 561-580.

Wouters, P. (1999b). The citation culture. Universiteit van Amsterdam.

Wu, Q. (2010). The w-index: A measure to assess scientific impact by focusing on widely cited papers. *Journal of the American Society for Information Science and Technology*, 61(3), 609-614.

Xia, X., Li, M., and Xiao, C. F. (1999). Author analysis of papers published in "Space Medicine & Medical Engineering" from 1988 to 1998. *Space Med Med Eng* (Beijing), 12(6), 431-435.

Ye, F (2012). H-inconsistency is not an issue in dynamical systems. *ISSI Newsletter*, 8(2), 22-24.

Ye, F (2011). A Theoretical approach to the unification of informetric models by wave-heat equations. *Journal of the American Society for Information Science and Technology*, 62(6), 1208-1211.

Zahedi, Z., Costas, R., and Wouters, R. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of 'alternative metrics' in scientific publications. *Scientometrics*, 101(2), 1491-1513.

Zhang, C.-T. (2009). The e-index, complementing the h-index for excess citations. *PLoS ONE* 4(5): e5429.

Ziemski, S. (1975). The typology of scientific research. *Zeitschrift für allgemeine Wissenschaftstheorie*, 6(2), 276-291.

Zitt, M (2005). Facing diversity of science: A challenge for bibliometric indicators. *Measurement,* 3(1), 38-49.

Zuckerman, H. (1987). Citation analysis and the complex problem of intellectual influence. Scientometrics, 12(5), 329-33.