

UNIVERSITY OF KENT
Kent Business School

Research topic:

**Novel approaches for evaluation of
research impact**

A thesis submitted to the University of Kent for the degree of Doctor of
Philosophy in Management Science

Doctoral Student:

Tianhui Gong

Supervisors:

Professor Shaomin Wu

Professor Wenbin Liu

Dr Zhen Zhu

January 2025

Acknowledgements

I would like to express my deepest gratitude and appreciation to my supervisors, Professor Shaomin Wu, Professor Wenbin Liu, Dr Zhen Zhu, and Professor John Mingers. Their invaluable guidance and encouragement have played a significant role in completing this thesis.

I am profoundly grateful to Professor Wenbin Liu, my former first supervisor, who continued to guide me even after he left the University of Kent. We met in China to discuss up-to-date work. During the data collection period, I studied on a beautiful campus in Zhuhai, China, and I could meet Professor Wenbin Liu regularly, just like we did at Kent University. His generous support and patient mentorship have been indispensable throughout my PhD journey. At the same time, I extend my heartfelt thanks to Professor Shaomin Wu, my current primary supervisor, whose unwavering support and insightful advice have been instrumental in bringing this thesis to its current form.

I wish to extend special thanks to my research colleagues and office mates, both in Canterbury and Zhuhai: Dr Guanjiang Wan, Miss Junru Ren, Mr Xinyuan Li, Dr Ken Cheng, Dr Xiaodong Feng, Mr Fuyuan Zhu, and Miss Luyang Liu. Their assistance and camaraderie have been invaluable, and I have greatly cherished the time we spent together.

Finally and most importantly, I express my sincere gratitude and thanks to my family for their constant love and encouragement. To my parents, Mrs Yang and Mr Gong, who have supported me in every possible way throughout my life and helped me overcome tough periods; to my mother-in-law, Mrs Li, who supported me and helped care for my baby; to my uncle, Professor Gong, who always encouraged me and cheered me up; to my husband, Dr Wang, who suggested several research advice and shared childcare and housework during my busy time; and to my adorable daughter, Hongyueh, who is the most important person in my life and also a big challenge during my PhD journey, I love you forever.

Thank you all!

Abstract

Research evaluation plays a vital role in the management of science and technology. This thesis is motivated by the fact that manipulation of citation data is becoming a common issue in scientific research evaluation. Thus, we aim to utilise multiple data resources, including scientometric and altmetric data, to develop evidence-based quantitative evaluating methods for research impact to increase the reliability of assessment results.

To achieve the research goals, we first examined various altmetric data sources and platforms through a comprehensive literature review and an empirical study to investigate the quality of altmetric data. We suggested ways to utilise altmetric data in research impact evaluation studies. Then, we developed an evidence-chain-based evaluation method for research impact. We explored the methods for evaluating the research impact of research publications. We analysed the impact generation cause-effect chains, which put together the academic and broader impacts (economy, society, culture, public policy or services, health, and more) of the research output as a final effect, and illustrated the cause-effect chains of the impact generation process across different logical stages. We collected a range of evidence data as indicators of the impact generation process that were carefully selected from the most representative database within their category, and quality was verified by small sample tests, clarifying the origins of data sources and the breadth and comprehensiveness of the database's publication coverage. We utilised the structural equation modelling approach to establish the research impact aggregation models. Finally, all observed publications' impact scores were calculated using the factor weights from disciplinary structural models for research impact.

This research offers several significant contributions to the evaluation of research impact. Firstly, it proposes an innovative quantitative method based on evidence chains, effectively addressing the limitations of traditional results-driven evaluation systems and yielding more reliable insights. Secondly, it pioneers a full-chain evaluation method that captures both

academic and broader dimensions of impact by incorporating every stage of the impact generation process, providing a novel perspective on impact assessment. Thirdly, the study develops a new way of thinking that employs cause-effect chains to analyse research impact, improving attribution and consistency. Additionally, an empirical analysis of altmetrics platforms offers a detailed understanding of altmetric data quality, reliability, and disciplinary differences, serving as a valuable reference for future use of the altmetric data. Lastly, while recognising the limitations of altmetrics as direct tools for impact evaluation, this research highlights their role in deriving meaningful indicators that propose a fresh approach to their application in research impact evaluations.

Keywords: research evaluation; research impact evaluation; evidence-chain-based evaluation; impact generation process; altmetric data; altmetrics; SEM;

Contents

Acknowledgements	i
Abstract	ii
List of Tables.....	viii
List of Figures	x
List of Abbreviations.....	xi
Publication	xiii
Chapter 1 - Introduction.....	1
1.1 Research motivation	1
1.2 Research questions	4
1.3 Research contributions	4
1.4 Thesis structure	6
Chapter 2 - Literature review on research evaluation	8
2.1 Rise of research evaluation.....	8
2.2 Development of research evaluation	10
2.2.1 Definitions of research evaluation	10
2.2.2 Value orientation of research evaluation	11
2.2.3 Methodologies	12
2.2.4 Frameworks	14
2.3 Evaluation of research impact	17
2.4 Summary.....	20
Chapter 3 - Literature review on traditional metrics and altmetrics for research impact	21
3.1 Introduction.....	21
3.2 Traditional metrics for academic impact evaluation	21
3.2.1 The classic citation indicators	22
3.2.2 The Hirsch-index (<i>h</i> -index)	23
3.2.3 The g-index	25
3.2.4 The h-dependent indicators.....	26
3.3 Indicators reflecting academic impact of journals	28
3.4 Altmetrics: the emerging metrics for impact evaluation	31
3.4.1 Emergence of altmetrics	31

3.4.2 Definition of altmetrics	35
3.4.3 Advantages of altmetrics	37
3.4.4 Disadvantages of altmetrics and research challenges	41
3.4.5 Difference between altmetrics and traditional metrics	43
3.5 Altmetrics platforms and data sources	46
3.5.1 Terms and classification	46
3.5.2 Altmetrics aggregators	47
3.5.3 Data sources used in the altmetrics	53
3.6 Some of the critical topics in altmetrics study	58
3.6.1 Cross-metric validation	58
3.6.2 Data quality validation	61
3.6.3 Applications of altmetrics	66
3.7 Knowledge gaps identified	69
3.8 Summary	71
Chapter 4 – Investigation of the altmetric data quality with Altmetric and PlumX as examples	73
4.1 Introduction	73
4.2 Related studies	74
4.3 Altmetrics platforms	76
4.4 Data collection	78
4.5 Coverage analysis	80
4.5.1 Publication coverage	80
4.5.2 Altmetrics coverage	81
4.5.3 Disciplinary differences of coverage	85
4.6 Correlation analysis	87
4.7 Data consistency and heterogeneity	90
(1) CiteULike	92
(2) Blogs	93
(3) News	94
(4) Twitter	95
(5) Facebook	96
(6) Wikipedia	97
(7) Mendeley Reader Count	97

4.8 Recommendation on the use of altmetrics	101
4.9 Summary.....	102
Chapter 5 - An evidence-chain-based approach to evaluating scholarly impact.....	106
5.1 Introduction.....	106
5.2 Research background	107
5.3 Evidence-chain-based quantitative evaluation method	109
5.3.1 Evidence chain and cause-effect chain	109
5.3.2 How to build an evidence chain?.....	112
5.3.3 Primary evaluation models based on the evidence chain	112
5.4 Empirical study - A case study of the impact evaluation of Journal A and Journal B	114
5.4.1 Construct the cause-effect chain and evidence chain for academic impact	114
5.4.2 Evidence data collecting and processing	116
5.4.3 Evaluation results and analysis.....	118
5.5 Limitations of the study	121
5.6 Summary.....	121
Chapter 6 - Evidence chain of impact of research publications.....	124
6.1 Introduction.....	124
6.2 The cause-effect chains of impact of research publications.....	126
6.2.1 Stage 1 - Knowledge generation (S1).....	128
6.2.2 Stage 2 - Knowledge publishing (S2).....	130
6.2.3 Stage 3 - First wave of knowledge dissemination (S3).....	131
6.2.4 Stage 4 - Knowledge understanding and extensive dissemination (S4).....	133
6.2.5 Stage 5 - Knowledge transformation and regeneration (S5)	136
6.3 The evidence data of measurement in impact evaluation	141
6.3.1 Measurement of impact generation process	141
6.3.2 Measurement of different dimensions of impact	145
6.4 Summary.....	146
Chapter 7 - A case study of comprehensive impact evaluation for research publications	148
7.1 Introduction.....	148
7.2 Database screening and data quality verification.....	148
7.2.1 Measurement data in S1	149
7.2.2 Measurement data in S2	153

7.2.3 Measurement data in S3	158
7.2.4 Measurement data in S4	160
7.2.5 Measurement data in S5	160
7.3 Data Collection	176
7.4 Methods on impact evaluation modelling	178
7.4.1 Model establishment	178
7.4.2 Model Identification.....	182
7.5 Model analysis and discussions	184
7.5.1 Structural model analysis.....	184
7.5.2 Disciplinary differences.....	188
7.6 The evidence-chain-based research impact evaluation.....	193
7.7 Summary.....	202
Chapter 8 – Conclusions and future research	205
8.1 Conclusions	205
8.2 Research limitations.....	209
8.3 Future research work	210
References	211
Appendix 1: The comprehensive schematic diagram of the generation of research impact for publications (large print).....	236

List of Tables

<i>Table 2.1 Brief summary of major methodologies.....</i>	<i>12</i>
<i>Table 3.1 Term interpretations in altmetrics (Haustein et al., 2016).....</i>	<i>46</i>
<i>Table 4.1 Altmetric data and sources studied.....</i>	<i>79</i>
<i>Table 4.2 Altmetrics Data Coverage Ratio and Overlapping Ratio on Altmetric and PlumX.....</i>	<i>82</i>
<i>Table 4.3 Altmetrics coverage between different disciplines.....</i>	<i>86</i>
<i>Table 4.4 Intra-correlation of data sources on Altmetric.....</i>	<i>88</i>
<i>Table 4.5 Intra-correlation of data sources on PlumX</i>	<i>88</i>
<i>Table 4.6 Cross-correlation of data sources on Altmetric and PlumX.....</i>	<i>90</i>
<i>Table 4.7 Statistics of overlapping data of each data source</i>	<i>91</i>
<i>Table 4.8 Statistics of Wilcoxon Signed Ranks Test</i>	<i>92</i>
<i>Table 4.9 Statistics of overlapping Mendeley data.....</i>	<i>99</i>
<i>Table 4.10 Non-parametric correlation coefficients</i>	<i>99</i>
<i>Table 4.11 Test Statistics^a</i>	<i>99</i>
<i>Table 4.12 Descriptive statistics of natural logarithm of Mendeley reader data</i>	<i>100</i>
<i>Table 5.1 Data information of Journal A and B.....</i>	<i>118</i>
<i>Table 5.2 Academic impact measurement results based on evidence chain for Journal A and Journal B (Scenario 1)</i>	<i>120</i>
<i>Table 5.3 Academic impact measurement results based on evidence chain for Journal A and Journal B (Scenario 2)</i>	<i>120</i>
<i>Table 6.1 Measurements of the impact generation process.....</i>	<i>141</i>
<i>Table 6.2 Examples of impact indicators</i>	<i>145</i>
<i>Table 7.1 Descriptive statistics of grant data sample</i>	<i>153</i>
<i>Table 7.2 Correlations of grant data sample</i>	<i>153</i>
<i>Table 7.3 Examples of leading preprint repositories.....</i>	<i>155</i>
<i>Table 7.4 Matching results of preprint repositories in Dimensions</i>	<i>157</i>
<i>Table 7.5 Leading database for comprehensive patent data</i>	<i>165</i>
<i>Table 7.6 Descriptive statistics of patent citations for Lens and Dimensions</i>	<i>170</i>
<i>Table 7.7 Correlations of patent citations for Lens and Dimensions</i>	<i>170</i>
<i>Table 7.8 Descriptive statistics of clinical citations for Dimensions and PlumX.....</i>	<i>172</i>
<i>Table 7.9 Descriptive statistics of policy citations for Dimensions, Altmetric and Overton.....</i>	<i>175</i>
<i>Table 7.10 Correlations of policy citations for Dimensions, Altmetric and Overton</i>	<i>175</i>
<i>Table 7.11 Data used in the evaluation of impact of research publications.....</i>	<i>177</i>
<i>Table 7.12 Standardised regression coefficients and significance level</i>	<i>187</i>
<i>Table 7.13 The squared multiple correlations on latent variables.....</i>	<i>188</i>
<i>Table 7.14 Standardised regression coefficients (β) and significance level for different group</i>	<i>189</i>
<i>Table 7.15 Standardised regression coefficients (λ) for different category group.....</i>	<i>191</i>
<i>Table 7.16 Factor score weights for all data.....</i>	<i>194</i>
<i>Table 7.17 Top 10 highest-impact research papers published in 2013.....</i>	<i>198</i>

<i>Table 7.18 Top 20 highest-impact journals published in 2013 in Operations Research & Management Science.....</i>	<i>199</i>
<i>Table 7.19 Factor Score Weights for the Engineering group</i>	<i>200</i>

List of Figures

<i>Figure 4.1 Altmetrics Coverage on Altmetric and PlumX</i>	<i>82</i>
<i>Figure 4.2 The schematic diagram for dominant coverage of the data source</i>	<i>84</i>
<i>Figure 4.3 The schematic diagram for different coverage of the data sources</i>	<i>84</i>
<i>Figure 4.4 The schematic diagram for similar coverage of the data sources</i>	<i>85</i>
<i>Figure 4.5 Histogram of Natural Logarithm of Mendeley Reader data</i>	<i>101</i>
<i>Figure 4.6 Q-Q Plot of Natural Logarithm of Mendeley Reader Data</i>	<i>101</i>
<i>Figure 5.1 The schematic diagram of causality and relationship between cause-effect chains</i>	<i>110</i>
<i>Figure 5.2 The schematic diagram of relationship between cause-effect chains and evidence chains</i>	<i>111</i>
<i>Figure 5.3 The schematic diagram of cause-effect chains and evidence chains for academic impact</i>	<i>115</i>
<i>Figure 6.1 The schematic diagram on the stages of generation of research impact</i>	<i>127</i>
<i>Figure 6.2 The schematic diagram of the process of knowledge generation</i>	<i>129</i>
<i>Figure 6.3 The schematic diagram of the process of knowledge publishing</i>	<i>131</i>
<i>Figure 6.4 The schematic diagram of the first wave of knowledge dissemination</i>	<i>133</i>
<i>Figure 6.5 The schematic diagram of process of knowledge understanding and extensive dissemination</i>	<i>135</i>
<i>Figure 6.6 The schematic diagram of the process of knowledge transformation and regeneration ..</i>	<i>137</i>
<i>Figure 6.7 The comprehensive schematic diagram of the generation of research impact for publications</i>	<i>140</i>
<i>Figure 7.1 The hypothetical model for the impact of research papers</i>	<i>180</i>
<i>Figure 7.2 Measurement models for the impact of research papers</i>	<i>183</i>
<i>Figure 7.3 The improved model for the impact of research papers</i>	<i>185</i>

List of Abbreviations

AGFI	Adjusted goodness of fit index
AHRC	Arts and Humanities Research Council
AIC	Akaike information criterion
AJG	Academic Journal Guide
ALM	Article-Level Metrics
Amos	Analysis of Moment Structures
API	Application Programming Interface
BBSRC	Biotechnology and Biological Sciences Research Council
CED	Crossref Event Data
CNIPA	National Intellectual Property Administration in China
CRS	Congressional Research Service
CSAB	Content Selection and Advisory Board
DOI	Digital object identifier
DPCI	Derwent Patents Citation Index™
EBSCO	Elton Bryson Stephens Company
EPO	European Patent Office
ERA	Excellence in Research for Australia
ESRC	Economic and Social Research Council
GFI	Goodness of fit index
HEFCW	Higher Education Funding Council for Wales
HEI	Higher Education Institutions
HTML	Hypertext Markup Language
IBM	International Business Machines Corporation
JIF	Journal Impact Factor
JPO	Japan Patent Office
KIPO	Korean Intellectual Property Office
MAREC	MAtrixware REsearch Collection
MRC	Medical Research Council
NSFC	National Natural Science Foundation of China

OFR	Open Funder Registry
ORCID	Open Researcher and Contributor ID
PCT	Patent Cooperation Treaty
PDF	Portable Document Format
PLOS	Public Library of Science
RAE	Research Assessment Exercise
RE	Research England
REF	Research Excellence Framework
RePORTER	Research Portfolio Online Reporting Tools
RMR	Root mean square residual
SEM	Structural Equation Modeling
SEP	Standard Evaluation Protocol
SFC	Scottish Funding Council
SJR	Scientific Journal Rankings
SNSF	Swiss National Science Foundation
SPSS	Statistical Package for the Social Sciences
SSM	Soft System Methodology
SSRN	Social Science Research Network
STAR METRICS	Science and Technology for America's Reinvestment: Measuring the Effect of Research on Innovation, Competitiveness and Science
NERC	Natural Environment Research Council
NIH	National Institutes of Health
NSF	National Science Foundation
USPTO	US Patent and Trademark Office
WIPO	World Intellectual Property Organisation
WoS	Web of Science

Publication

1. Lin, D., **Gong, T.**, Liu, W., & Meyer, M. (2020). An entropy-based measure for the evolution of h index research. *Scientometrics*, 125, 2283-2298. <https://doi.org/10.1007/s11192-020-03712-1>
2. **Gong, T.**, Liu, W. & Wu, S. (2022). An investigation of the quality of altmetric data with Altmetric.com and PlumX as Examples. *Proceedings of the twelfth international conference on business intelligence and technology*, 1-7. IARIA ISBN: 978-1-61208-955-3.
3. Gong, Q. & **Gong, T.** (2024) Review of the Development of Geographic Information Technology in China. *Geomatics & Spatial Information Technology*, 47(3), pp 1-7. (Written in Chinese)
4. Gong, Q. & **Gong, T.** (2024). Research on Comprehensive Land Consolidation in the Whole Region. *Modern Agriculture Research*, 30, 126-128. (Written in Chinese)
5. **Gong, T.** & Xu, F. (2019). *The analysis of major altmetric data providers* [Conference Paper]. The Third annual meeting of the evaluation methods and applications sub-society of the Chinese Society of optimization, overall planning and economic mathematics, Beijing, China (Written in Chinese)
6. Gong, Q., Li, D., **Gong, T.** & Li, D. (2018). Discuss on the computational thinking of building the geographic information sharing platform. *Engineering of Surveying and Mapping*, 27(1), pp.1-4 DOI: 10.19349/j.cnki.issn1006-7949.2018.01.001 (Written in Chinese)

Chapter 1 - Introduction

1.1 Research motivation

Science and technology are increasingly becoming the leading forces for the development of the economy and society. Many countries have significantly increased their investment in science and technology in order to be competitive in the global environment and have paid more attention to scientific research (Matyushok et al., 2021). Demand for research evaluation is growing in many countries. Research evaluation plays a vital role in the management of science and technology as it provides accountability to the public and supports decision-making to key stakeholders (Wanzer, 2021).

Research impact evaluation is a hot and challenging topic that has received increasing attention in recent years (Reed et al., 2021; Kassab et al., 2020). For instance, the Research Excellence Framework (REF) in the UK has started to assess research impact since 2014, and such assessment accounts for 25% of overall evaluation outcomes (REF 2021). The ACUMEN Consortium (in Europe) launched a new portfolio to assess academic researchers, emphasising the values and impacts of individual scientific output (ACUMEN, 2014). Various indicators in bibliometrics, webometrics and altmetrics are used in the portfolio to measure research impact and social impact. In the UK, the REF evaluates the research impact by considering different dimensions of influence, such as economy, culture, public policy, health and welfare, environment, quality of life, creativity, production, and international development (Samuel & Derrick, 2015). However, the term “research impact”, as defined by the REF, means the research influence beyond academia. In most other countries, the academic impact and socioeconomic impact of research are often viewed as one (Penfield et al., 2014). When the benefits from research are tangible, measurable and recognised outside of academia, the research is said to have an impact, and this impact could be economic, environmental or cultural (Schnitzler et al., 2016). This thesis argues that a research output’s impact should be comprehensive, encompassing both the academic impact generated by scholarly research outputs and broader impacts beyond academia. That is, the term “research

impact” in this thesis is defined as a comprehensive impact, which contains academic impact and broader impact (social, economic, etc.) generated together from the research outputs.

Specifically, the evaluation methods for research impact, apart from the qualitative methods such as peer reviews, citation-based scientometric indicators are the key measures to reflect the research impact (Mingers & Leydesdorff, 2015). Abramo (2018) summarised that scientometrics measures the impact of a research paper or a case study on future scientific and technical advancement, and the number of citations of a paper is a natural indicator of the impact of the paper. Therefore, citation-based indicators mainly reflect an academic impact.

It is easy to believe that the research output with high quality and high impact would be highly cited. However, many scholars argued that pursuing high citation numbers becomes their unique goal, and the citation count has become a number game (Fong & Wilhite, 2017; Urlings et al., 2021). Citation manipulation in scientific publishing is a widespread issue that threatens the integrity of academic research (Joshi & Pandey, 2024). The increasingly complicated citing motivations often lead to manipulated citation data (e.g., citing each other or citing by helpers), and thus, the reliability of citation-based indicators is declining (Fong & Wilhite, 2017). Relying on a single citation data source to evaluate impact may compromise the accuracy and fairness of the results. Because all desirable results (like citations) can be produced by improper actions/processes such as cheating. To increase the reliability of impact evaluations, it is important to evaluate not only the results of a piece of research but also its generating processes. For example, for a normal highly cited publication, there should be a series of evidence on high views (or abstract views), downloads, commentaries or discussions. When such evidential data are very small, we have reasons to doubt the authenticity of its high citation that represents the impact. If we can build an evidence-based model for impact evaluation, we will be able to produce more reliable assessment results.

We believe these ideas should also be applied to aggregate comprehensive research impacts, including academic and broader impacts. Therefore, in aggregating comprehensive research impacts of publications, we should utilise the evidential data that occurred in the normal processes of impact generation for a research publication, from publishing to spreading, understanding, and finally generating the impact. Of course, developing such novel evaluation methods is not trivial.

On the other hand, after altmetrics emerged in recent years, scholars started to discuss whether altmetrics could measure a wide range of impacts for research outputs, such as social impact (Melicherová et al., 2020; Reale et al., 2018). Although a common definition of altmetrics is still lacking (Haustein, 2016; Konkiel, 2016), altmetrics has been mostly used as online metrics to measure the impact of scholarly materials with an emphasis on social media platforms, and these metrics differ from the traditional citation-based indicators (Tahamtan & Bornmann, 2020). Altmetrics present good advantages compared to citations: they have potentials to measure research influence fast and real-timely, and able to cover diverse scholarly outcomes (Ezema, 2017). However, altmetrics are confronted with several challenges, such as the robustness of the underlying system, data quality, bias, lack of standards etc. (Bar-Ilan et al., 2018). Existing research reveals that social media metrics might not reflect societal impact as expected (Haunschild et al., 2019; Robinson-Garcia et al., 2018; Tahamtan & Bornmann, 2020) although they still provide a unique opportunity to observe the interactions and communications between academicians and broader audiences.

Without a doubt, the emergence of altmetrics has increased the possibility of applying multi-source data in the research evaluation and represented the thinking of carrying out the quantitative evaluation method across a wider range of data sources.

Therefore, this research aims to utilise multiple data resources, including scientometric and altmetric data, to develop novel approaches to quantitatively evaluating research impact

based on a chain of processes with supportive evidence and to improve the reliability of the evaluation results.

1.2 Research questions

To achieve the above research aims, we ask the three following core research questions and then propose research plans.

(1) How can we identify and select suitable altmetric data for evaluating the impact of research publications?

Scientometric data (e.g. citations) has been widely used to reflect academic impact of research publications. Although many altmetrics platforms provide flourishing indicators of research impact, it is unclear which are of good quality and can be used in real impact evaluations (e.g., which are more reliable). Chapter 4 in this thesis will systematically investigate these issues.

(2) How can we aggregate those altmetric and scientometric data based on a cause-effect analysis to develop more reliable metrics for the impact of research publications?

Traditionally, the result data, such as citations, are used directly or computed with suitable weights to form metrics for the impact assessment. As discussed above, such approaches are often problematic. One should aggregate impact based on suitable cause-effect chains to give more reliable results. To this end, Chapter 5 in this thesis will first develop an evidence-chain-based method for illustrating the evaluation of impact, and Chapter 6 will further study the cause-effect chains of impact generation for research publications.

(3) How can we apply the above approaches in real applications?

We will collect real data and carry out empirical studies to produce research impact indicators for real applications in Chapter 7.

1.3 Research contributions

There are six main contributions in this thesis:

- 1) An innovative evidence-chain-based quantitative evaluation method for the impact of research publications is developed in the thesis. We initially propose this methodology to evaluate the academic impact of research work and successfully recognise the journal with a more genuine impact. This method overcomes the drawbacks of a results-driven perspective in the current research evaluation system and can provide more reliable outcomes of assessment results.
- 2) This thesis further explores the evaluation methods for a comprehensive impact of research publications based on evidence chains. A conceptual model of the comprehensive impact of research publications is developed by analysing the cause-effect chains of the impact generation process. The identified stages of the research impact generation process provide a novel, clear and logical connection to analysing impact. It breaks through the original limitation of only considering the output's results data in the impact evaluation. Using the data collected at each stage to evaluate the impact could solve the attribution and consistency problem and make the final assessment solid and reliable.
- 3) The case study successfully evaluates the comprehensive impact of research outputs by utilising the cause-effect chain of the research impact generation process and constructing the impact evaluation models using a structural equation modelling approach. The newly established quantitative evaluation methodology for research impact could support various purposes of assessing requirements, including the publication and journal's impact within required categories, disciplines, and research areas.
- 4) Our evidence-chain-based evaluation method is a full-chain evaluation that logically connects with causal chains of the impact generation process. This is the first time in the literature that a full chain of evidence from the research knowledge generation to impact generation is applied to evaluate impact. By incorporating every stage of the

impact generation process into the evaluation, this approach can capture the influence of research papers on academic and broader dimensions, which reflects the comprehensive impact. Our method for research evaluation is a new way of thinking and provides an innovative insight into the impact evaluation system.

- 5) The empirical study on altmetrics platforms, Altmetric and PlumX, involves a detailed analysis of altmetric data coverage, disciplinary differences, data consistency, and the reliability of the data sources, providing a thorough understanding of the altmetric data sources and platforms. Findings and suggestions in this thesis provides a valuable references for subsequent use of altmetrics in the research.
- 6) Existing literature indicates that Altmetrics has certain limitations as a direct tool for research impact evaluation. These metrics primarily reflect public interest and preferences, making them less suitable for directly measuring impact, such as social impact. However, this thesis uses altmetric data to extract indicators, recognising its significant role in impact measurement. It offers a new perspective on the appropriate and effective use of altmetric data in the research evaluation.

1.4 Thesis structure

The remainder of the thesis is structured as follows.

Chapters Two and Three present a literature review of the thesis. Chapter Two reviews existing literature on research evaluations, including the rise of research evaluation, its development, and its impact evaluation.

Chapter Three reviews the literature on traditional metrics for academic impact evaluation and illustrates the most used indicators and their limitations. Then, it reviews popular indices that reflect the journals' impact. It continues with a systematic review of altmetrics, including concepts and definitions, altmetrics platforms and data sources, and critical topics in the altmetrics study. Finally, it identifies the knowledge gaps.

Chapter Four investigates the quality of altmetric data quality with Altmetric and PlumX as examples to answer the first research question.

Chapter Five proposes a novel evidence-chain-based method to evaluate the research impact to answer the second research question. It also includes a case study comparing two journals' academic impacts using the evidence-chain-based evaluation approach.

Chapter Six illustrates the cause-effect chains of comprehensive impact for research publications and demonstrates a conceptual model for the research impact generation process. It builds up the prerequisite of evidence-chain-based comprehensive impact evaluation methods for publications.

Chapter Seven is a case study that uses the structural equation modelling approach to construct indicators for comprehensive research impact based on cause-effect and evidence chains of the impact generation process. It demonstrates the various applications and potentials of the novel evidence-chain-based research impact evaluation methodology.

Chapter Eight concludes the thesis, summarises research limitations, and proposes future research.

Chapter 2 - Literature review on research evaluation

2.1 Rise of research evaluation

The demand for research evaluation is increasing in many countries. To be brief, research evaluation reveals the progress and achievement of academic research. The rise of research evaluation is mainly due to the growing demand for the use of scientific research evaluation results in society, the governments, and the scientific community. According to Gingras (2016), research evaluation has the following aspects of roles: (1) providing accountability to the public, (2) supporting policies and decision-making to the governments, and (3) using in the research management. The following paragraphs elaborate on these functions.

Since the middle of the 20th century, science and technology are playing an increasingly vital role in the international competition and societal development (see e.g., Pavitt, 1990). Many countries have recognized that advanced technology is capable to promote the development of economy and enhance the national strength. Advanced research production may improve a nation's scientific progression and worldwide competitiveness (Marginson, 2022). Since more public money has been spent on scientific research, expectations of investment return are also growing. The public asks the governments and the scientific community to explain where the public money has been spent, and how the research has been funded (Sanghani & MacDonald, 2024). By now it is widely agreed that Government-funded research should provide accountability. Research institutions and universities face being efficient and accountable for the research funds. These pressures make evaluation critical. In response to the public, many governments have tried to link funds to research performance to review whether public spending is “value for money” (Grant, 2006). Furthermore, to embody the importance of research activities and be responsible for taxpayers, many countries such as the United States, Japan and France have promulgated laws and regulations on scientific evaluation to ensure a healthy and systematic development of scientific research by legislation (Marginson, 2022).

When the public wishes (or requires) to monitor the budget of research input and evaluate the process and results of scientific research, governments must also demonstrate the results of publicly funded scientific research through assessment and support the decision-making on fund allocation at the same time (Benito & Romera, 2019). Consequently, the research evaluation outcome also supports decision-making on fund distribution. In the UK, the stated aims of implementation of research evaluation under the REF are to inform the selective allocation of research funding to higher education institutions, as well as demonstrate its benefits (REF 2021). Besides, the results of the research evaluation also offer indications for policymaking in scientific institute.

More importantly, research evaluation is a vital tool in the management of science and technology (Xu et al., 2013; Reale et al., 2017). Research evaluation could guide the development of scientific research, improve the shortcomings of scientific research organisations, and fairly manage individual researchers (Xu et al., 2013). Huang and Chang (2008) also believe the research evaluation could be an incentive for scholars to improve the quality of their research. According to Qiu et al. (2003), comparisons of indicators, including citations or impact factors among the academic entities (such as journals, organisations, institutes), are able to analyse their strengths, weaknesses and improvement methods. For decision-makers, the comprehensive evaluation information would provide valuable evidence to support project planning, project funding, and personnel arrangements, among others. For individual researchers, evaluation on their scientific output may significantly influence their promotions, tenures and academic careers. Furthermore, as scientific production increases, the traditional evaluation methods such as internal evaluation or peer reviews began to meet limitations on time consuming (Ernst et al., 1993; Kumar, 2009). It is therefore necessary to conduct study on the different evaluation methods. Overall, the development and renewal of the science and technology put forward new requirements for scientific evaluation constantly, which enrich the connotation, methods and means of research evaluation (Marginson, 2022).

To sum up, research evaluation is gradually rising with the increasing need for development in science and technology. It provides evidence-based information for decision-making in the management of scientific activities according to different evaluation objectives and demonstrates accountability to the public for the investment of public funds in research.

2.2 Development of research evaluation

2.2.1 Definitions of research evaluation

As people gradually realise the importance of scientific research outcomes, issues such as evaluating scientific outputs, analysing scientific activities, and improving research efficiency and performance become more of a concern for governments and scientific communities worldwide (Geuna & Martin, 2003; Guthrie et al., 2013). According to Wang (2006), the earlier formal scientific evaluation activities led by governments began in the United States in early decades of the 20th century. The Congressional Research Service (CRS) of the United State started to analyse and evaluate research related scientific activities from 1920s which can be regarded as the prototype of scientific evaluation (Wang, 2006). Then, some countries such as Japan, France, Germany and Canada both carried out the assessment of science and technology activities in the 1960s. In the 1980s and 1990s, along with the commercialisation and marketization of scientific and technological achievements, many Western countries, such as the UK, Switzerland and some developing countries, such as China, gradually paid attention to the development of science and technology and scientific research (Qiu et al., 2010; Wang, 2006).

The term “research evaluation” involves broad implications, and it is challenging to find an exact definition. According to Wanzer (2021), research evaluation refers to evaluating scientific or scientific research activities. The object of evaluation is diverse and could be any entity, people or event related to scientific activities (Greene, 2016). Research evaluations comprise the evaluation of scientific publications (including papers, journals, books, patents, standards, databases, etc.), scientific institutes (including research institutes, universities, etc.), research work (including research plans, research proposals, research projects, research achievements, research inputs and outputs, research performance, research capabilities,

research staffs, research management, research policies, etc.) and disciplines (such as the development status, stages, levels, prospective, and correlations between different disciplines, etc.) (Qiu et al., 2003). As researchers' contributions are usually present in the form of outputs such as publications and patents, the research evaluation thus means to reflect the value of researchers' scientific outputs (Qiu et al., 2003)

2.2.2 Value orientation of research evaluation

In recent years, the value orientation of research evaluation is that in addition to the evaluation of academic value on scientific research outputs, the consideration of social impacts beyond the academia has been increased (Grant, 2006; Penfield et al., 2014; REF 2014; REF 2021; Reale et al., 2018). The most representative example is the REF in the UK. Evaluations of the submitted unit's research impact and research environment account for 25% and 15% of overall outcomes (REF, 2021). The motivation behind the move of assessing social and economic impact is the increasing expectation from the research institutions and funding bodies to demonstrate how the research they fund offers apparent benefits to societies (Samuel & Derrick, 2015). Nowadays, most research funding bodies in the UK have explicitly required a summary of societal impact (also known as the statement of "pathways to impact") as a compulsory part in the grant application form (UK Research and Innovation, 2019). Such changes of criteria further response to the increasing expectation on social benefits from funding projects. Obviously, as Donovan (2011) remarks, impact is a strong weapon for making an evidence based case to governments for enhanced research support. Furthermore, Penfield et al. (2014) points out four benefits of assessing impact beyond the academia, which are (1) to enable research institutes to disseminate the contribution of their research to local, national and international communities; (2) let funding bodies to demonstrate the value of research to public, stakeholders and governments as the accountability; (3) to inform funding decisions for desired impact; and (4) to understand the methods and routes by which research lead better impact. Indeed, the core purpose of research should be generating and transforming the new findings to knowledge and then benefit the society as a whole.

2.2.3 Methodologies

Since the evaluating purposes and objectives are diverse, evaluating methods are also under constantly developing during the development of research evaluation. The early evaluating methods are mainly carried through peer reviews within the academic community (Brinn et al., 2000; Qiu et al., 2003). In recent years, quantitative methods have developed rapidly (Kothari, 2004; Hicks et al., 2015; Cronin & Sugimoto, 2014; Mingers & Meyer, 2017). In addition some comprehensive evaluation methods based on modern information technology have emerged. There are a massive number of methods in use in the practical assessment, likewise each of them has own advantages, disadvantages or adaptabilities. In this study, we only select some representative mythologies to review briefly, which summarised in the table 2.1 below. Generally, these evaluation methods can be clustered into three categories, qualitative evaluation methods, quantitative evaluation methods and data orientated comprehensively evaluation methods (Qiu et al., 2010; RAND, 2013). In addition, we have outlined the characteristics of each method, which are presented in the Table 2.1. There is a need to emphasis, each of the methods has its strength and shortcomings. It would be more appropriate to select specific evaluating methods in accordance with the needs in application.

Table 2.1 Brief summary of major methodologies

	Methodologies	Brief introduction
Qualitative	Peer review	Peer review refers to a process by which scientific research activities or outputs are evaluated by a group of experts in the appropriate field according to certain standards and procedures (Abramo & D'Angelo, 2011; Van-Raan, 1996).
	Case studies	Case studies are in-depth examinations of research through a narrative of research process and outcome. They are flexible to capture a wide variety of impacts, including the unexpected, and can provide full context around a piece of research, researcher or impact (RAND, 2007)

	Document review	This method refers to a general primary step for the review of existing documentation, reports or other relevant materials (RAND, 2013).
	Logic models	Logic models are graphic depiction of the logical relationships among multi elements of a system that illustrate a sequence of cause-and-effect relationships paths towards a desired result (Xu et al., 2013). A logical model is used to describe what a program intends to achieve and how it will achieve (Jordan, 2010).
	Soft System Methodology (SSM)	SSM is a systems-based approach to problem structuring and taking action in complex situations that has been developed various application in strategy and other systems methods (Mingers & Taylor, 1992; Mingers & White, 2010). It has successfully used to decompose organisational strategies and build indicator systems to measure scientific research performance (see Liu et al. 2012; Xu et al., 2013).
Quantitative	Bibliometrics	Bibliometrics is the use of statistical methods to analyze books, articles, and other publications. This method intends to provide evolutionary quantitative models and contains a range of techniques for assessing quantity, dissemination, content of publications, patents and so on in science, technology, and scholarship (White & McCain, 1989; RAND, 2013).
	Economic analysis	A method of comparatively analyzing costs and consequences in economics. Purpose is to assess whether benefits outweigh opportunity costs and whether efficiency is achieved. Three types: cost

		benefit analysis (CBA), cost-effectiveness analysis (CEA) and cost-utility analysis (CUA) (RAND, 2013).
Data orientated	Data mining	Data mining allows access to and understanding of existing data sets. Uses algorithms to find correlations or patterns and present them in a meaningful format, reducing complexity without losing information (RAND, 2013).
	Data visualization	A tool for data summarization, presenting large amounts of data in a visual format for human comprehension and interpretation (RAND, 2013).
	Artificial neural network	The new method under development that is using artificial neural network technology to acquire knowledge through BP algorithm and machine learning or training, and then evaluate the target by simulating human brain (Karayiannis & Venetsanopoulos, 2013).

2.2.4 Frameworks

Varied evaluation methods and continuously developed evaluation indicators enrich the diversity of evaluation frameworks (Geuna & Martin, 2003; Reale et al., 2017). When designing the framework, people start to concern themselves with the evaluation efficiency by comprehensively utilising different methods based on their advantages and disadvantages (Geuna & Martin, 2003). For example, one of the main reasons for the Research Assessment Exercise (RAE) being replaced by the REF in the UK is to control the total assessment budget by partly applying the bibliometric method to reduce some cost burden (House of Commons, 2011). In the US, the evaluation framework of STAR METRICS (Science and Technology for America's Reinvestment: Measuring the EffectT of Research on Innovation, Competitiveness and Science) has been promoted for using low-burden quantitative methods such as data mining and bibliometric indicators (Lane, 2010).

Since UK's REF is relatively mature, we will mainly review this framework in this thesis. Nevertheless, there are several well-known frameworks in use in the different countries. For example, the ERA (Excellence in Research for Australia) measures the performance of research in Australia using Bibliometrics and peer reviews. STAR METRICS is the framework under development in the US to assess the performance of research and researchers by data mining, data visualisation and economic analysis. Productive Interactions are the framework used in Europe across several countries and for multiple disciplines and comprise both qualitative and quantitative evaluating methods. AERES, an independent agency that evaluates French research and higher education institutions based on internal and external assessments. SEP (Standard Evaluation Protocol) in Netherlands assesses four aspects of research: quality, productivity, relevance, vitality of the research (Régibeau & Rockett, 2016; RAND, 2013).

Research Excellence Framework (REF)

The REF is UK's world-leading system for assessing the excellence of research in higher education institutions (HEIs). The REF replaced the former RAE since 2014. According to REF (2021), the evaluation system is conducted jointly by the four UK higher education funding bodies include Research England (RE), the Scottish Funding Council (SFC), and the Higher Education Funding Council for Wales (HEFCW) and the Department for the Economy, Northern Ireland (DfE). Its outcomes are used to allocate about 2 billion pounds per year of public funding for UK's universities' research, and affect their international reputations (Nurse, 2015). The assessment outcomes also provide benchmarking information for research institutes and public, as well as providing a basis for the strategic decision-making on national research priorities (REF 2021; Nurse, 2015).

According to the REF (2021), assessments are conducted at the level of subject-specific sub-panels, which roughly correspond to a 34 units of academic disciplines. Each sub-panel will examine the submissions based on three distinct elements: quality of research outputs (account for 60% of overall outcome awarded), impact of research (carry a weighting of 25%),

and vitality and sustainability of the research environment (carry a weighting of 15%) (REF, 2021a). Furthermore, the criteria for each element is, for instance the quality of outputs will be assessed in terms of 'originality, significance and rigour', with reference to international research quality standards. For impact, the sub-panels will assess the 'reach and significance' of impacts on the economy, society, culture, public policy or services, health, the environment or quality of life and so on. Regarding the element of environment, the criteria is in terms of 'vitality and sustainability' of the submitted unit, including the approach to enabling impact from its research, and its contribution to the vitality and sustainability of the wider discipline or research base (REF 2021). There is a need to emphasize, the assessments for the research outputs are conducted via expert panels' peer reviews, and supplemented by the bibliometric information (Sousa & Brennan, 2014; REF, 2021)

The final evaluation outcomes are presented in the qualified profile with stars in accordance with REF's standard for research level. The criteria of each level of REF's standards of research quality is: (1) 4* indicates the quality of the research output is world-leading in terms of originality, significance and rigour; (2) 3* means the quality that is internationally excellent in terms of originality, significance and rigour but which falls short of the highest standards of excellence; (3) 2* is recognized internationally in terms of originality, significance and rigour, and (4) 1* indicates quality that is recognized nationally in terms of originality, significance and rigour (REF, 2021). Additionally, for each of the three elements of the assessment - outputs, impact and environment – sub-panels will develop a sub-profile, showing the proportion of the submission that meets each of the four starred quality levels (Bornmann, Haunschild and Adams, 2019).

From RAE to REF

In the era of RAE, all research outputs were assessed by peer reviews (HEFCE, 2009). Although peer reviews are widely recognized as the most reliable method for evaluation, the significant drawbacks of this method are high labour intensive and time consuming (Brinn et al., 2000). Obviously, the UK's top research evaluation system suffers from those inadequacies as well.

As remarked in the House of Commons (2011), the cost of RAE was too high, while a bibliometric indicator-based process was expected to be much cheaper and less burdensome. Additionally, peer reviews also are criticized for other issues such as biases among experts, conflict in interest and Matthews' effect (Cole et al., 1981; Ernst et al., 1993; Kumar, 2009) etc. The government thus decided to change the system for assessing research quality to become mainly metrics-based in order to reduce the burdens imposed on universities (House of Commons, 2011). After the 2008 RAE, HEFCE (the Higher Education Funding Council for England), the former funding council of Research England (RE) before 1st April 2008 (RE, 2019), proposed to introduce a metrics-based system for subjects in science, technology, engineering and medicine (HEFCE, 2009), while this proposed change has caused significant debates and widespread criticisms in the academic community. There were substantial concerns on the potential implications of assessing the science related and non-science-related subjects differently, as proposed under the bibliometrics approach (RAND, 2013). The bibliometrics pilot exercise then backtracked after this trial (RAND, 2013). Reasons behind this unsuccessful move showed the lack of trust in metrics driven methodologies, to improve the current quantitative methodologies should be an incessantly researching objective.

2.3 Evaluation of research impact

The evaluation of the impact of the research is one of the key elements of the research evaluation. Mingers and Leydesdorff (2015, p.1) find that the main themes of scientometric research "include ways of measuring the quality and impact of research". Scientometrics is grounded in the quantitative analysis of scientific advances, mainly in the area of the "research results", for which it tries to measure impact for evaluative purposes (Abramo, 2018). When the benefits from research are tangible, measurable and recognised outside of academia, the research is said to have an impact, which could be economic, environmental or cultural (Schnitzler et al., 2016). The impact of the research could include changes to public policy, health care, or quality of life (Liu et al., 2023).

In the UK's REF, "research impact" is defined as the influence of research beyond academia (REF 2021). Academic impact and socio-economic impact are assessed as distinct categories.

REF evaluates research impact by considering various sectors of influence, including the economy, culture, public policy, health and welfare, the environment, quality of life, creativity, production, and international development (Samuel and Derrick, 2015).

In contrast, many other countries view academic and socio-economic impacts as a unified concept (Penfield et al., 2014). However, a clear distinction exists: "academic impact" refers to intellectual contributions within academia, while "external socio-economic impact" pertains to influences beyond academic circles (Penfield et al., 2014). Reale et al. (2018) further differentiates between scientific impact, dissemination, political impact, and social impact, suggesting that social impact represents the culmination of these earlier stages of research. Similarly, Milat et al. (2015, p.2) define research impact as "any type of output from research activities that can be considered a 'positive return' for the scientific community, health systems, patients, and society at large."

Furthermore, the research impact encompasses changes in practice, knowledge, understanding, attitudes, and ideas resulting from research (Morton, 2015). Morrow, Goreham and Ross (2017) build on this by exploring the pathways to impact in leadership, governance, and management research, identifying key components that enable it. The impact is generally viewed as the changes in research outcomes generated in academia, the economy, and society. However, attributing such changes to specific research poses challenges, including difficulty measuring impact, delays between research activities and observable effects, and determining whether the research was the primary driver of the changes observed.

Milat et al. (2015) highlight that research impacts are inherently complex, unpredictable, and non-linear. There is often a tendency to measure what is easily quantifiable rather than focusing on what is genuinely significant. The fundamental purpose of research, however, is to generate new findings, transform them into knowledge, and ultimately benefit society.

Additionally, some scholars distinguish the political impact from the broader impact of research. In Europe, political impact is often assessed separately, particularly in studies focusing on the relationship between science and policy and improving research outcomes' influence on policymaking (Reale et al., 2018). Political impact occurs when research findings are utilized by decision-makers or social actors to inform policies and actions (Flecha, 2014). While it overlaps with social impact, political impact specifically relates to policy development and policymaking changes (Lemay and Sá, 2012).

Quantitative methods for evaluating impacts are varied and lack a universally accepted “gold standard” checklist or standardized process (dos Santos Ribas et al., 2020). Experimental and statistical methods, along with evidence synthesis approaches, are typically applied in summative evaluations, while textual, oral, arts-based methods, systems analysis, and indicator-based approaches may be employed in both summative and formative evaluations. Certain evaluation designs aim to disentangle the contribution of research to observed impacts as part of a broader set of factors, whereas others focus on demonstrating a direct and sole attribution between research and its impact (Reed et al., 2021).

However, experimental and statistical methods for impact evaluation have notable limitations. Synthesized studies often lack common standards for data collection and reporting, reducing the reliability of evidence (Victoria et al., 2017). Additionally, metric-based approaches to impact assessment have been criticized for oversimplifying complex phenomena and producing partial or misleading results (Bayley & Phipps, 2019).

Garbarino and Holland (2009) argue that integrating qualitative and quantitative methods in impact evaluations offers dual benefits: it provides statistically reliable measures of the magnitude of impact while also delivering deeper insights into program effectiveness, including the factors contributing to their success or failure and areas for potential improvement. The primary distinction between qualitative and quantitative methods lies in

the nature of the data they generate. Quantitative methods yield numerical data, whereas qualitative methods produce descriptive data presented in textual or narrative forms.

Furthermore, standardised methodologies, tools, metrics, and processes for assessing broader research impacts remain underdeveloped. As a result, case studies are frequently adopted in evaluation frameworks (Tahamtan & Bornmann, 2020). While altmetrics are increasingly utilised, their limitations persist. Bornmann and Haunschild (2018b) caution that frequent mentions of an article on social media often reflect public interest and engagement rather than providing meaningful insights into its societal impact (see section 3.4).

2.4 Summary

This chapter presents a literature review on research evaluation, including its rise, development, and research impact evaluation. Each section thoroughly reviews relevant definitions, concepts, methodologies, and frameworks.

According to the existing literature, the use of quantitative evaluation methods in research evaluation is a growing trend, as they are economical and less burdensome. However, existing quantitative evaluation methods are not perfect enough. Quantitative evaluation methods for impact are still developing and lack a universally accepted standardised process, methodologies, or metrics. Furthermore, current bibliometric assessment mainly examines 'research results' to evaluate impact, such as citations. These approaches are results-only. Therefore, developing an evaluation method that integrates process and result is necessary to improve the assessment philosophy.

Chapter 3 - Literature review on traditional metrics and altmetrics for research impact

3.1 Introduction

Concerning quantitative methods and tools for research impact evaluation, scientometric indicators based on citations are the key measurements to evaluate academic impact in most literature (Mingers & Leydesdorff, 2015; Cronin and Sugimoto, 2014; Thelwall, 2016). Peer review judgments are the leading qualitative method for evaluating the academic impact. The peer-review process evaluates an author's scholarly work, research, or ideas through the judgments of experts in the same field and is regarded as essential for ensuring academic and scientific quality (Abramo & D'Angelo, 2011; Brinn et al., 2000).

For a broader impact beyond academic impact, the UK's REF uses case studies to illustrate the societal benefits of the research work to present impact (Thelwall, 2020). Regarding the quantitative methodologies for broader impact, there is no consensus yet. Since altmetrics emerged in recent years, scholars started to discuss whether altmetrics could measure a wide range of impacts for research outputs, such as social impact (Reale et al., 2018; Tahamtan & Bornmann, 2020).

This chapter will examine the traditional citation-based indicators used for impact evaluation and then comprehensively and critically review the altmetrics study to explore the research gaps of using various indicators for impact evaluation.

3.2 Traditional metrics for academic impact evaluation

The use of citation-related indicators in the assessment of publications' academic impact has been widely studied in the literature (Mingers & Leydesdorff, 2015; Li, Thelwall & Giustiini, 2012; Waltman et al., 2012). Bornmann and Haunschild (2016) noted that citation counts are commonly used as a measure of academic impact, reflecting the influence of a work based on how often it is cited by others. While citations are quantitative indicators of scientific impact,

they do not capture all forms of impact. A key benefit of using citation counts is their link to the publication process. Researchers need to contextualize their work within the existing literature through citations and bibliographic references, demonstrating their awareness of important developments in their field (Woelert, 2013).

Regarding scholarly impact evaluation specific to a research unit (including individual publication, researcher or institute and so on), quantitative evaluation indicators have evolved diversely from being simple to being compound. Traditional citation indicators typically include the total citation number, the max citation number, average citation and the proportion of highly cited papers, which mainly evaluate the overall impact of a research unit within a certain period (De Bellis, 2009; Bensman, 2007).

3.2.1 The classic citation indicators

Total citation (TC)

The Total Citation (TC) indicator represents the total citations of a set of publications in each time window. This indicator has been widely used in assessing individual researchers' work in earlier days (West and McIlwaine, 2002; Bensman, 2007). However, the indicator has been argued to be easily inflated by a few highly cited papers (Aksnes, 2006). Meanwhile, it also absorbed the undesirable cases where a high total citation accumulated from low-cited research papers.

Average Citation (C_{ave})

The Average Citation (C_{ave}) indicator calculates the mean number of citations for a set of publications. Several scholars have discussed its usefulness and limitations in research evaluations (Huang et al., 2008; Adler et al., 2008). Using this indicator alone in the evaluations may reduce the publishing motivations of research output, as a small number of highly cited papers would demonstrate a better mean number. It may also discourage young researchers from publishing their work freely.

Maximum citation (C_{max})

The Maximum Citation ($C_{\max}$) indicator represents the highest citation count in a set of publications. It also demonstrates the highest impact in a profile. Likewise, such an impact measure comes from one piece of work only. Therefore, it is deficient to evaluate the whole set of publications.

It is important to note that these classic indicators cannot effectively reflect the citation status for a set of publications, while more advanced indicators in later days do. Since 2005, when the h -index was proposed, the measurement for a group of publications has entered a new era.

3.2.2 The Hirsch-index (h -index)

Introduction of the h -index

There is no doubt for the scientific impact and achievements of Nobel Prize-winning scientists, while for most of other scholars, how to measure their cumulative academic impact? The classic citation indicators have advantages and obvious drawbacks. None of them could measure the different dimensions such as quality, productivity, age of citations at the same time in one calculation (Hirsch, 2005; Bar-Ilan, 2008; Ye & Leydesdorff, 2014). In 2005, Hirsch proposed the h -index, which quantifies the cumulative impact and relevance of a researcher's scientific output using only one index number. This index initiated the popularity of using citation-related indices to measure academic impact (Hirsch, 2005; Ye & Leydesdorff, 2014; Waltman & Van Eck, 2012; Knight, 2015; Merigó & Yang, 2017).

The h -index computes two dimensions which include the number of publications and the citation count for each publication for a researcher's output at the same time. It seeks the minimum citations of number n in at least n publications. The definition of the h -index is:

"A scientist has index h if h of his/her N publications have at least h citations each and the other $(N-h)$ publications have $\leq h$ citations each" (Hirsch, 2005, p.16569).

The h -index measures the broad impact of an individual's work. Different scholars' overall scientific impact can be compared through the h -index despite the difference in the total

number of publications or citations they have. As Hirsch (2005) argued, those who have a higher h number are likely to be more accomplished than others. He also indicated that the difference among scientists with high h indices appears to be much higher than average. There is a need to note, as Hirsch (2005) refers, whether the h -index is fair enough or applicable to all disciplines shall be further studied.

Verification of the h -index

Since the introduction of the h -index, some literatures started to verify the applicability of the h index in various disciplines and the limitations. For instance, Bornmann and Daniel (2005) compared the peer reviews for the applicants of post-doctoral research fellowships and found successful candidates do have higher the h -index than non-successful applicants on average. They believed the h -index is a promising rough measurement to assess the quality of a young scientist's work. Cronin and Meho (2006) compared the h -index and faculty rankings based on raw citations counts in the subject of information scientists and found the strong positive correlation between the two sets of rankings. They affirmed that the h -index is a reliable indicator to present the scholarly impact and the influence of scientists in the field of Information Science. More scholars carried out various papers to explore the h -index empirically and theoretically (Egghe & Rousseau, 2007; Egghe, 2006; Glanzel, 2006), and found different advantages and limitations.

Limitations of the h -index

Although the h -index has been regarded as a milestone in the development of the Bibliometrics, there are a lot of debates on its limitations correspondingly (Bornmann, Mutz & Daniel, 2008; Jin et al., 2007; Zhang, 2009). For instance, Hirsch further studied the consideration of career lengths of researchers, and he concluded that the h -index can only increase as time goes by. Thus, if a researcher stops to publish new papers or is not active any more, his h number remain the same. This could cause that the researchers in different working periods are not fairly compared (Hirsch, 2005; Kelly and Jennions, 2006). Furthermore, some scholars mentioned the discrimination power of the h -index is relatively

weak. As a result, various attempts to improve the h -type indices as well as some h -independent indicators generated that stimulated further explorations in bibliometric research in turns. We will demonstrate some indices of improvement or complement to the h index in the following literature review.

3.2.3 The g -index

The obvious advantage of the h -index is its using one number only to present the measurement of a researcher's academic production by taking account of both publication quantity and citations of each paper. Additionally, the h -index is insensitive to either lowly cited papers or minority highly cited papers for a researcher since the highly cited paper only count once in the h -index (Hirsch, 2005). However, scholars argued that highly cited papers are less considered whereas they should be the significant reflection for a paper's quality and impact (Egghe, 2006; Glanzel, 2006). For instance, if a paper has received 20 citations and they have been counted in his/her h -index, when ignoring the quantity change of publications, the h score may not change at all even such paper has further received 50, 100, or 200 citations. Another example is: if there are two scholars both have an h score of 3, the first person has three papers with 3 citations each, and the other person has three papers with citations of 2, 5, and 50. Do they have the same academic influence as they have the same h -index? Obviously, the latter has a larger academic impact. Therefore, Egghe (2006) proposed the g -index to improve the h -index by taking consideration of total citations of each paper.

Base on principles of the h -index, a set of publications' total citations (TC) are at least h^2 and the actual number would be much higher. Hirsh (2005) presents the TC equation as *the Number of TC* $= ah^2$, where a is defined as a proportionality constant, and the range of a is approximately between 3 and 5 through his experiments. Similarly, the g -index defines the highest number g for a set of papers when $\sum TC \geq g^2$. There is a need to emphasize that g represents a rank number, and all papers are supposed to be ranked in decreasing order according to each citation. In a sense, g is equal to or greater than h . when a group of researchers have a similar h -index number, whom has more citations in topper ranking papers

(in other words, some papers have been very highly cited), will present a higher g score. The g -index is particularly useful when comparing the similar h -index.

3.2.4 The h -dependent indicators

H-core

First of all, we would look at the concept of the h -core to help with understandings of all h -dependent indicators. In the h -index, rank h of a set of publications means at least h publications have h citations (Hirsch, 2005). Egghe and Rousseau (2006) proposed the set consisting of the first h articles as the Hirsch core (h -core). According to Ye and Rousseau (2010), all publications in the h -core could be recognized as the outstanding work for a scientist and major achievements in his scientific career. In the case, it is believed that the h -index loss many citations in the h -core (Ye & Rousseau, 2010; Jin et al., 2007)

The A-index and R-index

Egghe (2006) proposed the g -index, which corrected the less counted citations in the h -core, while Jin et al. (2007) proposed the A-index and the R-index, which were trying to achieve the same objective and provide their own advantages. The A-index is the average of citations received by the publications in the h -core. The name of this index is derived from the fact that it is just an average (A). Mathematically, it is written as:

$$A = \frac{1}{h} \sum_{j=1}^h Cit_j \quad (3.1)$$

The value of A is obviously equal or greater than h since $A \cdot h$ takes account to the total citation of all h -core publications. The drawback of this indicator is that it may punish the scientist with a higher h -index. Jin et al. (2007) then solved this issue by applying square root to the sum of total citations of h -core, that to form an improved indicator, R-index. Mathematically, it is written as:

$$R = \sqrt{\sum_{j=1}^h Cit_j} \quad (3.2)$$

The R-index carried forward the advantages of the h -index yet considered the total citations of all h -core papers. It could be used as a supplement indicator for the h -index. Jin et al. (2007)

also compared the R-index and the g-index. The former provides a computational advantage and the latter is rather relied on listing ranks and their square values.

The h^2 -lower, h^2 -center and h^2 -upper indices

Some scholars further researched the publications in the h -core, Bornmann (et al., 2010) proposed the h^2 -lower, h^2 -center and h^2 -upper indices to improve the accuracy of the h -index. Three indices are proposed as the h^2 -lower, h^2 -center and h^2 -upper indices, which describe the picture in the tail, the middle and the top of citation profile. Mathematically, they are defined as follows:

$$h^2\text{-lower} = \frac{100 \sum_{j=h+1}^N c(j)}{\sum_{j=1}^N c(j)} \quad (3.3)$$

$$h^2\text{-center} = \frac{100h^2}{\sum_{j=1}^N c(j)} \quad (3.4)$$

$$h^2\text{-upper} = \frac{100 \sum_{j=1}^N (c(j) - h)}{\sum_{j=1}^N c(j)} \quad (3.5)$$

The h_{rat} -index

The lack of discrimination power for the h -index has been argued for a long time. The value of h -index is only an integer, and two individuals may have totally different citation profiles while having the same h -index. When Ruane and Tol (2008) applied the h -index methodology to rank the economics departments in the Republic of Ireland, they argued that the h -ranking lacks a finer structure. In their research, they proposed rational h -indicators to distinct the cases with the same h -index value. It is defined as:

$$h_{\text{rat}} = h + 1 - \frac{n}{2h+1} \quad (3.6)$$

In the above formula, h is the original value of the h -index; n denotes the additional points (papers) needed to reach the h -index to $h+1$ in an observed publication set. Suppose we know a scholar's h index is h , and he wants to increase the h value to $h+1$, then he would need $h+1$ publications with $h+1$ citations at least each upon the h -index definition. If his total citations are just equal to h^2 , in other words there are no surplus citations in his h -core, the maximum citations needed would be equal to $2h + 1$ to increase h to $h+1$. Accordingly, $0 < n/(2h+1) \leq 1$.

The h_{rat} -index has the advantage of increasing a fraction part to h value that provides more granularity to an evaluation process. Two individuals with the same h value become comparable, and prolificacy would not be penalized any more. Nonetheless, the h_{rat} -index is a derivative of the h value and the scale of discrimination is limited due to its nature of maximum value.

There is a need to clarify, there are many other indicators are proposed to overcome the main drawbacks of the h index include the issues of citation distribution and excess citations in the h -core publications. For example, the j -index (Todeschini, 2011), the k -index (Ye & Rousseau, 2010), the Maxprod index (Kosmulski, 2007), the w -index (Wu, 2010), the π -index (Vinkler, 2009), the citation speed s -index (Bornmann and Daniel, 2010), and more indicators for field normalization such as the field-normalized total citations (Radicchia et al., 2008), the crown indicator (Van Raan, 2005), the n -index (Namazi & Fallahzadeh, 2010) and so on.

3.3 Indicators reflecting academic impact of journals

Journal evaluation also plays a vital role in the research evaluation field. Citation-based bibliometric indicators are still commonly used to reflect the relative impact of a scientific journal (Glänzel & Moed, 2002; Rousseau, 2002). The Journal Impact Factor (JIF) is calculated by the scientific division of Thomson Reuters and published annually in SCI Journal Citation Reports (JCR). Due to the popularity of the journal impact factor among science editors and research evaluators worldwide, JCR has become the legitimate authority for ranking scholarly journals (Bornmann et al., 2012). However, the journal impact factor also reveals some limitations (Fersht, 2009; Bornmann et al., 2012; Bollen et al., 2005 & 2009) and then many

scholars proposed complement indicators to improve it, such as the Eigenfactor™ metrics (Bergstrom, 2007), indicators of SCImago Journal Rank (González-Pereira et al., 2010), source normalized impact per paper (SNIP) (Moed, 2010), integrated impact (Wagner & Leydesdorff, 2012), and so on.

Journal Impact Factor (JIF)

The most well-known indicator for journal impact evaluation is the Journal Impact Factor (JIF). JIF has been used as a proxy of a journal's quality and scientific prestige (Leeuwen & Moed, 2005). The method was first developed by Garfield in 1972 (Garfield, 1972) and it calculates the average number of citations for research outputs published in a given journal in a two-year evaluation window (Garfield, 1971 & 1972). If an impact factor of a journal equals to 3 means that the articles published in this journal have been cited 3 times on average under the two year evaluation period. Let $IF_i(t)$ denotes the impact factor of journal i in year t . Mathematically it can be written as follows:

$$IF_i(t) = \frac{TC_i(t)}{P_i(t-1) + P_i(t-2)} \quad (3.7)$$

$IF_i(t)$ is the impact factor of journal i in year t and $TC_i(t)$ denotes the number of citations received to year t ; $P_i(t-1)$ and $P_i(t-2)$ are the number of papers published in journal i in year $t-1$ and $t-2$, respectively.

Today, JIF is calculated by the scientific division of Thomson Reuters and published annually in the SCI Journal Citation Reports, and due to the popularity of JIF among science editors and research evaluators worldwide, JCR has become the legitimate authority for ranking scholarly journals (Bornmann et al., 2012). Some arguments addressed that the two-year window is possibly too narrow to indicate the actual impact of a journal (Seglen, 1997). Therefore, JCR has also developed a five-year journal impact factor, which calculates the average number of citations for papers published in certain journals in a five-year evaluation period. Besides, JCR also introduced more journal evaluation indices such as:

- **The immediacy index**, which measures the average number of citations received by papers published in academic journals within a one-year evaluation period. It serves as an indicator of how quickly journal articles are cited.
- **The citing half-life**, which calculates the median age of publications cited by the journal in the JCR year. It serves as an indicator of how quickly journal articles are cited.
- **The cited half-life**, which counts the years from the current year that account for 50% of total citations of a journal receives. It identifies the typical age of the majority of articles cited in the journal.

However, JIF also reveals some limitations that scholars have criticized. For instance, various disciplines are not comparable (Agarwal and Agarwal, 2007), and the JIF can be easily manipulated by inauthentic citation data through self-citation, randomness or various motivations (Mingers & Xu, 2010; Yu et al., 2005; Agarwal and Agarwal, 2007; Chew et al., 2007) etc. Another drawback of JIF is that using JIF to evaluate individual papers is unreliable due to the uneven distribution of citations on each paper in most journals. The following literature review will discuss more improved indicators based on journal impact factors.

The Eigenfactor

Different from the JIF, the Eigenfactor index which is introduced by Bergstrom (2007) applies a weighted citation concept and gives more weights to highly ranked journals. According to Bergstrom (2007), the core methodology of Eigenfactor is to use so called the “Eigenfactor” algorithm, which allows readers to move from one journal to another along a chain of citations. This approach is similar to the Pagerank algorithm, which is adopted by the Google search engine to rank websites according to the importance (Fersht, 2009; Bergstrom, 2007). Additionally, the Eigenfactor is found to have strong correlation with the total number of citations for a journal (Davis, 2008), and it is believed to be more robust than journal impact factor as more considerations of weights on different journals (Bollen et al., 2009).

The SJR indicator

The SJR (Scientific Journal Rankings) indicator is developed by the SCImago research group which adopts the Pagerank algorithm to measure the prestige transferred from one journal to another. Basically, when a journal is cited by other prestigious journals in a given scientific area as one “incoming link”, it gains more prestige than a journal that is equally cited from less prestigious journals (Moed, 2012). The well-known bibliometric database, Scopus, applies this indicator to rank journal impact in their database. There is a need to note, it has been discussed there are inconsistent results from ranking journals using JIF and SJR indicators (Falagas et al., 2008), and the differences may be caused by the usage of databases from different sources.

Journal's h-index

The *h-index* could be applied in different level in applications such as rank the impact of academic journals (Braun et al., 2006). According to Braun et al. (2006), the *h-index* can be used for the case of research journals and article citations to be considered as a single measure of visibility of whole groups of publication units (each journal as a single unit). Bador and Lafouge (2010) carried out a comparative study between the *h-index* of journals and journal's impact factor and found that the ranking results are different. However, they still claimed that two indices could be complemented with each other for journal evaluations in the same discipline (Bador & Lafouge, 2010; Cuellar et al., 2008).

3.4 Altmetrics: the emerging metrics for impact evaluation

3.4.1 Emergence of altmetrics

The term "Altmetrics" was first denominated in 2010 by Jason Priem, a doctoral student in the School of Information and Library Science at the University of North Carolina. He used the hashtag "Altmetrics" on Twitter to name all online metrics of academic outputs, including social media networks (Roemer & Borchardt, 2015). He then cooperated with D. Taraborelli, P. Groth and C. Neylon to establish the website altmetrics.org, and published the “Altmetrics: a manifesto” claiming the origins, significance, value and future development paths of Altmetrics (Priem et al., 2010). “Scholars rely on evaluation filters to select the most relevant and significant academic literature today”, said Priem (Priem, Groth, et al., 2012). However,

the mainstream filters, such as peer reviews, citation counts, and journal impact factors, have drawbacks and limitations on time-consuming, gaming, and narrow measures. It is claimed that altmetrics are "tomorrow's filters", which can better reflect the impact of scholarship broadly and timely (Priem et al., 2010). This term then triggered an excellent study upsurge in the field of scientific research evaluation. The common concerns include, for example, the significance of the altmetrics concept (Galligan & Dyas-Correi, 2013), what they measure (Piwowar, 2013), and whether they can predict future citations (Barnes, 2015). There is no doubt that the appearance of altmetrics provides new research ideas and data sources for scientific research evaluation and increases the possibility of using multi-source data rather than citations only in quantitative research evaluation.

Traditional research evaluation methods such as peer reviews and citation-based metrics have been extensively applied in scientific research evaluation (Brinn et al., 2000; Guthrie et al., 2013; Penfield et al., 2014). However, their limitations are being criticised all the time. "We rely on filters to make sense of the scholarly literature", "but traditional filters are narrow and being swamped", said Priem et al. (2010, p.1). Peer review serves scholarship well, yet it is time-lagging and experts' energy-consuming (Goldbeck-Wood, 1999; Kumar, 2009; Priem et al., 2010). A large number of publications need to be appraised in peer reviews whilst the time and effort of experts are limited. Kumar (2009) points out a series of issues for peer reviews, such as personal bias on judgements, conflict of interest, delay in publication, cost of the process, and unspecialised selection of reviewers.

Citation-based metrics have been widely used in quantitative evaluation methods, yet they always have time-lagging problems (Zahedi et al., 2014). Publications need years to be cited to reveal the impact. Priem (2010) addressed that the *h*-index and other prevalent bibliometrics need even longer accumulated time to present the results. In addition, the influential work may remain uncited or less cited for an extended period. The examples are many so-called "sleeping beauty" articles (Ke et al., 2015).

Furthermore, journal evaluation indicators such as the Journal Impact Factors (JIF) have more critiques. For instance, the Journal Impact Factor is an average citation figure for the whole journal; thus, the impact of individual articles would be judged misleadingly (Dong et al., 2005). It has been argued that the JIF is easily manipulated relatively (Arnold, 2009; Bornmann et al., 2012; Reedijk & Moed, 2008). Reedijk and Moed (2008) summarised several editorial behaviours that help to manipulate the JIFs, which include encouraging (or forcing) self-journal citation, over citing specific articles such as "editorials", publishing distinct articles timely to fulfil a 2-year citing window, accepting more fashionable topics rather than qualified minorities and so on. Arnold (2009) revealed that mutual citations between journals and editors are common in his case study. He criticised the close connection between the *Journal of Chaos, Solitons and Fractals* and the *International Journal of Nonlinear Science and Numerical Simulation* as they have published 60 articles written by the same editor-in-chief, and the two journals cite each other frequently.

Moreover, the complex citing motivation is another crucial issue that may lead to a decline of authenticity in citation counting. Mingers and Xu (2010) found that the citation number is easily influenced by various factors (apart from quality), and sometimes a high citation number does not certainly represent a high impact or quality. Larivie're and Gingras (2010) investigated the citation difference when identical papers published in different journals and found that citation results from journals with high impact factors are twice those from journals with lower impact factors. In short, the citing motivation and quality of citation data were questioned.

Good news is that the study on quantitative evaluation of research impact in academia is not limited to the citation only. Using online metrics as alternatives to traditional metrics emerged in the 1990s. Cronin et al. (1998) raised that scholars' communication and collaboration started to be varied with the rapid emergence of online publishing. Many modalities of influence, such as the audience's ideas, thinking and general professional presence, are overlooked. Cronin et al. (1998) recognised a professor's impact through his name mentioned

in 11 identified outlets, including published articles, conference proceedings, external home pages, personal/parent organisation home pages, resource guides, etc. The digital traces the scholar left online reflect how academic influence is exercised and acknowledged, increasing the possibility of viewing influence from different perspectives (Xu, 2018). Holmberg and Thelwall (2014) pointed out that the emergence of the Internet has changed the formal and informal patterns of academic communication. The new communication platforms of research, such as social media networks and online reference management tools, may provide new indicators for measuring impact.

In 2009, the Public Library of Science (PLOS) launched the Article-Level Metrics (ALMs), which comprise the information of downloads, views, saves, discussions and citations of a paper, initiated the participation of usage data in the impact assessment (Neylon and Wu, 2009). It is believed that using such multi-source data would present a much fuller perspective of each article's impact and boost the emergence of new metrics (Tananbaum, 2013). The Altmetrics, proposed by Priem et al. (2010), claimed to be "tomorrow's filters" for selecting the most relevant and significant scholarly literature. Altmetrics present more diverse measures by tracing diverse activities in online tools and environments (Mingers & Leydesdorff, 2015). According to Galligan and Dyas-Correia (2013), the development of altmetrics framed with capturing broader impacts of academic papers based on crowdsourcing data in social media platforms such as Twitter, Facebook, blogs, online peer review systems, collaborative bookmarking tools and reference management services, which are beyond the traditional citation-based indicators. Altmetrics are recognised as the complement to the traditional citation-based metrics, providing an interesting option to measure impact beyond the science (Tahamtan & Bornmann, 2020). Bornmann (2014) also argued that using online information for alternative metrics starts with "webometrics" or "cybermetrics". The difference between webometrics and altmetrics is the content that is being analysed. The former mainly concerns web page analysis as text (e.g., blog analysis in Shema et al., 2015), and the latter emphasises usage data collected from common social platforms. Overall, various types of metrics have been added to the evaluation toolbox, and they are expected to complement the traditional

bibliometric indicators by presenting a much fuller perspective of research impact (Bar-Ilan et al., 2018).

Throughout the literature above, we could summarise that altmetrics emerge from two main aspects. On the one hand, researchers are exploring alternatives to the limitations and failings of traditional evaluation methods and indicators. On the other hand, the use of online tools and social media network for the research communication create massive amounts of online data, and these data can be used to form indicators, as they are countable now. Hence, the change in academic communication behaviours and shortcomings of traditional evaluation methods drives the study of altmetrics.

3.4.2 Definition of altmetrics

Although the relevant research on altmetrics has been in full swing, a universal definition of altmetrics is still lacking (Erdt et al., 2016; Haustein, 2016; Sugimoto et al., 2017). When the term altmetrics is proposed initially, Priem et al. (2010, p.2) address that “this diverse group of activities (that reflect and transmit scholarly impact) forms a composite trace of impact far richer than any available before. We call the elements of this trace altmetrics.” This representation is introducing altmetrics rather than defining it. Priem et al. (2012, p.1) further propose altmetrics as “the study and use of scholarly impact measures based on activity in online tools and environments.” Adie and Roe (2013) also define altmetrics from a study perspective, “the creation and study of new metrics based on the Social Web for analysing and informing scholarship”. In addition, as Erdt et al. (2016, p.1117) state, “Altmetrics is an emergent research area whereby social media is applied as a source of metrics to assess scholarly impact.”

From an approach perspective, Piwowar (2013b, p.159) states that “scientists are developing and accessing alternative metrics, or ‘altmetrics’ - new ways to measure engagement with research output.” Weller (2015, p.262) defines altmetrics as “the evaluation methods of scholarly activities that serve as alternatives to citation-based metrics”, specifying the measurements in social media environments.

Several scholars define the term from a measurement perspective and emphasise the metrics' scope in a social media environment. Haustein et al. (2014, p.1146) define that "Altmetrics, indices based on social media platforms and tools, have recently emerged as alternative means of measuring scholarly impact". Shema et al. (2014, p.1019) define altmetrics as "a term used to describe web-based metrics for the impact of scholarly material with an emphasis on social media outlets as a data source". Barnes (2015, p.122) states these new metrics are "based on the social web" for analysing and informing scholarship. Haustein (2016) criticises and suggests that the term should be replaced by social media metrics consequently. Besides, Rousseau and Ye (2013) believe that altmetrics are generated from the intersection of network and information metrics and should thus be named as influmetrics or web-based social influmetrics. Haustein et al. (2016) further propose a definition of altmetrics with an emphasis on data accessibility, "events on social and mainstream media platforms related to scholarly content or scholars, which can be easily harvested (i.e., through APIs), and are not the same as the more 'traditional' concept of citations" (p.374).

Bornmann (2014) recognises altmetrics as the complement to the traditional citation-based impact measurements, which are considered an interesting option for assessing the societal impact of research. Galligan and Dyas-Correia (2013) expect altmetrics to provide an alternative measure of impact to distinguish the Journal Impact Factor, "which has been categorically misused and fails to respond to the digital environment that scholarship takes place in today" (p.56). Indeed, altmetrics, outside the scope of traditional measurements, expect to quantify web-driven academic interactions, such as how research output is tweeted, blogged or bookmarked (Howard, 2013). Altmetrics also include the usage of data such as viewing and download statistics, although such data have been accessible longer than web-based media applications (Haustein et al., 2016). Altmetrics provide more interpretations of specific impact, which differ from the citation-measured impact from an academic perspective but more like the online impact of research (Zahedi et al., 2014). Citation count is a valuable sign of conventional scholarly impact, while altmetrics can tell reactions of public

and non-academic insightful impact, which can be used to predict later citations. Thus, both measurements are needed to prove a comprehensive influence of research (Stacy et al., 2016). It seems altmetrics are complementary rather than an alternative to citation-based indicators currently, and they may outlive the traditional metrics subjecting to the standardising states of the new metrics (Galligan & Dyas-Correia, 2013; Haustein, 2016).

Through the above literature, the main differences in current definitions are to define the altmetrics from which perspective, for instance, from the perspective of study, method, and measurement. In this thesis, we focus on the measurement perspective of the altmetrics. From the measurement perspective, a common understanding is that altmetrics are new or alternative metrics for measuring the impact of research output, which differ from traditional metrics like citations (Erdt et al., 2016; Haustein et al., 2016; Shema et al., 2014). The arguments, however, are where altmetrics originate —activities in online tools and environments (Priem et al., 2012), based on social media platforms and tools (Haustein et al., 2014), social and mainstream media platforms (Haustein et al., 2016), or social media outlets (Shema et al., 2014), and what impact they measure — reflecting scholarly impact (Priem et al., 2012; Haustein et al., 2014) or more interpretations of research impact (Bornmann, 2014; Zahedi et al., 2014).

In this thesis, we apply a clear and straightforward definition of altmetrics based on a measurement perspective: “altmetrics are indices based on social and internet media platforms and tools (Haustein et al., 2014, p. 1145)”. Regarding the impact that altmetrics could reflect, we hope altmetrics and traditional metrics could be applied together to demonstrate a much fuller impact of research output (relevant research, see Chapters 6 & 7).

3.4.3 Advantages of altmetrics

Altmetrics could provide several potential advantages. Piwowar (2013) believed that altmetrics is superior in terms of timeliness, diversity of indicators and diverse audience compared to traditional citation-based metrics. Wouters and Costas (2012) advised four arguments about the advantages of altmetrics, including the diversity of sources and filtering,

speed, openness, and usage in measuring scientific performance. Most literatures addressed the advantages of altmetrics from the following aspects:

1. Broadness

Bornmann (2014) referred to "broadness" as a broader impact beyond science that altmetrics could measure. Altmetrics are expected to provide new insights into the impact of research and change the situation by only using citation-based indicators to explain the scientific impact (Konkiel, 2013; X. Wang, Mao, et al., 2014). Zahedi et al. (2014) were concerned that altmetrics could represent the online impact of research work because of its appearance on social media and online platforms. Moed and Halevi (2015) believed that altmetrics might provide a broader public with some perspectives on research popularity and research interest. Tahamtan and Bornmann (2020, p.17) pointed out, "altmetrics may capture a different aspect of societal impact (which can be called unknown attention) to that seen by reviewers (who are interested in the causal link between research and action in society)." Furthermore, when deciding on fund support, some scientific institutes request that the candidate project provide information on the potential impact that might be generated (Stacy et al., 2016). Bornmann (2014) revealed that the US National Science Foundation has criteria for societal impact potentials for the candidate applicant. It is argued that altmetrics are expected to become the supplementary indicators for justifying research project budgets or as a reference to support decision-making on project funding, recruitment and promotions for research institutes (Bornmann, 2014). Piwowar (2013) explained the broad impact from the perspective of audiences. The audiences of indicators should be not only the scholars but also "the practitioners, clinicians, educators and the general public". It emphasises the broad users who may benefit from the research findings. Mohammadi et al.(2015) also mentioned that clinical medicine papers might quickly generate an impact in clinical practice because they are favoured to be read by everyone. However, only scholars who cite the paper could prove the academic impact generation, whilst possible clinical practice's impact is difficult to trace. Therefore, altmetrics may indicate this kind of broad impact complementally (Mohammadi et al., 2015).

2. Diversity

Altmetrics are diverse in data and assorted data sources. Altmetrics track the online activities of scholarly product on various platforms, which include online reference management software such as *CiteULike*, *Mendeley* and *Zotero*, social media networks such as *Twitter* and *Facebook*, scholarly social network such as *Research-Gate* and *Academia.edu*, academic and non-academic Blogs, online encyclopaedia such as *Wikipedia*, and more different online tools for research communication (Priem et al., 2012). The diverse group of online activities forms a composite trace of impact far more affluent than the traditional citations and cited sources (Priem et al., 2010). Konkiel (2013) addressed the diverse features of altmetrics from the perspective of the user group (i.e. scholars, researchers, students, librarians, policymakers and institution managers) and the types of engagement in scholarly context (i.e. views, saves, discussions, formal references and recommendations). Konkiel (2013) also believed the diversity characterised the "impact flavours". Piwowar (2013) pointed out that altmetrics are diverse in audience, i.e. not only scholars but also medical practitioners, clinicians, teachers, students, and the public. Bornmann (2014) stated that altmetrics offer the opportunity to evaluate a greater diversity of scholarly products where citation could not apply, for example, the datasets, copyrights, software, algorithms, videos, grey literature, slides, as well as a variety of scholarly services such as teaching, online courses, reading lists and more. In cases where the citation does not apply, there may be more aspects of using altmetrics to determine the impact.

3. Speed

One primary advantage of altmetrics is that they gather timely reactions to new scientific research outputs as soon as they are released, even before the formal publication (Patthi et al., 2017; Weller, 2015). According to Priem, Piwowar and Hemminger (2012), the collection of altmetrics data is based on open programming interfaces (APIs), and many social media networks and online tools offer free access to these real-time data. Hence, the impact of a scientific work can be tracked immediately after its publication. Citation

count is a reliable and effective measurement for evaluating academic impact. However, its most significant limitation is time lagging. One altmetric data, Mendeley readership, has been validated to be significantly positively correlated with the citation count, and it starts to accumulate data right after the publication of the research output (Bornmann & Haunschild, 2016a; Li et al., 2012; Mohammadi & Thelwall, 2014; Mike Thelwall & Wilson, 2016). Citing and waiting to be published take years; therefore, altmetrics can be used as substitutes in some informal evaluations (Mohammadi & Thelwall, 2014). Furthermore, micro-blogging social media platforms such as Twitter and Facebook may assist scholars in publishing research findings before completing a paper to gather prior reactions from peers and readers before publishing (Darling et al., 2013). Additionally, preprint papers can gather early and timely comments as well. Although speed has been seen as a characteristic feature of the altmetrics, accumulating velocity of each altmetric data differs (Fang & Costas, 2018). Fang and Costas (2018) classified altmetrics data with fast and slow sources and argued that the events labelled as slow sources should not share the usefulness of speed.

4. Openness

Altmetrics collects data through open Applications Programming Interfaces (APIs), and integrating data and sources from different online platforms becomes easy when more data providers agree to access their specific APIs freely (Priem, Piwowar, et al., 2012). The format of altmetrics supports a trend of open data for every user. Besides, altmetric data comes from definite software or platforms such as reader count in Mendeley or likes in Facebook, which has clear boundaries and data types to help with data analysing and the interpretation of findings (Bornmann, 2014). Wouters and Costas (2012) argued the openness of altmetrics from a more political perspective, the balance between openness and transparency. As Wouters and Costas (2012) illustrated, Google Scholar is more open and accessible for scientific work searching, while the Web of Science database requires subscriptions, and the reality is that the Web of Science has a more transparent construction of the database and reliable indicators than Google scholar does. Despite the fact that openness is a valuable resource, free online-based tools do not always have

superior performance than those profitable data services. If using altmetrics as a research evaluation tool, transparency and reliability of data might be a higher priority than free accessibility to beat the dependency on current commercial databases (Wouters & Costas, 2012).

3.4.4 Disadvantages of altmetrics and research challenges

Altmetrics has drawbacks, just as the benefits they offer. Bornmann (2014) pointed out four main disadvantages of altmetrics: commercialisation, data quality, missing evidence and manipulation. Haustein (2016) discussed the challenges of altmetrics in three aspects: heterogeneity, dependency and data quality. Adie (2014) proposed more practical issues for future study, for instance, a clear understanding of altmetrics and how they ought to be used, ongoing development of altmetrics tools and practice within the community, and the availability and reliability of altmetric data. The following paragraphs will discuss these arguments in detail.

Firstly, regarding the lack of systematic theory, Priem (2014) believed that the study of new sources for impact measurements (i.e. patents, affirmations, mentorships, and academic hyperlinks) usually started before developing foundations for hypothetical theory. A systematic theory of altmetrics ought to be needed for the new field, yet optional in the early stage of the study (Cronin & Sugimoto, 2014). Haustein (2016) also believed that heterogeneity, one feature of altmetrics, leads to difficulty in building up a universal definition or theoretical framework due to the variety and complexity of social media acts that altmetrics are based on. Besides, it is unclear whether altmetric data and indicators obtained from different media sources could reflect academic influence, public interest or something else.

Secondly, the ease of manipulating or gaming of altmetric count has been mentioned by several papers (Cronin & Sugimoto, 2014; Erdt et al., 2016; Gamble et al., 2020; Patthi et al., 2017; Thelwall, 2020). Manipulation in traditional metrics has been discussed for a long while. For altmetrics, manipulating is even easier. Altmetric data count could be inflated through

numerous anonymous online accounts without difficulty, so gaming is a prominent issue on the negative side of altmetrics. It is suggested designing an effectively recognising mechanism to restrain gaming as patterns are easily identified in big data environments (Galligan & Dyas-Correia, 2013), whilst Gamble et al. (2020) believed that not all suspicious actions are detectable. Gamble et al. (2020) revealed that there is 2.4% of tweets were carried by bot accounts for tweeting dental journals in the US. Moreover, another area to be easily gamed is to vote "like" for a tweet or other online material (Gamble et al., 2020). It is not easy to identify that a "like" supports the original paper's content, arguments in the comment, research idea or something else. The meaning of this element is very uncertain and easily to be gamed. Bornmann (2014) also criticised the social media networks today that lack quality control and processes for linked users when offline, which would produce high altmetric scores for articles.

Thirdly, altmetrics are prone to biases towards younger scholars with more web visibility (Galligan and Dyas-Correia, 2013). It is argued that many researchers rarely utilise online media, so bias potentially exists within the measurements. Conversely, Priem et al. (2011) found that the PhD students are not more dynamic than university lectures on Twitter. Disciplinary difference is also a factor for bias. Cronin and Sugimoto (2014) suggested that in fields such as the digital humanities, more academics use Twitter and generate a higher number of tweets than the conservative technical disciplines.

More importantly, data quality issues are becoming more significant in recent years. Several papers have addressed the concern about data accuracy, consistency, reliability, replicability, data coverage, and durability (Bornmann & Society, 2014; Haustein, 2016; Konkiel, 2013; Ortega, 2018b; Zahedi & Costas, 2018). Altmetric data are more accessible to obtain through open APIs. In contrast, different times of collection, accessibility of particular papers, or varied data collection methods can lead to inconsistency among data collectors. As altmetrics are especially subject to the accessibility of APIs and DOIs, they would emphatically rely on data suppliers and aggregators, conceivably affected by the technical limitations of operating

platforms (Haustein, 2016). Indeed, altmetric data rely upon the characteristics of the datasets and the frameworks where they are filed. According to Erdt et al. (2016), using the correct search query is also important, as not all research publications have DOIs. Besides, different sites and platforms may have different URLs for the same research paper. Hence, only relying on DOI to track data is not accurate. However, resolving all links related to the research output could be more challenging (Erdt et al., 2016). Liu and Adie (2013) were concerned with data collection issues such as missing links or duplicated content. It is particularly challenging to capture meta-data in multi-media sources such as recordings or videos, which do not have DOIs or links but only verbal content.

We can see that the data quality and reliability issues are becoming more critical as the use of altmetrics in research evaluation grows. It is essential to understand the extent of data reliability of altmetrics before using it as an evaluating tool. Therefore, we will consider these issues in the following research and provide our analysis (Chapter 4).

3.4.5 Difference between altmetrics and traditional metrics

Traditional evaluation methods for research impact mainly referred to qualitative approaches such as peer reviews and quantitative metrics such as the journal impact factors and the *h*-index (Aung et al., 2019; Ezema, 2017). Priem et al. (2012) claimed that the traditional methods for selecting significant literature are failing, as the volume of academic material is surging today and scholarly communication has become moved online. Altmetrics, which provide alternative measures of impact through tracking the readership, transmission, and use of scholarly publications in online tools and social media, are promising metrics (Priem et al., 2012). Compared to the traditional methods, altmetrics have the potential to measure research influence faster, real-timely, and widely cover diverse research outcomes, sources and a broad range of non-academic audiences (Ezema, 2017).

As discussed in Section 4.1.1, traditional metrics suffer from the time-lagging issues. Aung et al. (2019) argued that the time to accumulate citations for a piece of scholarly work could be years, obstructing the timely measurement of impact. Altmetrics are usually timely data

indicating proof of impact in days rather than years (Piwowar, 2013). Altmetrics collect interaction data of scholarly output from online and social media tools such as Mendeley, ImpactStory, Twitter, Facebook, blog post and so on immediately after the interactive events happened, which makes estimation of influence quicker and shorter. Altmetrics could measure the impact of a scholarly work just a couple of days or weeks after it has shown up (Mohammadi & Thelwall, 2014). On the other hand, Galligan and Dyas-Correia (2013) believed that abundant bookmarks and discussions implied great public support. Altmetrics can supplement the traditional method of peer review prospectively and takes less time. As Priem et al. (2010) addressed in their manifesto, the speed of altmetrics increases the chances of creating real-time recommendations and synergistic sifting systems.

According to Galligan and Dyas-Correia (2013), the traditional metrics count citations in academic publications, whilst the altmetrics count the number of mentions in blogs, Twitter and other online media, which are non-publishing carriers. Hence, altmetrics have different dimensions of measurement compared to the traditional methods. The traditional methods also ignore many non-journal forms of research outputs such as scientific blog posts, slideshows, datasets and scholarly dialogues; in contrast, altmetrics measure mentions in new forms of research output that jump out of the journal level (Galligan & Dyas-Correia, 2013). Additionally, altmetrics determine the impact of scholarly work by tracking its relevant events on social media platforms; such measuring perspective is novel and allows discovering new insights into the impact (Ezema, 2017).

Another significant difference is that journal metrics are refereed. The current commercial database such as Web of Science (WoS), safeguards the quality and reliability of their indicators based on referees of core journals (named as the core collection on its platform¹). The traceable path of referees increased the transparency and reliability of data quality. However, there is no such core referee for the altmetrics. The quality of a paper is no doubt the essence of the research assessment, whilst altmetrics could not determine the quality of

¹ <https://clarivate.com/webofsciencegroup/solutions/web-of-science/>

a scholarly output as competent as traditional metrics do. Although the research on altmetrics is increasing, the absence of normalisation across datasets and framework impede its generalizability (Konkiel, 2013).

Lastly, altmetrics measure more dimensions of impact than citation-based metrics. Several empirical studies have shown that Mendeley's readership is positively and moderately correlated with citation, reflecting a certain value of academic impact (Li et al., 2012; Torres-Salinas et al., 2017; Zahedi et al., 2014). Mohammadi et al. (2015) investigated Mendeley's user categories and believed Mendeley could monitor students' reading preferences. Therefore, it probably reflects the educational value of research outputs. Mohammadi et al. (2016) further conducted a survey of motivation for using Mendeley's bookmarking, and the results indicated that Mendeley's bookmarking count might reflect a combination of scholarly impact and wider educational impact.

Altmetrics could measure social engagement by capturing discussions of scientific papers online (Bornmann, 2014). Robinson-Garcia et al. (2018) believed social media metrics could help identify the social interactions between researchers and people outside the scientific field so that they have the potential to assess societal impact. Pulido et al. (2018) carried out a content analysis of posts on Facebook and Twitter and only found minor evidence to prove the societal impact, which is not sufficient to say social media metrics are good indicators of social impact. Although there are no constant opinions on what altmetrics are measuring, the idea that social media indicators provide different dimensions of measurement for research impact (Costas et al., 2015; Jobmann et al., 2014; Martín-Martín et al., 2016; Pulido et al., 2018). Altmetrics may capture a different aspect of societal impact, such as public engagement (Khazragui & Hudson, 2015), popularity of research (Xia et al., 2016) or unknown attention (Tahamtan & Bornmann, 2020). To sum up, altmetrics measure a wider impact beyond science, possibly complementing the citations to show an integrated impact of research output.

3.5 Altmetrics platforms and data sources

3.5.1 Terms and classification

The altmetrics, as an emergent research field, covers various social media applications as data sources to assess scholarly impact (Erdt et al., 2016). Before understanding what kind of data, sources and hosts are contained in the scope of altmetrics, it is necessary to interpret the relevant terms clearly. Table 3.1 lists the interpretation of generic terms of altmetrics illustrated by Haustein et al. (2016). In addition, examples of the terms have been added to the table.

Table 3.1 Term interpretations in altmetrics (Haustein et al., 2016)

Term	Interpretation	Examples
Research object	A research output or research agent for which an event can be recorded	journal article, conference paper, patent, individual scholar
Event	a recorded act/activity related to the research object	download, bookmark, tweet
Host	the place where research objects are made available and exposed to potential events;	Erdt et al. (2016) remark that 'hosts' often act as 'data sources'. For example, Mendeley recorded the readership of papers.
Source	a platform where events are available	Mendeley, Twitter
Consumer	a party that collects and uses events to research objects	Publisher, altmetrics provider, researcher
Aggregator	a type of consumer who collects and provides events to research objects with a specific methodology	Altmetric.com, ImpactStory, PLOS ALM, Plum Analytics
End-user/ Audience	a type of consumer who uses and applies events in a specific context and intention	Researcher, librarian, research institute, funding body.

According to Haustein et al. (2016), altmetrics are based on events created on social media platforms, and those events are formed by a series of acts (e.g. saves, reads, downloads) that are related to the research object (e.g. scholarly output). Lin and Fenner (2013) grouped the acts relevant to the research object into five categories based on types of engagement: *viewed*, *saved*, *discussed*, *recommended* and *cited*. When PLOS introduced the Article-level-Metrics (ALM), their indicators were sorted into five categories. Haustein et al. (2016) classified the acts that led to the online events into three categories: *accessing*, *appraising*,

and *applying*, and argued that these three categories have covered all stages and aspects of use and interaction with the research outputs.

Regarding the other classification of altmetrics, Glänzel et al. (2019) classified the existing online tools from the functional perspective, which includes *social bookmarking tools*, *microblogging tools*, *blogs* and *encyclopaedia*. Liu (2016) addressed several levels of altmetrics based on the types of research output, such as journal paper level, book levels, databases and software levels, video levels, conference presentations and poster levels and so on. Jin et al. (2019) classified the source of altmetrics into two categories: original data sources and data aggregating tools. The original data sources category includes direct altmetric data sources such as Google+, Facebook, etc. Data aggregating tools such as Altmetric.com and PLOS ALM provide aggregated altmetrics from other data sources. The characteristic of data aggregated tools is that they do not generate data themselves whilst providing integrated altmetric data and analysis.

3.5.2 Altmetrics aggregators

Over the past few years, several altmetrics aggregators emerged to serve as altmetrics providers. One could gather sources independently or capture data from other hosts to become a secondary or tertiary altmetrics provider or aggregator (Erdt et al., 2016). The aggregator focuses on the host that collects different events and aims to aggregate new alternative metrics. In contrast, some hosts, such as Mendeley, which generates events locally, would be classified as sources or data sources. Below, a few popular altmetrics aggregators in use are introduced.

(1) *Altmetric.com*

Altmetric.com (<http://www.altmetric.com/>) was founded by Euan Aide in 2011. It tracks online activities related to a research output from blogs, Facebook, Twitter, major news and other different social media networks, and calculates the impact of the research output into a score, which is called "Altmetric Attention Score" based on weighting the data gathered

from different social media source². Altmetric.com now provides its products to publishers, institutes, individual users and researchers (through *the altmetric bookmarklet*) to check the attention score for the relevant scholarly content and all attention data collected. Several famous publishers such as Springer, Taylor & Francis, Wiley and Elsevier have illustrated altmetric score and altmetric data of published papers on their websites (Tahamtan & Bornmann, 2020). According to the official website of Altmetric.com, the impact of more than 9 million academic documents by the end of 2018 has been tracked. The current data sources that Altmetric.com traced include Twitter (named as X from 2023), Facebook, Policy documents, News, Blogs, Mendeley, Scopus, Post-publication peer reviews, Reddit, Wikipedia, Q&A (Stack Overflow), F1000Prime recommendations, YouTube, Open Syllabus, Web of Science and Patent Citations (Wu et al., 2016).

(2) PlumX

Plum Analytics (<http://www.plumanalytics.com>) was founded by Andrea Michalek and Mike Buschman in 2012. The vision of Plum Analytics is to provide modern and timely indicators collected from social media networks, open access repositories, and other online tools to determine the valuable research output, aiming to break through the dependence of traditional evaluation systems on citation count³. Plum Analytics was acquired by EBSCO in 2014 and then Elsevier in 2017. Now, it has embedded in the Scopus literature database to present available metrics alongside the Scopus documents⁴. PlumX metrics (hereinafter referred to as PlumX) contain a comprehensive set of article-level metrics to demonstrate how people interact with academic outcomes online beyond the traditional citation metrics. Plum Analytics also provides a customised service of altmetrics for paid institutions. According to the figure on its website, Plum Analytics has collected about 52.6 million literature data by the end of 2018.

² www.altmetric.com

³ <https://plumanalytics.com/about/>

⁴ <https://www.scopus.com/>

(3) PLOS

The Public Library of Science (PLOS) is a non-profit open-access scientific publishing program that aims to create an open-access library of journals and other scientific literature under the open content license. It started implementing Article-Level Metrics in March 2009 to measure the performance of research papers published by PLOS (Lin & Fenner, 2013). PLOS provides five categories of metrics. The indicators in the 'viewed' category refer to the activities of users retrieving online papers, including HTML web browsing, PDF downloads and XML downloads. Meanwhile, it combined with PubMed Central to provide metrics such as page views, downloads and Figshare browsing data. The 'saved' category includes saving a paper to an online literature management tool. The other categories are 'discussed', 'recommended' and 'cited'.

(4) ImpactStory

ImpactStory (<http://impactstory.org/>) is a non-profit and open-source network platform that aims to help researchers discover and share the online impact of their research⁵. Heather Piwowar and Jason Priem founded the ImpactStory, which is currently receiving funding from the National Science Foundation (NSF) and the Alfred P. Sloan Foundation. The ImpactStory has established a new academic reward system to encourage researchers to reveal their research achievements via online data. The data sources used by ImpactStory include Twitter, Mendeley, CiteULike, Delicious, F1000 reviews, SlideShare, GitHub, and the metrics collected from altmetric.com, including blogs, Facebook, Google+, Wikipedia and other online sources (Galligan, 2012). It is argued that ImpactStory demonstrates well the impact of individual researchers as it has a high coverage of data on Twitter, particularly (Kraker et al., 2015).

Additionally, ImpactStory distinguishes the impact measurement from the perspective of the audience. It separately measures research outputs' scholarly and public impacts (Peters et al., 2014). The scholarly impact measurements refer to the indicators derived from specific platforms, for example, Mendeley and Scopus, and the public impact measurements are

⁵ <https://profiles.impactstory.org/>

those indicators demonstrating a broader interest of the public, such as the count of Twitter, Facebook, or Wikipedia.

(5) Crossref Event Data (CED)

Crossref Event Data (CED) (www.eventdata.crossref.org/) is a Crossref service that provides unique records of the relationship between a research publication and its relevant online events (academic and social activities). CED aims to capture the online activities for specific scholarly content such as publications' discussing, sharing, and commenting on social webs and then distribute these data to various users. According to the CED official website, event data has several distinctive purposes: 1) serving scholars to understand how their research outputs have been used and mentioned; 2) helping readers to access more information about published articles; 3) to understand the research impact beyond citations; 4) more users such as publisher, service provider, or data analysis organisation can access a variety of sources that provide opinions on research articles and use their tools to analyse, interpret, and report data.

In CED's broad sense, anyone mentioning a research publication with a registered DOI at any place online is considered an event. However, CED could not record all the online events due to its limitation on tracking websites and the way of tracking (only the publication's DOI or URL could be tracked). CED now collects raw data from a selection of sources and agents, including Crossref metadata, DataCite metadata, Faculty Opinions, Newsfeed, Hypothes.is, Wikipedia, Twitter, Reddit, Reddit links, Stack Exchange Network and Wordpress.com⁶. It is seen that the sources collected by CED are still limited. CED's own explained that some resources did not have public access, and some academic sources were not easy to establish a suitable agent. We must note that CED only provides raw data without integration or metrics, so how to use these data depends on the users. Since all data is automatically collected, the CED has referred that errors, delay, and linkage mistakes may be included in the original data, and data quality cannot be guaranteed sometimes for specific use cases⁷.

⁶ <https://www.crossref.org/services/event-data/#00354>

⁷ <https://www.crossref.org/services/event-data/#00354>

(6) Dimensions

Dimensions (<https://www.dimensions.ai/>) is a new scientific information platform launched by Digital Science in January 2018. The Dimension platform aims to create an open and powerful "research data tool" for users to explore all linked data of a research lifecycle in one single platform. Dimensions collect various types of academic content data within a research lifecycle, for example, research grants, datasets, publications, policy documents, clinical trials and patents, so that it is possible to trace research from the beginning of the research idea to the generation of research impact⁸. The Dimensions covers over 120 million research publications from 87,000 journals, more than 1.5 billion citations, 11 million datasets, 6 million grants, 188 million online mentions, 743k policy documents, 671k clinical trials, 140 million patents etc.⁹ It is said to be the world's largest linked research information dataset. The Dimensions now has five main products: the accessible version of the Dimensions platform, Dimensions Analytics, Dimensions Profiles, Dimensions on Google BigQuery and Dimensions Lift Sciences & Chemistry, and aims to serve researchers, academic institutes, publishers, government and funders, corporate research and development for understanding the research with a broader picture of life cycle and comprehensive impact.

The Dimensions platform is not limited to the publication data as most platforms do but is more focused on the data arising in the research process. Williams (2018) outlined the Dimensions' data into pre and post-publication stages. For instance, the former category includes data on research (idea), grants, conferences, and data sets, and the latter includes tweets, blogs, citations, clinical trials, patents, etc. The wide range of data enrichment processes could help to track and understand a research cycle and assist the research results to become more comprehensive and meaningful. The good news is that these process data are easy to reach for Dimensions' users. It is free to check information through web interfaces, and metrics API are made accessible to construct off the batch data.

⁸ <https://www.dimensions.ai/why-dimensions/>

⁹ <https://www.dimensions.ai/>

Orduña-Malea & Delgado-López-Cózar (2018) produced a coverage analysis for Dimensions and found that the annual growth of scientific literature (number of journal paper records per year) is currently higher in *Dimensions* than in *Scopus* and *Web of Science Core Collection*. *Google Scholar* has a slightly higher coverage of papers than *Dimensions*, whilst its annual growth of papers is relatively unstable between the years 2012 and 2018. For the citation counts of the analysing sample, it is found that Dimensions' record is slightly lower than that of Scopus. However, the metrics correlations at three levels, including author, document, and journal, are very high between *Dimensions* and *Scopus*. Orduña-Malea & Delgado-López-Cózar (2018) believed that the Dimensions combines the favourable functions of classic bibliographic databases (e.g. availability of metrics, search filters, and sort results functionality) with some of the characteristics of Google Scholar (e.g. superior coverage, simple search tool, fast and precise, and free) so that it is an emerging platform with potential. Bornmann (2018) questioned the reliability and validity of the field classification system on Dimensions, which used machine learning techniques to assign the field of a paper rather than the traditional method on journal disciplines. The results showed that most papers in the sample had been misclassified. Thus, the new feature of Dimensions on field classification may lead to a mistaken result, especially the field-normalised citation scores demonstrated in the Dimensions (Bornmann, 2018). More validation studies on Dimensions would be needed, and this new scholarly search database does broaden the use of scholarly information from a new perspective.

(7) Mainstream publishers

It is worth noticing that many world-leading journal publishers such as Springer, Elsevier, Taylor & Francis, Wiley, PLoS, Sage, IEEE, Science and Nature both presented altmetrics (or Altmetric Score) for their published articles. Mainstream publishers attach importance to altmetrics, reflecting the usefulness of altmetrics for impact measurement (Tahamtan & Bornmann, 2020; Mike Thelwall, 2020). Erdt et al. (2016) classified the publishers to be altmetrics aggregators as they demonstrate the different usage of events. In 2015, Springer and Altmetric.com cooperatively developed the platform "Bookmetrix", which provides the metrics of *Downloads*, *Readers*, *Mentions* and *Citations* for Springer's publications in book

level and chapter level¹⁰. Publishers such as Taylor & Francis Group, SAGE and IEEE embed the Altmetric Score directly alongside the link of papers on their websites, including page views, downloads and citations together for readers. The publishers' conversion may reflect that measuring the impact of a journal paper by citation alone is already changing.

3.5.3 Data sources used in the altmetrics

Various data sources are famous for being tracked by altmetrics aggregators. The collection of data sources is determined by each aggregator according to its strategy and designed indicators (Jin et al., 2019). A wide range of data sources can make alternative metrics diversified. Barnes (2015) listed a few popular forms of data sources that altmetrics trace, including (1) micro-blogging or short-message services (e.g. Twitter), (2) social networking sites (e.g. Facebook, Research Gate), (3) blogs (e.g. WordPress, Blogger), (4) social bookmarking networks (e.g. Delicious), (5) academic bookmarking platforms (e.g. CiteULike, Mendeley), (6) peer review services (e.g. F1000Prime), (7) academic networks (e.g. Academia.edu), and (8) collaboratively edited online encyclopaedias (e.g. Wikipedia).

According to Wouters et al. (2019), the typical social media platforms used as data sources include Twitter, Facebook, various Blogs, News, Wikipedia, YouTube, Policy documents, etc. Some data sources had been used before, whilst gradually discontinued in later days, such as LinkedIn and Connotea. Many data sources are rarely used, such as Flickr, Technorati, Sina Weibo (a Chinese microblog), and data from different organisations, for example, CRIS (Current Research Information System), WIPO (World Intellectual Property Organisation), QS (Quacquarelli Symonds—a global provider of specialist higher education and careers information and solutions), SSRN (Social Science Research Network), ADS (Astrophysics Data System), bit.ly (a URL shortening service), and COUNTER (Counting Online Usage of Networked Electronic Resources) (Erdt., 2016; Wouters et al., 2019).

¹⁰ <https://www.digital-science.com/press-releases/springer-and-altmetric-to-launch-bookmetrix-a-new-platform-for-book-impact/>

The following sections review a few data sources (or platforms) that commonly used among altmetrics aggregators.

Mendeley

Mendeley is a well-known reference management software that helps users manage and share research papers and engender bibliographies. It was founded in 2007 and then acquired by Elsevier in 2013¹¹. According to Mendeley's official website, the latest version of the software is able to plugin citations in a desired style (e.g. APA, IEEE, etc.) in a Word document, organise and read all references in own created library, and collate all personal comments and highlights of multiple PDFs on Mendeley Notebook¹². It offers "readership" statistics which capture the number of Mendeley users who have saved a given publication (together with their academic statuses, countries and disciplines). Mendeley is also a scientific networking platform where scholars can connect with peers online and collaborate easily (Rodgers & Barrow, 2013).

Although several other reference management tools, such as RefWorks, Zotero, and EndNote, are also popular, Mendeley has the highest coverage in use (Tahamtan & Bornmann, 2020). Mendeley software is free for users, which progresses the widespread usage. . The readership data is also easy to gather via open API so that many altmetric platforms provide Mendeley Reader Count as a readership metric (Alhoori & Furuta, 2014). Zahedi et al. (2014) studied the different altmetric data collected from Impact Story and found Mendeley is the major source for altmetrics. Mendeley readership data covered 62.6 % of all papers studied, whilst the other source such as Twitter (1.6%), Wikipedia (1.4%) and Delicious (0.3%) has much lower data coverage. Zahedi et al. (2015) compared Mendeley.com, Lagotto and Altmetric.com with 30,000 random DOIs and concluded that Mendeley is the highest altmetric event with a coverage of 69% for all studied publications.

¹¹ <http://www.mendeley.com>

¹² <https://www.mendeley.com/reference-management/reference-manager>

It needs to be noted that Mendeley's coverage varied by discipline. Mohammadi and Thelwall (2014) analysed the extent of correlation between Mendeley readership and citation count in different disciplines and found that social science subjects had a higher correlation coefficient than humanities overall. The science-related discipline has the highest correlation coefficient among their research, while lower correlations are of traditional humanity subjects. However, they argued that the current correlation result could not demonstrate Mendeley's readership measure in the same aspects of research impact as citations do. Mendeley may capture a wider range of academic activities from different readers' perspectives than citation counts (Mohammadi & Thelwall, 2014). Mohammadi et al. (2015) further compared the Mendeley coverage in Physics, Engineering, Chemistry and Technology and concluded that papers in Clinical Medicine have the highest coverage in their study.

Blogs

Blogs, as a new form of web-based media, emerged in the late 1990s, and their expanding speed has been startling (Garden, 2012). It is argued that Blogs flourished rapidly after their emergence because it allows individuals to post personal opinions and analysis freely and engage in online conversations easily (Kenix, 2009). According to Fausto et al. (2012), many prestigious scientific journals such as Nature, PLoS and Guardian have research blogs to release the latest scientific news, and famous scholars such as Paul Krugman and Gary Becker, the Nobel Laureates in economics, both run popular blogs. The motivation behind research blogs mainly includes influencing their audiences, sharing findings, knowledge and skills and exchanging ideas about their research (Shema et al., 2015).

Fausto et al. (2012) believe blogging effectively allows scientists to exchange ideas with peers or the public. It could promote research production through scientific information sharing and discussion and lift the dissemination of complex scientific findings through simplification and explanation. Indeed, general readers without expertise may feel obscure to specific academic knowledge, and blogs serve as a bridge. Upon the reduction of the public's barriers to scientific expertise, blogs also bring more practical information services to the public. Shema et al. (2015) discussed the popularity of medical blogs and addressed that the attention from

mainstream media and the public also boosts medical experts and skilled people to use blogs to discuss health issues.

Shema et al. (2014) studied the correlation between a journal article's "blog citations" and the Web of Science citations. "Blog citation" means if a paper is reviewed and discussed in ResearchBlogging.org, a peer-reviewed blog site, it receives a "blog citation". Their statistical results supported the hypothesis that the articles with "blog citations" at ResearchBlogging.org receive more WoS citations than those articles in the same journals without "blog citations". Therefore, they proposed that blog citations could be considered as an alternative metric source.

Glänzel et al. (2019) summarised that blogs, particularly scientific blogs and blogging tools focused on peer-reviewed research, are emergent platforms for disseminating scientific materials to academia and the general public. Blog mentions of scholarship (for example, mention of researchers or academic institutes) and blog citations (mention of publications) are the main metrics that can be obtained from these platforms. In addition, possible available data could be gathered through the web interface or API (Glänzel et al., 2019).

Twitter

Twitter is a social networking service where people can share their thoughts and information by posting comments, photos, videos and links. The information people share on Twitter sometimes includes scholarly content, for example, papers and conferences, research news, and research ideas, which provide an opportunity to use these data to evaluate the impact of related scholarly works.

It is claimed that Twitter has an average of 330 million monthly active users and creates more than 400 tweets each day¹³. Barthel et al. (2015) believe that Twitter is a very useful data source for altmetrics because it contains links to academic papers and they are traceable. The context information in tweets can be further used for content analysis. Haustein et al. (2015)

¹³ <https://www.statista.com/statistics/970920/monetizable-daily-active-twitter-users-worldwide/>

showed that the coverage of articles on Twitter was 21.5%, followed by other social media platforms such as Facebook (4.7%) and blogs (1.9%). A Twitter user survey by Mohammadi et al. (2018) showed that 45% of the respondents who have tweeted journal articles do not work in academia. The large number of non-academic users who tweeted scholarly papers on Twitter provides evidence to support using tweets count to measure the non-academic impacts of scholarly research. In an analysis of 1968 tweets on research papers from the top 70 journals in the discipline of Psychology, Htoo & Jin-Cheon (2017) have found that the tweets from non-academic users are twice as many as tweets from academic users.

However, there are issues and limitations in using Twitter data to measure the societal impact of scholarly works. Robinson-Garcia et al. (2017) mentioned that in the study of tweets linked to research papers from 47 dentistry journals, 74% of articles were tweeted using the bots mechanism. These tweets are generated automatically without users knowing the content of the papers. Htoo & Jin-Cheon (2017) suggested that for using Twitter data as a more valid data source, it would be helpful to provide information on users' characteristics, such as gender, geographic distribution, etc. The additional information about the users would increase the validity of the data and enhance the analyser to provide more insights and interpretation from the data.

Wikipedia

Although articles were rarely cited on Wikipedia (2.6%), compared to Twitter or Mendeley (Mas-Bleda & Thelwall, 2016), it still has the potential to be used as an altmetrics source as an open-access encyclopaedia. However, it should be used cautiously as Wikipedia content can be edited by anyone (Serrano-López et al., 2017).

To sum up, it is predictable that data sources of Altmetrics will increase in later days when new forms of communication platforms appear. Altmetrics may be further diversified.

3.6 Some of the critical topics in altmetrics study

When the term “altmetrics” appeared not long ago, most scholars focused on the concept and connotation of altmetrics, indicators and tools, and the relationship with traditional citation metrics (see Section 3.4 & 3.5). Over the past recent years, the research on altmetrics has gradually turned to more cross-metric validation of altmetrics and validations of data quality which include the study of coverage of altmetrics, disciplinary differences, inconsistency in altmetrics aggregators, effectiveness of specific data source and data normalisation (see Section 3.6.1 & 3.6.2). The practical application of altmetrics is also widely discussed and attempted (see Section 3.6.3).

3.6.1 Cross-metric validation

Not long after the concept of altmetrics emerged, scholars began investigating the relationship between these new metrics and traditional metrics, especially the citation. The early cross-metric validation between altmetrics and citations started from bookmarking data in reference management tools (e.g., Haustein & Siebenlist, 2011; Li et al., 2012). Then, a mass of studies have been carried in various data sources such as Twitter, Facebook, Blogs, Google+, Reddit, Policy files, Q&A, and more (Barnes, 2015; Bornmann, 2014a; Costas et al., 2015; Haustein et al., 2014; Meschede & Siebenlist, 2018; Repiso et al., 2019a).

Haustein and Siebenlist (2011) proposed four article usage indicators, including *Usage Ratio*, *Usage Diffusion*, *Article Usage Intensity* and *Journal Usage Intensity* retrieved from bookmarking data in CiteULike, Connotea and BibSonomy for the use of journal evaluation. Li et al. (2012) found a significant correlation between bookmarking count and citation count. Their results suggested that online reference managers may provide valuable data to help measure scholarly impact. Zahedi et al. (2014) studied the performance of Mendeley, Twitter, Wikipedia, and Delicious and their correlation with bibliometric indicators. The results suggested that Mendeley readership is the highest coverage of altmetrics in between, and the correlation to the citation is moderate ($r = 0.49$), whilst other altmetric data sources have low metrics coverage to provide limited information. Thelwall et al. (2013) applied a sign test with a score of success or failure to examine the relationship between 11 altmetrics and the

Web of Science citation. They found a significant association between the higher metric score and the higher citation in at least six altmetrics: Tweets, Facebook posts, Blog mentions, Research highlights, media mentions and forum posts. Kraker et al. (2015) and Peters et al. (2015) studied Figshare (<http://figshare.com>) more deeply. Several papers also addressed the disciplinary difference in altmetrics and the varying extent of effectiveness of data (see, e.g. De Filippo & Sanz-Casado, 2018; HTOO & Na, 2015; Ortega, 2018a; Peters et al., 2015, 2016).

Studies also explored if the Altmetric Attention Score (from Altmetric.com) correlates with the citation count (Mukherjee et al., 2018; Mike Thelwall, 2018; Mike Thelwall & Nevill, 2018). Altmetric.com provides the Altmetric Attention Score as a comprehensive indicator of the impact that is received for research output (Elmore, 2018). The score is derived from a weighted count of attention in various social media sources, including Twitter, Facebook, News, Wikipedia, public policy documents, multi-media (i.e., YouTube), patents, and more. Barbic et al. (2016) revealed a mild correlation between Altmetric Attention scores and citations in Emergency Medicine and other biomedical journals with a scope of 200 most cited papers. Chang et al. (2019) failed to find the correlation between the Altmetric Attention Score of top-cited articles and the Journal Impact Factor. However, the weak association with individual paper's citations has been proved in paediatric surgery journals. Gumpenberger et al. (2016) criticised that the "all in one" indicator could have several issues in transparency, weight of measurement among traced data sources, reliability and inconsistency of data sources, completeness of data, and result stability. Therefore, using the Altmetric Attention Score directly to reflect the impact is not appropriate. Araujo et al. (2018) explored associated factors with high altmetric scores in clinical trials. They found four potential determinant variables: the journal impact factor as journal characteristics, descriptive title, published years, and normalised citation count as article characteristics. The research idea of the paper looks contrary to the common research direction; however, their results could be useful when the attention score is more important than citation one day.

Bornmann and Haunschild (2018) empirically studied the altmetrics by principal component analysis and factor analysis and found that Mendeley and F1000 Prime data can provide academic impact evaluation. Their results also suggested that Facebook and Twitter could provide a more comprehensive range of readers while reflecting a broader impact from a separate dimension. Bornmann (2014) verified that the research paper recommended by peer review experts in F1000 also has a higher citation count and pointed out that F1000, as a qualitative indicator, may help altmetrics play a better role in evaluating scholarly impact. Kolahi & Khazaei (2016) comprehensively selected 50 dental papers with top impact by using peer review indicators in PubPeer, Publons and experts' scores in F1000, which indicated that altmetrics could evaluate the influence of academic papers in a short time. Some scholars found that social media metrics such as Tweets and Facebook post has lower or no correlation to the citation, whilst the fact indicating the impact reflected by social media metrics are different dimension (Barnes, 2015; Bornmann & Haunschild, 2015; Costas et al., 2015; Yu et al., 2017).

Thelwall and Nevill (2018) found that Mendeley Reader is a reliable predictor for future citation. For a specific journal, *Advanced and Specialised Nursing*, weak predictors could include CiteULike, Blogs, News, Google+ and Tweets for future scholarly impact. Overall, it is revealed that the alternative metrics could predict future citations, although the current result is not good enough. However, it is possible to predict future citations better using both altmetrics and the journal impact factor (Mike Thelwall & Nevill, 2018). Mohammadi et al. (2016) conducted a user motivation survey for Mendeley users. They found that about 85% of the respondents were adding articles in Mendeley Library for future citing, 50% for use in scientific research activities, 25% for teaching, and 13% for education purposes. Although scholars used different data samples and citation databases, the common result is that there is a positive correlation between Mendeley readership and traditional citation indicators (Didegah et al., 2018; Li et al., 2012; Mike Thelwall & Nevill, 2018; Mike Thelwall & Wilson, 2016). Therefore, Mendeley Readership could reflect the scholarly impact to a certain extent.

Barnes (2015) criticised that several studies used Pearson's correlation coefficient in the analysis, which needs to be corrected since altmetrics data demonstrate a skewed distribution. Bornmann (2015) claimed that the value of altmetrics could not be demonstrated by correlation to bibliometrics only, but more studies on content analysis to interpret the behaviours behind are needed. Remarkably, he suggested that Tweets and blogs are suitable for content analyses. Erdt et al. (2016) compared over 80 existing studies of altmetrics and summarised the results of cross-metric validation by a meta-analysis. They found that the common bibliometric sources used to compare with altmetrics are WoS, Scopus, and Google Scholar Citations.

To summarise the above literature, the relationship between altmetrics and citations has been fully studied. Many correlation analysis results showed that reference management tools (e.g. Mendeley, CiteULike) and the post-publication peer-review forums (e.g. F1000, PubPeer and Publons) are positively and significantly correlated with the citations. Other alternative metrics were weakly or negatively correlated with citation, meaning the impact they reflect might differ from that of citation.

3.6.2 Data quality validation

The importance of data quality, reliability and validity must always be emphasised in the bibliometrics study, and so must the altmetrics (Zahedi & Costas, 2018). As recent research gradually recognises the particular effectiveness of altmetrics in use of reflecting the research impact (Bornmann & Haunschild, 2018b; Mike Thelwall, 2020; Mike Thelwall & Nevill, 2018), the data quality of altmetrics would become the top priority. However, altmetrics must still build trust and confidence in the scientific community before applying it practically (Carpenter et al., 2016). National Information Standards Organisation (NISO) has started to propose standards for altmetrics analytics, which shows that altmetrics are approaching mainstream acceptance (Carpenter et al., 2016; Yang & Dawson, 2018).

According to Haustein (2016), the data quality issues in altmetrics are more challenging than in bibliometrics because they are beyond the known errors and biases of citation data and

face a range of dynamic data sources rather than static materials. The main issues of data quality may include accuracy, consistency, and replicability of data at the level of data providers, data aggregators, and end users (Haustein, 2016).

Torres-Salinas et al. (2013) believe that the traditional citation indicators have stable data sources, while altmetrics rise from the complex web environment where too many interferences to guarantee a long-term and continuous measurement. Mas-Bleda et al. (2014) pointed out that altmetrics indicators rely on social media and social networks, and the functions of these platforms might be highly uncertain. For example, they may disappear like Yahoo one day. There is a lack of standardised and accurate validity tests for these indicators. Sud and Thelwall (2014) suggested that different types of study and validation would be needed to give enough evidence to support these new academic-related metrics in scientific evaluation. With altmetrics-related research becoming a hot exploratory area, many scholars have raised concerns about the quality of altmetric data and its sources. In the literature, many papers have started to validate the coverage of altmetrics, data consistency between altmetric data aggregators, quality of specific data sources, data normalisation, and so forth.

Altmetrics coverage

Some papers investigated the coverage of different altmetrics or data sources. For a given data source, coverage represents the percentage of articles in the data source within the studied data sample (Erdt et al., 2016). Mendeley is one of the most prevalent data sources and has been identified with high coverage (Alhoori & Furuta, 2014; Haustein et al., 2014; Zahedi et al., 2014). Then Twitter and Facebook contribute moderate coverage (Hammarfelt, 2014; Xia et al., 2016). The other data sources like CiteULike, Google+, F1000, Reddit, News, Wikipedia, Blogs and so on showed a low coverage of below 10% (Alhoori & Furuta, 2014; Meschede & Siebenlist, 2018; Repiso et al., 2019a). Studies have been carried out to gain a more in-depth understanding of the coverage and characteristics of each altmetric source among different aggregators. For example, Ortega (2020) intended to discover the differences in the coverage of Blogs and News and found the biases which affect the impact of altmetrics from three criteria: language, country and theme. He found that more than 90% of the blogs

and news are present in English, more than 75% of blogs and news come from English-speaking media, and more than 45% of the media outlets come from the United States. These geographic and linguistic biases significantly underestimate the impact of alternative metrics in publications from other languages or non-English speaking countries.

Some scholars have studied the coverage of Altmetrics in different disciplines and found that altmetrics have noticeable disciplinary differences in coverage. Banshal et al. (2019) found that medicine and biology subjects account for more than 50% of all altmetrics cases. Regarding the altmetrics coverage, more than 60% of articles in Biology, Medicine and Multidisciplinary Science have altmetrics. In contrast, the coverage rate for articles in Engineering, Mathematics and Materials Science is less than 25%. Ortega (2018a) believed that general discipline, which is more related to public interest, such as the Social Science subject, of a paper attracted more attention from the media. He found that the usage rate of social science articles is higher than that of physical science articles, and the citation and bookmarking rates of general articles are higher than that of health science and social science articles. Yu (2017) studied the count type and user category of altmetrics and found that disciplinary difference exists for different user categories. Banshal et al. (2018) conducted a study using ResearchGate data and found disciplinary variations in the presence of articles.

Altmetrics aggregators

Jobmann et al. (2014) evaluated the altmetrics coverage and metric counts of five disciplines between four altmetrics aggregation platforms: ImpactStory, Altmetric.com, Plum Analytics and Webometric Analyst. The study revealed that the coverage and metrics count of several altmetrics vary across disciplines and different altmetrics data aggregators. Ortega (2018b) studied three altmetrics tools (Plum Analytics, Altmetric.com, CrossRef Event Data) for the same set of publications and found that altmetrics are varied. Their results suggested that Altmetric.com has better altmetrics on Twitter, News, and blogs, PlumX has better altmetrics on Mendeley, and CED has better altmetrics on Wikipedia. Zahedi et al. (2015) compared the data consistency in three platforms (PLOS ALM, Altmetric.com and Mendeley) and explained the reasons for differences across different platforms from the perspective of update time

and tracking data methods. Zahedi and Costas (2018) carried out an extensive study for Altmetric.com, CrossRef Event Data, Lagotto open source application and Plum Analytics to understand the difference in tracking data and methodologies of aggregating among these aggregators. It is found that PLOS and PlumX provide a high value of Mendeley readership, and such a result is similar to studies in (Ortega, 2018) and (Zahedi et al., 2015). The social media metrics also vary across four different major altmetric aggregators, according to Zahedi and Costas (2018), and they concluded that the inconsistency of data among different aggregators mainly comes from three factors: the differences in data collection approaches, aggregation and reporting approaches, and updating approaches (Zahedi & Costas, 2018). These results revealed that no single platform or aggregator could reliably reflect the alternative impact. Ortega (2018b) suggested that when using altmetrics, one could select a specific metric from the specific platform according to certain advantages (e.g. highest coverage) or combine several platforms to analyse a series of altmetrics.

Obviously, the difference across the data providers of altmetrics challenges alternative indicators' reliability. Data consistency is urgently needed for future applications. The current situation is adhering to both altmetric providers and publishers (Zahedi et al., 2015).

Validation of altmetric indicator and data source

Verifying the validation of each altmetric indicator or data source is the principle for altmetrics to play its value and role. Hammarfelt (2014) found that Mendeley and Twitter are the most popular data sources in Altmetrics, while Mendeley occupies the first place regarding coverage in different disciplines. Holmberg and Thelwall (2014) found that researchers share links of research and tweet research papers more actively than average users during their research process. Haustein et al. (2014) conducted a survey on the coverage of papers published by metrology experts. They found that 82% of the authors used Mendeley to collect literature, and 50% of the respondents had Twitter accounts. Currently, Mendeley and Twitter are the two most studied indicators across the literature. In Section 3.5.3, we have reviewed a few popular data sources used in the altmetrics, which include Mendeley, Twitter, Blogs, and Wikipedia. Therefore, we will only review limited data sources here.

Scholars are still actively exploring new data sources for altmetrics. For example, Holmberg and Bowman (2019) evaluated the feasibility of using Google search results as an alternative data source, and it is found that the Google search results lack open API and exist the false data problem, which is not good enough to be an alternative source of econometric data. However, they believe that combining a variety of search engines to analyse the large-scale data sets may improve the current analysing results. Yu et al. (2020) systematically evaluated the accuracy of the data sources of policy documents on Altmeteric.com. They carried out an in-depth coding analysis, which contains 2079 policy documents extracted from 79 policy document sources, and the results showed that more than 70% of records have errors in transcriptions, links or false mentions. It reveals that the authority and credibility of policy document altmetrics need to be questioned further. Erdt et al. (2016) pointed out that some data sources have not been investigated much, for instance, Zotero, YouTube, SlideShare, and Academia.edu. Additionally, a few platforms, which could be a potential data source or aggregator for altmetrics, have not been well known, for example, AMiner¹⁴ – a search and mining service that extracts researchers' profiles from web sources and social networks (Erdt et al., 2016).

Data normalisation

Data collected from Altmetrics systematically vary depending on the research papers' different disciplines and publication time (Michael Thelwall, 2016; Mike Thelwall, 2017). Ortega (2018a) has found that specialised disciplines are less referenced and mentioned than general disciplines in social media. On Mendeley, it was found that disciplines like social sciences and humanities publications had the most significant number of readers among all disciplines compared to mathematics and computer science, which have the lowest number of readers (Mohammadi & Thelwall, 2014). Shema et al. (2014) showed that biological and medical disciplines had different numbers of posts on Blogs in 2009 (67%) and 2010 (74%). Due to these findings, time- and field-normalising altmetrics data is necessary for systematic

¹⁴ <https://www.aminer.org/>

comparison, and it has been suggested to achieve this through benchmarking altmetrics (Liu & Adie, 2013).

A percentile-based indicator for normalising is proposed by Bornmann and Haunschild (2016) for Twitter counts of published papers in different years and disciplines. The comparisons of units in different fields (such as researchers, institutions, and countries) can be achieved using the Twitter percentile indicator. Based on the Mean Normalized Citation Score (MNCS) in bibliometrics, Haunschild and Bornmann (2016) also proposed the Mean Normalized Reader Score (MNRS). The MNRS can be used to compare the impact of a paper on Mendeley across subject categories and publication years. For the studied journals, they found a high correlation ($r = 0.70$) between MNRS and MNCS. Bornmann and Haunschild (2018a) further proposed another field- and time-normalising indicator called Mantel-Haenszel quotient (MHq) for altmetrics data (e.g. tweets and Facebook posts). This indicator compares the proportions of papers shared on social media (e.g. Twitter or Facebook) with those of non-mentioned papers (not shared on social media). It has been validated that MHq could distinguish between different quality levels defined by peer assessments (Bornmann & Haunschild, 2018a).

3.6.3 Applications of altmetrics

Use of Altmetrics for Impact Assessments

Since the emergence of altmetrics, the focus of concern in the scientific community is whether altmetrics can measure research impact. The current empirical results (see Section 4.3.1) show that the different altmetric indicators could reflect the different impact dimensions. For example, Mendeley and F1000 could reflect certain academic impact, and social media metrics could reflect public engagement and interest (Bornmann et al., 2019; Stacy et al., 2016; Tahamtan & Bornmann, 2020; Xia et al., 2016). How to utilise altmetrics, its data, or data sources to measure the impact of research output is still widely discussed. Nevertheless, selecting appropriate indicators to match the evaluating goals is essential, and altmetrics should be gathered, aggregated, and interpreted with care when in actual applications (Mike Thelwall, 2021).

Zhang et al. (2019) believe that emergent interdisciplinary study such as Artificial Intelligence develops fast, the traditional metric could not reflect the impact of recent research quickly, and more timely alternative metrics would be needed. They verified the general validity of altmetrics in the AI field. It is found that altmetrics are a good indication for highly cited publications and publications with the potential to be cited increasingly. In addition, the results showed that altmetrics are more effective in high-level AI journal papers and AI conference papers (Zhang et al., 2019). Their study supports the effectiveness of altmetrics from a new perspective, and it has important implications for the research on the impact of social media on the scientific community. Dardas et al. (2019) applied the altmetric attention score to interpret the online attention of scientific papers published in nursing. The author also suggested that we should be prepared for a changing measurement environment in research quality and impact in the digital age. Oberst (2017) attempted to analyse the social impact of 148 publications produced in ACSL (African Studies Centre in Leiden) using metrics from Altmetric.com and believed that altmetric data showed a valuable mention of ASCL research output that the institute failed to notice. However, results from altmetric.com could not supply a reasonable societal impact for observed research output. More studies attempted using altmetrics data to evaluate the famous research work in Neurosurgical journals (Justin Wang et al., 2017), Radiology journals (Rosenkrantz et al., 2017), and pharmacoepidemiology research (Gamble et al., 2020).

It is criticised that many studies did not explain well the meanings of altmetrics measures, whether these indicators serve as a proxy for societal impact or some impact in early scientific attention, educational utility, or popularity (Robinson-Garcia et al., 2018). Sugimoto et al. (2017) addressed that social media is a newly opened channel for informal discussions among researchers rather than “a bridge between the research community and society at large”. Bornmann (2014) assumed that altmetrics might measure social engagement by capturing discussions of scientific papers online. Tahamtan and Bornmann (2020) believed that social media data represents the discussion of scholarly works, and those corresponding altmetric

indicators may reflect the public interest, attention, and engagement rather than societal impact or academic quality. Thelwall (2021) states that measuring the social impacts of research is challenging for various reasons. Fundamental research may occasionally culminate in applied research and practical applications, but the transversion of fundamental research to impact can be protracted. Moreover, the realm of science is an intricate web of interconnectedness, making it challenging to quantify the individual contribution of a single research endeavour.

According to Robinson-Garcia et al. (2018), social media data can reflect the interactions between inside and outside the scientific fields, enhancing our understanding of scholars' community beyond the scientific dominion. They suggested that societal impact assessment should develop a different methodological approach—based on interaction approaches. Robinson-Garcia et al. (2018) proposed using Twitter data to analyse the network of researchers to capture dissimilar degrees of public engagement. Their methodology brought a mapping perspective between contexts and social impact. Pulido et al. (2018) proposed a social impact coverage ratio (SICOR) to identify the percentage of tweets and Facebook posts that revealed the potential social of specific research output by calculating the amount of social media data. Their results showed that social media users are intermediaries to deliver visible and assessing evidence of social impact.

Use of Altmetrics for new indicators

With the expectation of the application exploration, several studies have begun developing other impact indexes with altmetric data. Fang and Costas (2018) proposed the *Velocity Index* to compare the velocity of data sources in Altmetric.com across document types and subject fields. They found that Twitter, Facebook, Reddit and News data sources accumulated fast in the early days after publishing the content. At the same time, Policy documents, Wikipedia, Q&A, and Peer review posts have low velocity values in their study. Most data sources' velocity degree also changes by document types and subject fields. Fang and Costas (2020) further proposed *altmetric half-life* and *altmetric time delay* to show that different altmetric data sources have different data accumulation speeds. Hassan et al. (2018) proposed the *Alt-*

index, which is analogous to the *h*-index (Hirsch, 2005) that identifies highly cited publications with a more accurate indicator formed using Altmetric scores. The indicator has been defined as "a scholar has an alt-index of *a*, if *a* of her/his *N_p* papers have at least *a* social mentions, and the other (*N_p*-*a*) papers have no more than *a* mentions each"(Hassan et al., 2018, p.750). Alhoori and Furuta (2014) defined a new metric, the Journal Social Impact (JSI), to indicate the online attention of the scholarly outputs. They found that the online attention of research papers has a significant correlation with the rankings of journals and tends to the journals with a long history of academic impact. Wang et al. (2014) proposed a sustainable, dynamic, and composite paper evaluation system by integrating the period of action of various indexes.

Other applications of altmetrics

Regarding the attempts of impact measurement for different scholarly outputs, Torres-Salinas et al.(2017) applied PlumX indicators to analyse the different dimensions of impact for books. Wong and Vital (2017) attempted to use PlumX to indicate researchers' scholarship, creativity and engagement in a university's faculty. Zhao and Wang (2019) used citation and altmetrics indicators comprehensively compare the influence of international open-access journals between China and the USA. There is more area that could utilise altmetric data and indicators for research. For example, McGowan (2017) called for the research of using altmetrics as a citation metric to measure the impact of an MIT/declarative title.

3.7 Knowledge gaps identified

To summarise the research state of altmetrics, the emphasis on altmetrics has shifted from the conceptual study of definition and background to the empirical analysis of the effectiveness of alternative metrics and is developing towards the direction of further validation of data quality and exploration application. How to use altmetrics data and sources to serve research evaluation and whether altmetrics could be used for a wider impact analysis for scholarly output could be the next stage of the research focus. Through a systematic review of altmetrics, including its concept (see Section 3.4), tools and data sources (see Section 3.5) and the research states in recent years (see Section 3.6), we could draw the following conclusions for the research of altmetrics.

Firstly, the relationship between altmetrics and citation has been almost fully studied. Numerous studies on correlation analysis showed a significant positive correlation between Mendeley readers and citation counts. The post-publication peer-review forums (e.g. F1000, PubPeer and Publons) are also positively correlated with the citations. Other alternative metrics were weakly or negatively correlated with citation, meaning the impact they reflect might differ from academic impact. It is suggested that Mendeley Reader and some social media metrics (Tweet and Facebook posts) could be normalised by standard methodologies from bibliometrics. Consequently, such altmetrics have great potential to be used in future research evaluation. However, more fine-grained research on different types of alternative metrics and disciplinary differences would be needed to consolidate and improve the current conclusions.

Secondly, with the gradual recognition of the effectiveness of altmetrics for measuring impact, applying altmetrics in the research evaluation would become a trend in the future. Then, sufficient study on data quality, availability and reliability is the top priority before applying these new metrics. Results from current studies are not enough. For example, some scholars have raised concerns about the reliability of altmetric data sources and potential problems for altmetric data aggregators (see Section 3.6.2). The accuracy, consistency, and replicability of altmetrics are considered the main challenges to data quality due to the dynamic nature of social media events (Haustein, 2016). A few studies have addressed the differences in data count across different altmetric platforms, whereas the explanation of the discrepancy is not enough. The technical affordances of underlying platforms potentially influence the altmetric data (see Section 3.6.2), while the problem is how to select and use altmetric data in practice still needs to be answered. More studies are needed to investigate the data quality issues in the altmetrics.

Thirdly, altmetrics are diverse. Many data, sources and platforms await to be considered or studied deeply (see Section 3.5). The forms of scholarly engagements that can be measured

are diverse, for example, downloads, favourites, comments and recommendations. How to use such data and data sources has the potential to be explored.

Fourthly, the development of altmetrics not only provides a series of new indicators for scientific research evaluation but also brings a new perspective of data and data sources that can be used in quantitative research evaluation. Altmetric data are traces of scholarly activity in the process of generating research impact (Moed, 2016). The rational utilisation of altmetric data may help us build a new quantitative evaluation method for research impact.

Fifthly, although some studies have discussed that altmetrics may not reveal the societal impact as expected (see Section 3.6.3), social media metrics provide a unique opportunity to observe the interactions and communications of research output between academicians and non-academicians, which can be widened to network approaches showing relationships and dependencies (Haunschild et al., 2019; Robinson-Garcia et al., 2018; Tahamtan & Bornmann, 2020). Using altmetrics data on network approaches would be another application of altmetrics in future research.

Regarding the research gaps discussed above, we conducted an in-depth analysis of altmetric data aggregators to detect the reason behind the data inconsistency and answer the question of which altmetric data is robust and reliable that can be used in practice (see Chapter 4). Our research results would supplement the existing literature for data quality validation. In Chapters 5, 6, and 7, we utilised altmetric data as the evidence of behaviours and results that participated in generating research impact to construct an evidence-chain-based quantitative research evaluation method. This evaluation method would bring a fresh idea on utilising altmetrics and multiple sources of data to evaluate the research impact.

3.8 Summary

This chapter begins with a review of traditional metrics for evaluating academic impact, highlighting commonly used indicators and discussing their limitations. It then examines widely utilized indices that measure journal impact. Following this, the chapter conducts a

systematic review of altmetrics, covering key concepts and definitions, an overview of altmetrics platforms and data sources, and an analysis of critical issues in altmetrics research. Finally, the chapter identifies existing knowledge gaps in the field.

Chapter 4 – Investigation of the altmetric data quality with Altmetric and PlumX as examples

(The main content of this chapter has published in the BUSTECH 2022 conference proceedings: Gong, T., Liu, W., & Wu, S. (2022). An Investigation of the Quality of Altmetric Data with Altmetric.com and PlumX as Examples. The Twelfth International Conference on Business Intelligence and Technology, 1–7. IARIA ISBN 978-1-61208-955-3.)

4.1 Introduction

Altmetrics has become a subject matter of widespread interest since its inception. From being included in the hot topics of several eminent international academic conferences (Yu et al., 2020) to the constant publishing of scholarly articles related to this field by various information science journals, altmetrics has gradually emerged as a prevalent research field for contemporary scientometrics and informatics (Qi et al., 2017). Modern-day scientific researchers are accustomed to the mainstream trend of online academic exchanges. The development of various user-oriented online social media platforms, such as blogs, Twitter, and Facebook, has not only attracted ever more researchers to utilize online search tools for literary works but has also given them increasing opportunities to exhibit their scientific research results, share and comment on accessible academic articles, communicate, discuss, and collaborate with others online.

When the term ‘altmetrics’ was first coined in *Altmetrics: A Manifesto*, Priem et al. (2010) started to call for using social media network indexes to reflect the spread and influence of scholarly output. They believe scholarly output’s impact can be determined by measuring mentions, sharing events, downloads, and comments on online platforms. Compared with traditional metrology based on citations, altmetrics features more prosperous indicators, extensive data sources, diverse audiences, and is timely. There is still debate in the academic community on whether altmetrics are appropriate as social (or broad) impact indicators to evaluate scholarly output. Undoubtedly, the advent of altmetrics has not only led to

innovative research ideas and data sources for the academic evaluation system but is also expected to improve the single evaluation method relying on citations only and compensate for the shortcomings of traditional metrological methods on time lag.

Data quality remains the core concern for the analysis and application of altmetrics. The altmetrics platforms are crucial in providing reliable data and guaranteeing data quality. There has been a series of altmetrics platforms in use, such as Altmetric.com (Altmetric), Plum Analytics (PlumX), ImpactStory and more. As we reviewed in Section 3.5, each platform has advantages, disadvantages, and limitations in terms of data coverage and data sources. With the gradual advancement of the research associated with altmetric data, several authors have expressed concerns about data quality, consistency, and reliability (Jin et al., 2019; Miles & Price, 2023). Current studies for altmetric data and platforms appear relatively scattered, with the applicability of the indexes still being a matter of debate and the presence of divergent views about the data consistency and variation (as reviewed in Section 3.6). Therefore, this study uses the data from Altmetric.com (hereinafter referred to as “Altmetric”), and Plumanalytics.com (hereinafter referred to as “PlumX”), as examples to carry out empirical research on their data coverage, consistency, and reliability and fill our research gap mentioned in Section 3.7. The study aims to provide a better understanding of the associated altmetric data and furnish valuable references for subsequent research.

4.2 Related studies

The term "altmetrics" was initially coined by Jason Priem in 2010. Subsequently, discussions on significance, cross-metric validation, and potential application of altmetrics appear constantly in the field of scientometrics (Dardas, et al., 2023; Erdt et al., 2016; González-Valiente et al., 2016; Patthi et al., 2017). Compared with traditional metrology based on citations, altmetrics features timely reactions, prosperous indicators, extensive data sources, open access and diverse audiences (Ezema, 2017; Weller, 2015).

The challenges of altmetrics are data quality, missing evidence, commercialisation, ease of

manipulation, easy bias and more (Bornmann, 2014; Thelwall, 2020). Since altmetrics count remarks in blogs, tweets and other online media that are non-publishing carriers, they provide different dimensions of impact measurement that jump out of the journal level (Galligan et al., 2013). Regarding the application exploration, scholars have engaged in developing different impact indexes under altmetrics, such as Alt-Index (Hassan et al., 2018) and Velocity Index (Fang and Costas, 2020). Studies attempted to identify the high influential research work by altmetrics in various subjects, for example, Humanities and Social Sciences data papers (McGillivray et al., 2022), Neurosurgical journals (Wang et al., 2017), Radiology journals (Rosenkrantz et al., 2017), pharmacoepidemiology research (Gamble et al., 2020), and interdisciplinary study like Artificial Intelligence (Zhang et al., 2019).

The importance of relying on reliable and robust data cannot be overemphasised at all times in the bibliometrics study, so does altmetrics (Sud & Thelwall, 2014). The National Information Standards Organisation (NISO) has started to propose standards for altmetrics analytics, which shows that altmetrics are approaching mainstream acceptance (Carpenter et al., 2016; Yang & Dawson, 2018). If altmetrics are used as a quantitative evaluation tool, the quality of the altmetric data needs to be further studied. The concern about altmetric data coverage, accuracy, consistency, replicability, and durability is arising with the gradual advancement of the research (Haustein, 2016). Altmetric data are easier to obtain through open APIs, while different collecting methods and times, as well as accessibility of particular papers, can lead to inconsistency among data collectors. The altmetrics platforms are crucial in providing data and guaranteeing the data quality.

Regarding the altmetrics platforms studies, Liu et al. (2016) researched the utilisation value of altmetrics indexes based on the PLOS platform, revealing that altmetrics indexes express variations in coverage, stability and timeliness. Zahedi et al. (2014) compared the data consistency of three of the most popular altmetrics platforms, including PLOS ALM, Altmetric, and Mendeley, which brought to the fore the evident differences between the three platforms in terms of update time and data tracking methods. Jobmann et al. (2014)

suggested the occurrence of considerable disparities in the measurement results obtained when multiple altmetrics platforms quantify the same set of papers simultaneously. Ortega (2018) analysed three altmetrics platforms - Altmetric, PlumX, and Crossref Event Data (CED), and concluded that Altmetric presented a better count of Twitter, news, and blogs, while PlumX offered a finer Mendeley count, and CED performed better with the Wikipedia count. Jin et al. (2019) investigated and sorted out six altmetrics platforms (Altmetric, PLOS ALM, Plum Analytics, Kudos, ImpactStory and Webometric Analyst) and raised two major issues: one about the consistency of the data and the other about the reliability of the data. The quick expansion and prevalence of research associated with this field have led to increasing concerns by various scholars about the data quality of altmetrics, such as the lack of consistency in data obtained from different data aggregating platforms, as well as the absence of unified standards recommended for data sources and measurement methods. The index systems provided by some of the mainstream altmetrics platforms also show dissimilarities.

In summary, contemporary comprehensive studies about altmetrics data and platforms appear relatively scattered. The applicability of the indexes is still a matter of debate, and divergent views about data consistency and variation exist. Therefore, this study aims to use the most common altmetrics platforms, Altmetric and PlumX, to conduct empirical research on their data quality and offer a reference for further understanding and utilising altmetric data.

4.3 Altmetrics platforms

Several platforms offer multi-source altmetric data, including PLOS ALM, Altmetric, Plum Analytics, ImpactStory, Crossref Event Data (CED) and Webometric Analyst. A brief introduction of each platform is discussed below.

PLOS ALM¹⁵ is the most primitive altmetrics platform offering Article-Level Metrics. It is characterized by distinct, rich, and extensive data sources with clear indexes, but it only

¹⁵ <http://almreports.plos.org/>

provides metrics for a limited number of PLOS journals. Altmetric¹⁶ was founded in 2011, which aims to track the online activity of a single academic output in real-time and count the number of mentions on various social media platforms. It determines the influence of scholarly outputs by assigning weights to different indexes and applies the “Altmetric Attention Score” to grade the level of influence. Altmetric provides accessibility to other platforms through its Application Programming Interface (API). Clients of Altmetric.com include publishers, institutions, researchers, and fund sponsors. Plum Analytics¹⁷ was created by Andrea Michalek and Michael Buschman in 2012. Its vision is to equip academia with more contemporary scholarly output measurement indexes, which essentially use the internet and multimedia in combination while utilizing and analysing the impact of scientific research outputs, aiming to overcome the dependence of traditional evaluation systems on the number of citations and journal impact factors. Plum Analytics was first acquired by EBSCO in 2014, followed by the Elsevier Group in 2017 and has since been embedded in Elsevier’s document retrieval tool Scopus. PlumX Metrics, in principle, are a series of measurement indexes offered by the altmetrics platform Plum Analytics that illustrate an inclusive online interaction between people and scholarly outputs. This platform also provides special altmetrics measurement services that assist in making payments to academic institutions. ImpactStory¹⁸, founded in 2011, seeks to help researchers determine and share the impact of their research outputs. Nevertheless, an API key is currently required to download the data from ImpactStory and is not accessible to non-paying users. Crossref Event Data (CED)¹⁹, officially released in 2017, offers free data access via a public API. Unlike other platforms, it does not process the collected data sources or provide any specific indicators but exhibits altmetrics information linked to the DOI identifier. Webometric Analyst²⁰, also known as the explorer of LexiURL, utilizes URL references or title mentions to replace hyperlink searches on the Internet and generate link impact reports. Although Webometric Analyst can track

¹⁶ <https://www.altmetric.com/>

¹⁷ <https://plumanalytics.com/>

¹⁸ <https://profiles.impactstory.org/>

¹⁹ <https://www.crossref.org/services/event-data/>

²⁰ <http://lexiurl.wlv.ac.uk/>

original data sources, such as Mendeley or Bing, and the altmetrics platforms like Altmetric, it is still limited by the data sources collecting channels.

After carefully evaluating each platform's advantages, disadvantages, and characteristics, this research focuses on a comparative study of Altmetric and PlumX, which are considered the two most dynamic altmetrics platforms currently and appear to be more inclusive and accessible than other platforms. Altmetric currently provides 16 altmetric data sources, and PlumX currently provides 44 impact measurement indexes comprising more than 50 data sources. More details of these two platforms are reviewed in Section 3.5.2.

4.4 Data collection

Data was collected from three platforms, Web of Science, Altmetric and PlumX, in March 2021. To ensure the variety of papers to a certain extent and to explore the differences between disciplines, we initially targeted four categories for data collection: (1) humanities, (2) business management, (3) chemical engineering, and (4) medicine. According to the Journal Citation Reports (published by Clarivate and accessed from the Web of Science), we started to choose journals from the highest impact factor in the corresponding category and then downloaded journal papers' basic information, including DOIs, from the Web of Science. Each category collected 3000 papers, so the total number of papers is 12,000 in this sample.

Through each paper's DOI, we acquired the corresponding altmetric data for all documents via the APIs of Altmetric and PlumX, respectively. Regarding the missing data, as neither of the platforms publishes any media data with a value of 0, we leave such data value blank without replacing it with any assignment (such as 0), indicating such altmetric data is temporarily unavailable or has not been tracked. It should be noted that some Mendeley data on the two platforms displays 0 when the API is exported, which signifies that the number of readers is 0. We reserved 0 in the Mendeley data.

The initial data sample contains abundant altmetrics, and the data sources provided by

Altmetric and PlumX are largely different. To examine the data quality in a targeted manner, this study mainly focuses on the data sources that are mutually available on both platforms, as shown in Table 4.1.

Table 4.1 Altmetric data and sources studied

No.	Data	Data Sources Interpretation	
		Altmetric	PlumX
1	Blogs	Altmetric curates a list of blogs and collects links to relevant academic content through RSS feeds.	PlumX curates a list of blogs and collects the number of blogs mentioning the scholarly output.
2	CiteULike	The number of users of relevant scholarly outputs saved in CiteULike (not included in Altmetric Score, only in Altmetric Explorer).	The number of scholarly outputs added to the literature management tool-CiteULike.
3	Facebook	Only tracks the posts on Facebook's public pages and prioritizes popular pages.	Counts the number of times the links to scholarly outputs have been shared, commented, or liked on Facebook.
4	Mendeley	The number of users who save scholarly outputs to Mendeley Library is recorded as the number of readers. Attention Score no longer counted, only data provided.	The number of users who add scholarly outputs to Mendeley Library.
5	News	Altmetric curates a list of news sources and the third-party providers and RSS feeds provide data directly.	PlumX curates a list of news sources and collects the number of news mentioning the scholarly outputs.
6	Twitter (changed to X since July 2023)	Count the public tweets, reposts or quoted tweets of related scholarly outputs on Twitter, and monitor suspicious activities.	The number of tweets and retweets mentioning scholarly outputs via Gnip.
7	Wikipedia	Collects the mentions of scholarly outputs in Wikipedia references; only for English Wikipedia.	The number of references cited in Wikipedia.

4.5 Coverage analysis

The coverage of altmetrics determines the range of its applications. Understanding the coverage of the altmetrics platform helps researchers utilize the data in a targeted manner and remains a prerequisite for constructing indicators and evaluation applications. Accordingly, we will first examine the publication coverage on Altmetric and PlumX and then compare the altmetric data coverage on both platforms.

4.5.1 Publication coverage

Altmetric provides an altmetric ID for each recording publication, and PlumX also endues unique IDs for all covered documents. Therefore, these non-reduplicative IDs help us quickly identify how these two platforms have covered the paper sample (N =12,000) we targeted.

It was initially found that 17 papers on PlumX were missing. After meticulously inspecting these missing papers, we learned that certain documents had incorrect DOIs. Since the DOIs were derived from the WoS database, we tracked back to the WoS and found that 16 papers had errors in DOI. Unexpectedly, we found notable mistakes in the data records on WoS. After revising these DOIs, 11,999 documents were found on PlumX, so the document coverage (for this sample) was approximately 100%.

On the other hand, 8,091 papers in the sample were found on Altmetric, and the remaining 33% of the documents have no corresponding data. Is the publication coverage on Altmetric lower? The answer may not be. We must be aware that Altmetric's primary service generates the "Altmetric Attention Score" for scholarly outputs. This score is determined by assigning different weights to the number of times that academic papers have been mentioned on online media. Therefore, each paper on the Altmetric possesses at least one altmetric data and value of the Altmetric Attention Score. When valid altmetric data is not collected, the platform displays no indexes or scores for corresponding documents. In this sampling, about 33% of the papers from Altmetric were missing. It indicated that such papers have no valid altmetric data but could not determine whether Altmetric failed to cover these documents.

A few previous studies (Yu et al., 2020; Zahedi et al., 2018) pointed out that PlumX's coverage is very high (>90%), while Altmetric's coverage appears relatively low (39%—61%). However, these studies failed to discuss the reason behind the coverage calculation.

Hereby, we provide a supplementary explanation to the similar results. PlumX provides citation indexes that are sourced from many traditional citation databases, including CrossRef, PubMed, Scopus, etc. PlumX's approximately 100% paper coverage is due to these high-coverage citation databases. Altmetric only focuses on social media indexes and does not supply citation indexes or display articles without altmetric data. Therefore, the publication coverage on Altmetric should be higher than it looks.

4.5.2 Altmetrics coverage

The coverage of each altmetric data source for the two platforms was computed, as shown in Table 4.2. The data source with the highest number of valid readings on Altmetric is Mendeley, which has 8,031 data readings, and the coverage ratio in this sample is 66.9% ($8,031/12,000 = 66.9\%$), followed by Twitter, with a total of 7,691 valid data readings and coverage ratio of this sampling is 64.1%. The coverage ratios of Blogs, CiteULike, Facebook, and News on Altmetric are 13.6%, 4.8%, 26.1%, and 18.5%, respectively, while the coverage ratio of Wikipedia is the lowest at about 2.7%. On PlumX, Mendeley has the highest number of valid readings, with a coverage of 97.6% ($11,713/12,000 = 97.6\%$) in this sampling statistic. Then is Twitter, with a coverage of 54.7%. The coverage of Blogs, CiteULike, Facebook, and News on PlumX are 8.3%, 4.2%, 13.2%, and 14.1%, respectively, while the coverage of Wikipedia is the lowest at about 3.4%. Figure 4.1 compares each altmetric data source for the two platforms, and we can see that Altmetric has higher coverage of social media data sources, such as Blogs, Facebook, News, and Twitter; PlumX has higher coverage on Mendeley and Wikipedia.

Table 4.2 Altmetrics Data Coverage Ratio and Overlapping Ratio on Altmetric and PlumX

	Altmetric		PlumX		Altmetric - PlumX		
	#	%	#	%	Overlapping #	ratio in A	ratio in P
Blogs	1630	13.6%	1000	8.3%	633	38.8%	63.3%
CiteULike	578	4.8%	500	4.2%	423	73.2%	84.6%
Facebook	3135	26.1%	1587	13.2%	903	28.8%	56.9%
Mendeley	8031	66.9%	11713	97.6%	7859	97.9%	67.1%
News	2214	18.5%	1693	14.1%	1524	68.8%	90.0%
Twitter	7691	64.1%	6564	54.7%	6471	84.1%	98.6%
Wikipedia	329	2.7%	405	3.4%	300	91.2%	74.1%

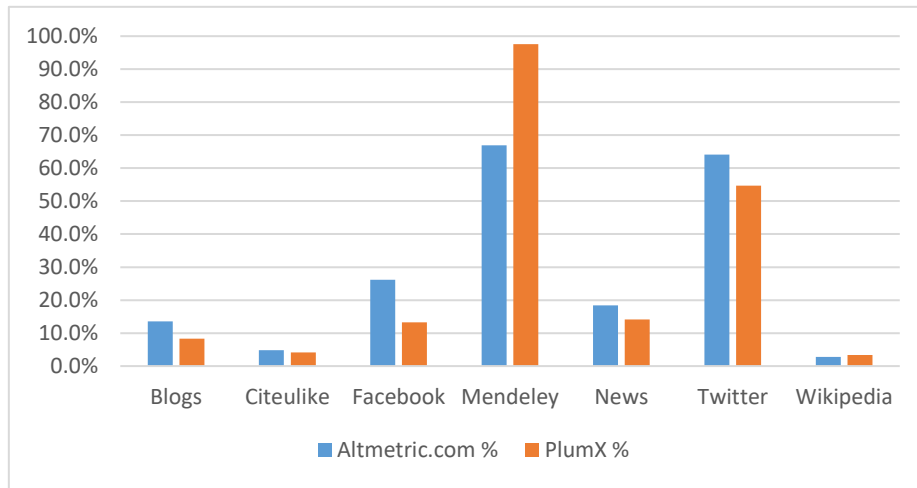


Figure 4.1 Altmetrics Coverage on Altmetric and PlumX

Furthermore, we counted the number of overlapping data (intersection) of each altmetric data source for the two platforms. In the meantime, we calculated the proportion of the overlapping data (intersection) in their respective data source of the two platforms. The results are shown in Table 4.2. According to the proportion of overlapping data in each data source, the cross-coverage status of each data source in the two platforms can be derived. We must emphasize only the proportion of overlapping data discussed, not the data value, as our focus is the cross-coverage status. According to the results, we illustrated the status of the intersection of each altmetric data source in schematic diagrams and divided them into three types:

- 1) **Dominant coverage:** means one platform has obvious coverage advantages. As shown

in Figure 4.2, Twitter, News, Mendeley, and Wikipedia present obvious advantages of data source coverage on a single platform. According to Table 4.2, both platforms have 6,471 overlapping readings from Twitter, accounting for about 98.6% of PlumX's Twitter coverage. This indicates that Altmetric covers 98.6% of the Twitter data on PlumX, and Altmetric holds more Twitter data than PlumX. Therefore, Altmetric has a dominant Twitter coverage compared to PlumX. Regarding the News data, the platforms have 1,524 overlapping data readings, accounting for 90% of the News data coverage on PlumX and 68.8% on Altmetric. This implies that more News data is found on Altmetric, and 90% of PlumX's News data source has been covered. Hence, Altmetric has a dominant coverage of News data sources compared to PlumX. On the other hand, PlumX appears to have data coverage advantages in Mendeley and Wikipedia. The intersection data of Mendeley in the two platforms is 7,859, which accounts for 97.9% of Altmetric coverage and 67.1% of PlumX. This means PlumX has covered 97.9% of Mendeley data collected by Altmetric. Therefore, PlumX's Mendeley data coverage is more prominent than Altmetric. Similarly, PlumX's Wikipedia data coverage is the dominant coverage compared to Altmetric.

- 2) **Different coverage:** means each platform has unique data sources, and the cross-coverage is limited. For Blog and Facebook, the intersection of the same altmetric data source of the two platforms only accounts for a small proportion in any of the platforms, as shown in Figure 4.3. The two platforms have 633 overlapping Blog data readings, accounting for 38.8% and 63.3% of data coverage on Altmetric and PlumX, respectively. Although Altmetric collects more data than PlumX, the latter possesses different data coverage from Altmetric, meaning PlumX has its unique Blog data source. A similar situation applies to Facebook data: the two platforms' data source coverage seems dissimilar, and certain overlaps and unique parts exist. This means the two platforms' data sources only represent a limited crossover range and have different data source coverage.

- 3) **Similar coverage:** means all platforms have similar data source coverage. For instance, the intersection of CiteULike data sources accounts for similar proportions in both platforms, as shown in Figure 4.4. The two platforms have 423 overlapping data for CiteULike, which accounts for 84.6% of PlumX's coverage and 73.2% of Altmetric's coverage. Although Altmetric has more CiteULike data on values, the coverage of the CiteULike data on both platforms tends to be similar.

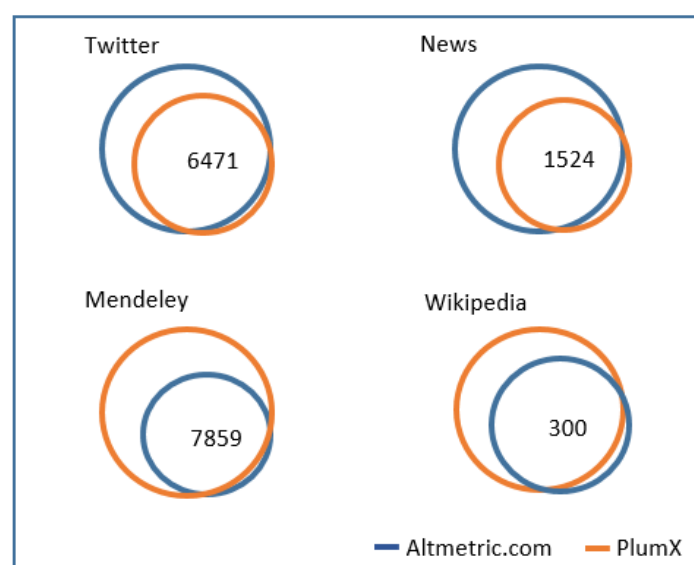


Figure 4.2 The schematic diagram for dominant coverage of the data source

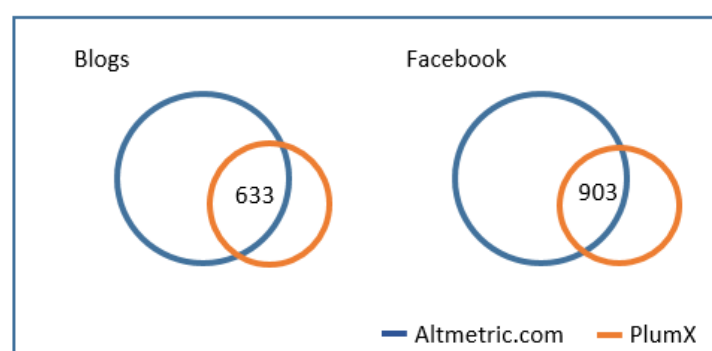


Figure 4.3 The schematic diagram for different coverage of the data sources

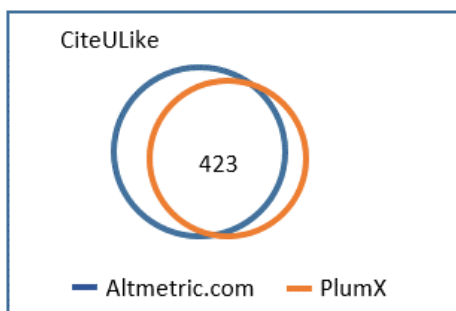


Figure 4.4 The schematic diagram for similar coverage of the data sources

4.5.3 Disciplinary differences of coverage

Furthermore, we compared the altmetrics coverage between different disciplines on the two platforms, and the results are formulated in Table 4.3. The coverage ratio is calculated by the number of available data over 12000 papers.

From the perspective of disciplinary differences, all altmetric data in medicine have higher coverage than in other disciplines, as shown in Table 4.3. Chemical engineering has more altmetric data on Blogs, News, Facebook, and Twitter than the humanities and business management disciplines, but it is lower than medicine. Interestingly, there is more discussion of the chemical discipline in mainstream social media than in humanities and business subjects. However, the chemical discipline retains the lowest coverage in CiteULike among the four disciplines. The humanities discipline has a higher Wikipedia coverage than business management and chemical engineering, but its News coverage is the lowest among the four disciplines.

For business management disciplines, altmetrics coverage is lower than that of the other disciplines and is lowest for the data sources of Twitter, Facebook, and Blogs. Business-related studies seem not keen to be discussed on mainstream social media. Compared to other disciplines, Wikipedia data in business management also stands at the lowest. This may be attributable to the broader channels through which business papers spread social impact, such as videos. The common perception is that business management disciplines are relatively social and are expected to have more social media data readings. However, this

study's statistical results provide a different answer. The disciplinary differences of altmetrics ought to be further studied in the future.

Table 4.3 Altmetrics coverage between different disciplines

	Humanities		Business Management		Chemical Engineering		Medicine	
	<i>Altmetric</i>	<i>PlumX</i>	<i>Altmetric</i>	<i>PlumX</i>	<i>Altmetric</i>	<i>PlumX</i>	<i>Altmetric</i>	<i>PlumX</i>
Blogs	259 (8.63%)	131 (4.37%)	152 (5.07%)	98 (3.27%)	563 (18.77%)	291 (9.70%)	656 (21.87%)	480 (16.00%)
CiteULike	67 (2.23%)	69 (2.30%)	187 (6.23%)	187 (6.23%)	48 (1.60%)	37 (1.23%)	276 (9.20%)	207 (6.90%)
Facebook	453 (15.10%)	441 (14.70%)	186 (6.20%)	236 (7.87%)	791 (26.37%)	75 (2.50%)	1705 (56.83%)	835 (27.83%)
Mendeley	1977 (65.90%)	2900 (96.67%)	1239 (41.30%)	2978 (99.27%)	1991 (66.37%)	2957 (98.57%)	2824 (94.13%)	2878 (95.93%)
News	169 (5.63%)	141 (4.70%)	200 (6.67%)	150 (5.00%)	440 (14.67%)	311 (10.37%)	1405 (46.83%)	1091 (36.37%)
Twitter	1895 (63.17%)	1662 (55.40%)	1105 (36.83%)	958 (31.93%)	1902 (63.40%)	1827 (60.90%)	2789 (92.97%)	2117 (70.57%)
Wikipedia	71 (2.37%)	90 (3.00%)	47 (1.57%)	50 (1.67%)	55 (1.83%)	72 (2.40%)	156 (5.20%)	193 (6.43%)

Lastly, combining the disciplinary differences with the platform differences, both platforms have the highest coverage of altmetrics on medical discipline, and the coverage of other data sources of the other three disciplines varies between the two platforms, for example, the Facebook data. Although Altmetric's Facebook data coverage is generally higher than PlumX's, PlumX's Facebook coverage in business management is higher than Altmetric's. Besides, the two platforms' Facebook data in chemical engineering disciplines differ greatly. This shows that the Facebook data tracked by the two platforms have significant disciplinary differences. For Mendeley data, the coverage of the two platforms in the medicine discipline field is not very different; nonetheless, in the other three disciplines, PlumX's Mendeley data coverage is significantly higher than Altmetric. The above findings highlight the differences in data tracking between the two platforms across different disciplines.

4.6 Correlation analysis

The internal relevance of altmetric data and the correlation of the two platforms' cross-data were investigated in the following sections. We first tested the normality of the data and found that all the variables are not normally distributed. When a linear correlation exists between the two continuous variables, Pearson's product-moment correlation coefficient is generally used. Otherwise, Spearman's rank correlation coefficient can be employed to measure the nonlinear dependence between two random variables (Myers & Lorch, 2013). Spearman's correlation coefficient, also known as Spearman's rank, is a non-parametric statistical method that utilizes the rank value of two variables for linear correlation analysis and does not require the distribution of the original variables (Artusi et al., 2002). Since all the variables are not normally distributed, the Spearman's Rank test was appropriate for correlation analysis in this section.

Correlation is an important index to quantify the consistency of changing trends among different variables. To clarify, when the correlation coefficient is less than 0.1, indicating that there is no correlation between the two variables; when the correlation coefficient is 0.1-0.4, the variables are weakly correlated; when the correlation coefficient is 0.4-0.6, the variables are moderately correlated; when the correlation coefficient is 0.6-0.8, the variables are relatively strongly or strongly correlated; when the correlation coefficient is 0.8-1, the variables are strongly or highly correlated.

We then conducted the intra-correlation coefficients of each data source for the two platforms, respectively, and the results were presented in Tables 4.4 and 4.5. In Table 4.4, the most relevant data sources on Altmetric are Facebook and Twitter ($r=0.632$). Evidently, the impact reflected by these two data appears similar. Blogs, News, Facebook, and Twitter on Altmetric are all moderately correlated. Noticeably, they are all social media data sources. The moderate correlation between these four data sources may indicate that some papers discussed on these media platforms overlap, such as hot articles. In other words, different online media outlets all discussed hot articles, whilst the four media outlets'

fundamental differences in user groups and communication methods have led to the difference in impact reflected by each data type.

Table 4.4 Intra-correlation of data sources on Altmetric

	Alt_Blogs	Alt_Citeulike	Alt_Facebook	Alt_Mendeley	Alt_News	Alt_Twitter	Alt_Wikipedia
Alt_Blogs	1.000	.133	.542**	.406**	.539**	.498**	.239**
Alt_Citeulike		1.000	.244**	.294**	.073	.343**	.166
Alt_Facebook			1.000	.380**	.520**	.632**	.199**
Alt_Mendeley				1.000	.302**	.368**	.165**
Alt_News					1.000	.465**	.210**
Alt_Twitter						1.000	.156**
Alt_Wikipedia							1.000

** . Correlation is significant at the 0.01 level (2-tailed).

Table 4.5 Intra-correlation of data sources on PlumX

	Plu_Blogs	Plu_Citeulike	Plu_Facebook	Plu_Mendeley	Plu_News	Plu_Tweet	Plu_Wikipedia
Plu_Blogs	1.000	.038	.454**	.443**	.584**	.478**	.109
Plu_Citeulike		1.000	.242**	.205**	.007	.325**	.207
Plu_Facebook			1.000	.173**	.455**	.402**	.104
Plu_Mendeley				1.000	.354**	.434**	.141**
Plu_News					1.000	.376**	.098
Plu_Tweet						1.000	.170**
Plu_Wikipedia							1.000

** . Correlation is significant at the 0.01 level (2-tailed).

In the view of information independence, if the correlation between data sources is weak, it may be easy to highlight the different dimensions of impact measured by the different data sources. For instance, Wikipedia on Altmetric is weakly correlated with Blogs ($r=0.239$), Facebook ($r=0.199$), Mendeley ($r=0.165$), News ($r=0.210$), and Tweets ($r=0.156$) at a significant level of 0.01, indicating that Wikipedia presents a different dimension of impact than other data sources. We can also see that most data sources have weak or moderate correlations in between. The weak correlation between different data sources indicates that

there is still a small correlation in the influence measured by different data. It shows that there is always a small number of articles that have an impact on all platforms. In addition, when an article is spreading on different social media networks, there should be a certain extent of academic impact generated first (because scholars are the key stakeholders in spreading scholarly papers at first), and then more public/social impact generated following. Since different media have different public influences, these altmetric data sources are more likely to present a certain correlation in between.

Table 4.5 presents the correlation coefficients of the data sources on PlumX. The four social media data sources on PlumX are also moderately correlated, for instance, Facebook and Twitter ($r=0.402$), Blogs and News ($r=0.584$). Mendeley shows a weak correlation with social media data, and the correlation coefficient with Facebook appears to be the lowest ($r=0.173$). CiteULike is not correlated with News but weakly correlated with other data sources. Wikipedia has a weak or very weak correlation with all other data sources. The correlation results of data sources on PlumX are similar to Altmetric, while the overall correlation coefficients are slightly lower.

Internal correlation explained the similar or different dimensions of impact measured by the different altmetric data sources. Both platforms illustrate varying degrees of internal relevance.

Next, we explored the correlation of the same altmetric data source between Altmetric and PlumX. The cross-correlation coefficients were summarized in Table 4.6. We found that Mendeley in the two platforms has the highest coefficient ($r=0.977$), followed by Twitter ($r=0.840$), Wikipedia ($r=0.763$), CiteULike ($r=0.753$), and News ($r=0.710$). These data sources present a strong correlation. Blogs and Facebook only exhibited a moderate correlation. Apparently, the data from the same data source on the two platforms differ. In Section 4.5.2, we discussed the findings of coverage difference in Blogs and Facebook data sources and demonstrated the intersection status in Figure 4.3. Therefore, the moderate cross-

correlations in Blogs and Facebook data sources were unsurprising.

Table 4.6 Cross-correlation of data sources on Altmetric and PlumX

	Blogs	CiteULike	Facebook	Mendeley	News	Twitter	Wikipedia
Altmetric-PlumX	.576**	.753**	.413**	.977**	.710**	.840**	.763**

** . Correlation is significant at the 0.01 level (2-tailed).

The cross-correlation of the two platforms' altmetrics demonstrated the differences between the altmetric data sources on each platform. In the following sections, we will delve into the differences between the data sources of the two platforms in more detail.

4.7 Data consistency and heterogeneity

This section implements a statistical method, the paired hypothesis test, to evaluate the data difference between the two platforms. Statistical hypothesis testing is used to check whether a difference exists between the two sets of experimental results and whether the difference is significant (Blair et al., 1985). If there is a significant difference between the two paired samples, the samples under comparison are assumed to be from different populations. Since unavoidable differences between individuals always exist, such as experimental errors caused by artificial factors or several uncontrollable aspects, it is essential to determine whether these differences are merely random errors or caused by different treatment effects or collection methods. For example, there might be some inconsistent data values when comparing the same data source between two platforms. On the other hand, the difference in data value between the two platforms may be caused by uncontrollable factors such as manual input errors, collection errors, and delayed updates, all of which account for random errors. In this case, we can still assume that the data of the two platforms is consistent.

Conversely, if the data of the two platforms appear significantly different, the data source of the two platforms is assumed to be from two different populations. In other words, the data inconsistency may be due to the difference in the data collection mechanism used by the platforms. As a result, we cannot assume the data (and source) in the two Platforms are

consistent. Before the hypothesis test, we tested whether the data followed the normal distribution. The result showed that the difference between the two platforms fails to follow the normal distribution. Therefore, we should employ a non-parametric test method, the Wilcoxon signed rank test method, to conduct the significance test of difference (Swift and Piff, 2014).

As for data processing, we matched each altmetric data source with the same article and only retained the overlapping data of the two platforms. For example, if an article has Blog data on Altmetric but not on PlumX, such Blog data was not included. Table 4.7 displays the basic statistical information on the overlapping data in the seven altmetrics studied. The Wilcoxon signed rank test was performed on the overlapping altmetric data sources of the two platforms, and the test results are shown in Table 4.8, in which Z stands for the value of the test statistic.

Table 4.7 Statistics of overlapping data of each data source

	N	Mean	Std. Deviation	Minimum	Maximum
CiteULike - Altmetric	423	1.29	0.758	1	6
CiteULike - PlumX	423	1.28	0.719	1	6
Blogs - Altmetric	633	3.17	3.687	1	31
Blogs - PlumX	633	2.55	3.153	1	33
News - Altmetric	1524	15.93	28.443	1	276
News - PlumX	1524	8.11	21.142	1	311
Twitter - Altmetric	6471	34.39	172.117	1	8128
Twitter - PlumX	6471	29.01	163.116	1	9819
Facebook - Altmetric	903	7.02	17.245	1	324
Facebook - PlumX	903	328.99	2684.636	1	39422
Wiki - Altmetric	300	1.46	1.359	1	14
Wiki - PlumX	300	1.68	1.682	1	15
Mendeley - Altmetric	7859	70.71	110.001	0	2372
Mendeley - PlumX	7859	70.60	109.527	0	2382

Table 4.8 Statistics of Wilcoxon Signed Ranks Test

	CiteULike	Blog	News	Twitter	Facebook	Wikipedia	Mendeley
Z	-0.039 ^b	-7.158 ^b	-24.801 ^b	-5.803 ^b	-21.550 ^b	-5.575 ^b	-11.313 ^b
Asymp. Sig. (2-tailed)	.969	.000	.000	.000	.000	.000	.000

a. Wilcoxon Signed Ranks Test

b. Based on positive ranks.

The following sections discuss the consistency and heterogeneity of each altmetric data source corresponding to the two platforms in detail and analyse the reasons for the heterogeneity.

(1) CiteULike

As shown in Table 4.7, the overlapping CiteULike data for the two platforms is 423 ($n = 423$). The average data value of Altmetric is 1.29, and PlumX's is 1.28, showing a very minute difference. The standard deviations of data of the two platforms are 0.758 and 0.719, respectively, showing similar dispersions. The result of the Wilcoxon signed rank test (Table 4.8) for CiteULike is $Z = -0.039$, and the two-sided value of asymptotic significance is $P = 0.969$. This suggests that we fail to reject the null hypothesis and conclude that there is no significant difference between the CiteULike data on Altmetric and PlumX.

The CiteULike data of the two platforms being not significantly different indicates that the data comes from the same population. Although minor differences could still be noticed in the values of CiteULike on the two platforms, it may be an error caused by artificial reasons, which could be considered a random error. Based on the current test results, the CiteULike data on the two platforms appears consistent. Additionally, the cross-correlation result in the previous section shows that there is also a strong correlation between CiteULike on the two platforms ($r=0.753$). Hence, it does not affect the statistical results when using CiteULike data from either platform under consideration. Therefore, when using CiteULike data for research and analysis, the CiteULike data of the two platforms may be used at will.

(2) Blogs

As shown in Table 4.7, there are a total of 633 papers that have overlapping Blog data on the two platforms. The average Blog data value of Altmetric is 3.17, and PlumX's is 2.55, seemingly less than the former. The standard deviations of data of the two platforms are 3.687 and 3.153, respectively, and the degree of dispersion shows a minor difference. According to Table 4.8, the result of the signed rank test is $Z = -7.158$, and the two-sided value of asymptotic significance is $P = 0.000$, which is lower than the significance level ($\alpha=0.05$). Thus, the null hypothesis should be rejected, and the conclusion is that there is a significant difference between the two matching variables. Therefore, the differences are not random errors such as manual input errors but induced by factors such as different data sources or ways of data collection among the two platforms.

When analysing the reasons for the difference between the two platforms, it is initially noticed from Table 4.1 that the Blog data tracked by the two platforms comes from their own curated list. Since neither of the platforms discloses their Blog list information, the Blog lists may differ. According to the websites of Altmetric.com, the counting method of Altmetric Blogs is to search for articles' citations and references (especially links) across each blog automatically based on their curated blog list. Moreover, Altmetric only searches for links related to the research outputs through the information in the RSS feed, which indicates that it can only search for the content in the blog text, excluding the mentions in the sidebar of the Blog²¹. PlumX also cites on its official website that counts the mentions of Blogs by tracking all the links related to scholarly outputs in its own curated Blog list²². Nevertheless, neither platform has yet to publish its blog lists.

The data difference between the two platforms is believed to arise from different data sources, while further explanation is incompetent. Ortega (2019) compared the Blog data of Altmetric, PlumX, and Crossref Event Data and found that the overlap rate between these

²¹ <https://help.altmetric.com/support/solutions/>

²² <https://www.elsevier.com/insights/metrics/plumx>

platforms was shallow. His study also fails to disclose detailed information regarding the blog list for Altmetric and PlumX. Besides, Shema et al. (2014) examined the peer-reviewed Blog posts discussing academic papers on Researchblogging.org, compared the WoS citations of related papers, and found that “Blog citations” are significantly associated with journal citations. There are few studies to discuss blog data so far and it is difficult to clarify the blog list coverage status of Altmetric and PlumX with limited information and low transparency. Therefore, the current Blog data quality is questionable and should only be used with discretion. If it is unavoidable to use Blogs data, it is preferable to adopt the data of both platforms simultaneously for complementation.

(3) News

The two platforms have 1,524 overlapping News data, as shown in Table 4.7. The average News data value of Altmetric is 15.93, and that of PlumX is 8.11. The standard data deviations of the two platforms are 28.443 and 21.142, respectively. The News data readings on the two platforms are different. The Wilcoxon signed rank test result for News data is $Z = -24.801$, the two-sided value of asymptotic significance is $P = 0.000 < 0.05$, the null hypothesis is rejected accordingly, and the conclusion is two matching variables appear significantly different. Therefore, the difference in News data between the two platforms is not a random error. That is, the two platforms may possess different data sources or data collection methods.

When tracking the reasons behind the heterogeneity, it is noticeable that the two platforms use their edited News lists (see Table 4.1), which may lead to varied third-party data sources. We found that Altmetric has exposed its primary third-party news sources and tracking methods on its websites, including identifying links to academic papers in news reports and tracking information in the news content. It is conducive to enhancing the transparency and reliability of the News data. PlumX has yet to publish a specific list of its news sources on its website, instead, it disclosed its third party news data provider, Newsflo, which covering more than 55,000 diverse news sources (Allen, 2017). On the other side, Altmetric announced that the platform has been cooperating with the news service provider Moreover Technologies since 2015, and has expanded its track list from approximately 1,300 news media to more

than 80,000 news media. However, Ortega (2020) pointed out that Lexis-Nexis later acquired Moreover Technologies, and the cooperation with Altmetric ended, leaving 19% of the links invalid, so Altmetric only manages 2,900 news media. Altmetric's official website updated the information about News data sources on 17 Sep 2020 and stated that it still tracked more than 5,000 global mainstream news media portals (Solutions, 2023). As the two platforms possess diverse data sources with evident differences, direct comparison is not recommended. In this experimental study, Altmetric's news data is higher than PlumX's in terms of total coverage, coverage of disciplines, and data values, and Altmetric's news sources are more appreciably transparent than PlumX's. Therefore, our results revealed that Altmetric's news data is deemed more reliable.

(4) Twitter

The total overlapping Twitter data of the two platforms is 6,471. The average data value of Altmetric is 34.39, and that of PlumX is 29.01, seemingly less than the former. The Wilcoxon signed rank test results show that the null hypothesis is rejected, and there is a significant difference in the Twitter data of the two platforms. Therefore, the difference in Twitter between the two platforms cannot be considered a random error but may have different methods for collecting or processing data.

The first reason for the difference in Twitter data could be attributed to the two platforms' different measurement standards for the data. As the statistical methods of the two platforms in Table 4.1, the Twitter count of Altmetric = the number of tweets + the number of retweets + the number of quoted tweets, while the Twitter count of PlumX = the number of tweets + the number of retweets. Therefore, statistically, Altmetric additionally includes quotation tweets. Indeed, the average Twitter data value of Altmetric is much higher than PlumX's. The second reason is that the two platforms have different collection resources for Twitter data. Altmetric's Twitter data comes from a third party, which is not disclosed on the official website. While PlumX's data comes from data provider GNIP, the tweet data provided by GNIP does not contain quoted tweets (Michalek, 2016). Evidently, the Twitter data source of Altmetric is different from that of GNIP. Therefore, we suggest that both platforms' Twitter

data come from different third parties. Furthermore, the different measurement standards of the third-party data sources may also have different counting methods. Hence, we assume that the difference in the Twitter data of the two platforms is essentially due to the two different third-party data sources, and the measurement standards of the data sources for Twitter counting may differ.

(5) Facebook

For Facebook data, the overlapping data in the sample is 903, and the average data value of PlumX is 328.99, much larger than Altmetric's 7.02. The standard data deviation on the two platforms also appears quite different. According to the Wilcoxon signed rank test results in Table 4.8, $Z = -21.550$, $P = 0.000 < 0.05$, the null hypothesis is rejected, so the Facebook data from the two platforms is significantly different. Therefore, Facebook data from the two platforms may have different data sources or utilise different data collection methods.

The first reason for the difference in Facebook data could be the fact that the two platforms possess different definitions of Facebook data measurement. According to the official websites of the two platforms, Altmetric only counts Facebook's public posts mentioning links to research outputs and prioritises statistics of popular pages, excluding individual posts on the timeline and the number of likes. PlumX counts the total number of shares, likes, and comments on the research link on Facebook. Furthermore, PlumX's early counting for Facebook data was to incorporate the shares and likes of links related to the scholarly outputs in public pages into social media data and links in comments connected into the category of mention data. However, since August 2016, PlumX added the number of comments on the personal page and integrated all of them into one Facebook data count. This integrated data only records the number of comments and does not disclose the content of the comments, so it enhances the measurement of the attention score of an academic link to the greatest extent and also enhances the visibility of impact measurement as well as complying with the privacy regulations (Allen, 2016). Therefore, the average Facebook data value of PlumX is much larger than that of Altmetric in the overlapping data. In association with the results in previous sections, we can note that the two platforms demonstrate different data source

coverage.

(6) Wikipedia

There are 300 overlapping Wikipedia data readings for the same article on the two platforms. The average Wikipedia data value of Altmetric is 1.46, and that of PlumX is 1.68, appearing only minutely different. The standard data deviations of the two platforms are 1.359 and 1.682, respectively. The Wilcoxon signed rank test result is $Z = -5.575$, $P = 0.000 < 0.05$, so the null hypothesis is rejected, and the conclusion remains that the two matching variables are significantly different.

Since the Wikipedia data from the two platforms are significantly different, we first check the measurement method of the two platforms. As shown in Table 4.1, Altmetric only counts the references to a particular scholarly output in the reference section of English Wikipedia, while PlumX counts all scholarly outputs cited as references in Wikipedia, that is, the latter possesses a broader measurement range. According to Ortega (2018), the differences in Wikipedia data between different platforms are systemic errors caused by coverage issues. The survey by Zahedi and Costas (2018) revealed that Altmetric obtains its data via Wikipedia API. PlumX obtains its data by combining different search engine results, including Wikipedia full-text search and literature citation searches. Therefore, the Wikipedia data value of PlumX is more than that of Altmetric. As data reveals in this study, the average Wikipedia data value of PlumX indeed appears slightly higher than that of Altmetric.

(7) Mendeley Reader Count

The overlapping Mendeley data is 7,859 for the same article among the two platforms. The average Mendeley data value of Altmetric is 70.71, and that of PlumX is 70.60. The standard data deviations of the two platforms are 110.001 and 109.527, respectively. From the view of descriptive statistics, there is not much difference between the two platforms. The Wilcoxon signed rank test result is $Z = -11.313$, $P < 0.05$, so the null hypothesis is rejected, and the conclusion remains that the two matching variables are significantly different.

There is no evident difference in the description of the measurement method between the two platforms in Table 4.1, and both platforms obtain the data directly from the Mendeley API. However, the question remains as to why there is a difference. Ortega (2018) assumed that when PlumX counts the number of Mendeley readers, it may duplicate records of documents of similar titles, years, and authors, while Altmetric counts the number of readers based on a unique identifier and eliminates duplicate entries. Besides, Zahedi et al. (2018) compared the data of the journal PLOS One on different altmetrics platforms and found that 97.9% of the Mendeley counts recorded by Altmetric were the same as Mendeley.com, while PlumX had only 30% of the exact count. In other words, the two platforms may perform differently when processing Mendeley data.

In order to find out whether the Mendeley data of Altmetric is more accurate than PlumX, we further collected data from Mendeley.com (hereinafter referred to as “Mendeley”). The same data sample (N=12000) from Mendeley is downloaded, and we also re-downloaded the data of Mendeley readers on Altmetric and PlumX to unify the time point. In this data cleaning, the Mendeley data excluded the 0 value as-signed data, and only the non-zero data was retained. After ensuring the data acquisition time of the three platforms is consistent, the numbers of Mendeley readers of the three platforms (Mendeley, Altmetric, and PlumX) were compared once more.

The statistics of the overlapping data of Mendeley are shown in Table 4.9. Then, the correlation between the three data sets was tested using the non-parametric test method of Spearman’s Rank, and coefficients are drawn in Table 4.10. The correlation result shows that Mendeley readers on the three platforms have a strong association. Finally, the three sets of Mendeley reader data were tested for significant differences in pairs, and the result in Table 4.11 shows there are significant differences between the three sets of overlapping data. Consequently, no statistical proof clarifies that the number of Mendeley readers of Altmetric or PlumX is consistent with Mendeley itself.

Table 4.9 Statistics of overlapping Mendeley data

	N	Mean	Std. Deviation	Min	Max
Altmetric	6114	81.58	122.192	1	2874
Mendeley	6114	76.82	114.250	1	2868
PlumX	11213	66.69	108.006	1	2765
Mendeley	11213	66.86	111.138	1	2965

Table 4.10 Non-parametric correlation coefficients

	Mendeley	Altmetric	PlumX
Mendeley	1.000	.952**	.950**
Altmetric		1.000	.968**
PlumX			1.000

** . Correlation is significant at the 0.01 level (2-tailed).

Table 4.11 Test Statistics^a

	PlumX - Altmetric	Altmetric - Mendeley	PlumX - Mendeley
Z	-51.635 ^b	-49.980 ^c	-33.832 ^b
Asymp. Sig. (2-tailed)	.000	.000	.000

a. Wilcoxon Signed Rank Test

b. Based on positive ranks.

c. Based on negative ranks.

In our sampling, Altmetric only has 6,114 (51%) reader data that overlap with Mendeley, while PlumX has 11,213 (93.4%) overlapping data. Obviously, PlumX possesses more comprehensive reader data than Altmetric. Furthermore, according to the information on Altmetric's official website, Mendeley data is no longer included in the "Altmetric Score". Therefore, Altmetric has reduced its interest in the Mendeley data. The above facts may explain the presence of missing data on Altmetric even when using the same Mendeley API. We should emphasise that the Mendeley reader counts in Altmetric and PlumX are both from the Mendeley platform, and the number of Mendeley readers of the three platforms is also strongly correlated ($r > 0.95$), so the origins of data sources are the same. Then, when comparing Altmetric and PlumX, it is considered that the latter acts better. However, since

the number of readers on PlumX and Mendeley could not prove any statistical difference, collecting data from the Mendeley platform is preferable in practical application.

In addition, we found that although the reader count in Mendeley does not follow the normal distribution, an approximately normal distribution can be obtained after logarithmic transformation. Table 4.12 shows the statistical information of Mendeley readers after logarithmic conversion on Mendeley data. As seen in Table 4.12, the coefficients of Skewness and Kurtosis are both as small as less than 0.5. The skewness coefficient is -0.324, implying that the variable distribution is left-skewed, that is, the data on the left side of the mean is more than the right side. The kurtosis coefficient is only 0.325, meaning that the distribution is relatively flat. Figure 4.5 illustrates the histogram of the variable frequency, and Figure 4.6 is the Q-Q plot to demonstrate the comparison between the log-transformed distribution and the normal distribution of the Mendeley reader data. The above findings provide an understanding of Mendeley data and reference the subsequent use of such data in statistical analyses.

Table 4.12 Descriptive statistics of natural logarithm of Mendeley reader data

	N	Min	Max	Mean	S.D	Skewness		Kurtosis	
	Statistic	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Std. Error
Log_MenReader	11730	.00	7.99	3.5336	1.25487	-.324	.023	.325	.045

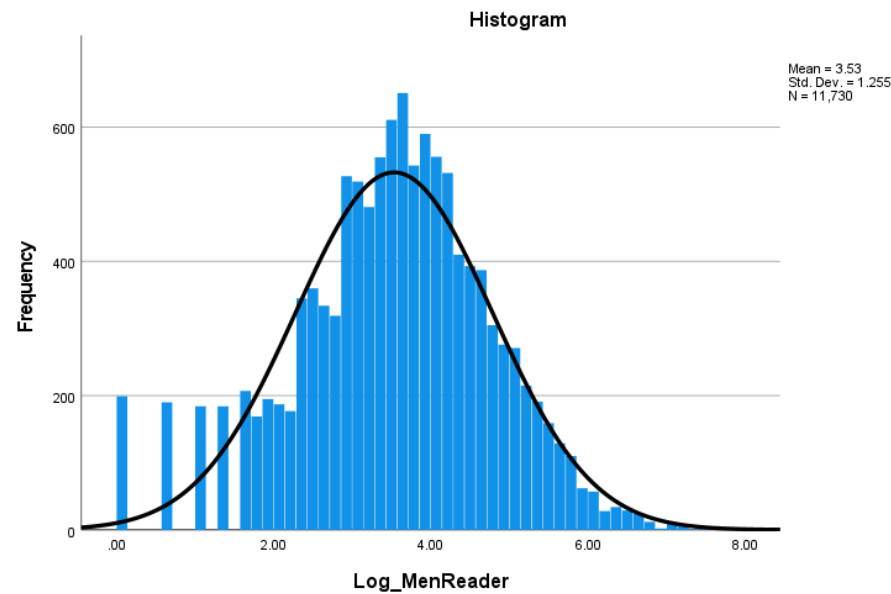


Figure 4.5 Histogram of Natural Logarithm of Mendeley Reader data

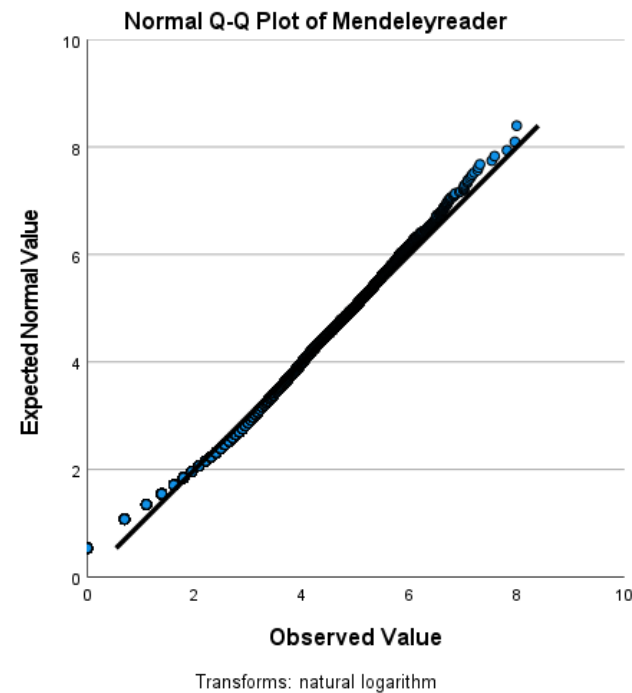


Figure 4.6 Q-Q Plot of Natural Logarithm of Mendeley Reader Data

4.8 Recommendation on the use of altmetrics

Based on the analysis results in Section 4.5 to 4.7, we propose the following suggestions for using the altmetric data of the two platforms.

1. Twitter and News data on Altmetric are significantly better than PlumX in terms of data coverage, data source coverage, and the transparency of data sources' origins. Therefore, using Twitter and News data from the Altmetric platform is recommended.
2. As for the Wikipedia data, the data coverage and data source coverage of PlumX are far better than those on Altmetric. Hence, it is advised to use PlumX's Wikipedia data.
3. In terms of Mendeley data for the two platforms, PlumX is appreciably better than Altmetric, and the Mendeley data for PlumX is recommended. However, the Mendeley data on the two platforms appear statistically significantly different from the Mendeley platform. Therefore, when conditions permit, we recommend directly collecting and employing Mendeley data from its own platform.
4. There is no substantial difference being noted between the CiteULike data of the two platforms in this statistical test, and the correlation seems firm, so either Platform's CiteULike data can be used.
5. The two platforms' Blog data sources have a limited crossover range. In addition, both platforms possess unique data sources. Therefore, the Blog data of the two platforms can be combined to complement each other.
6. The two platforms' Facebook data sources also maintain a limited crossover range, and each one has its own characteristics. However, the two platforms demonstrate very different definitions of Facebook counts. For example, Altmetric only counts the links mentioning scholarly outputs in public posts, while PlumX counts the total shares + likes + comments of links mentioning scholarly outputs. Therefore, it is not convenient and accurate to merge the data of the two platforms, and they can only be selected with discretion.

4.9 Summary

This chapter comprehensively compares and analyses the data quality of two significant altmetrics platforms, Altmetric and PlumX. The study involved a detailed analysis of altmetric data coverage, disciplinary differences, data consistency, and the reliability of the data sources, providing a thorough understanding of the altmetric data sources and platforms. The

findings in this chapter are summarized below.

Firstly, regarding the altmetric data coverage on the two platforms, the coverage of Almetric on social media data, including Twitter, Facebook, Blogs, and News, appears higher than that of PlumX. The data coverage of CiteULike on Almetric also appears slightly higher than on PlumX, while PlumX shows a higher Mendeley and Wikipedia data coverage than Altmetric. By evaluating the proportions of the overlapping data on the two platforms, we found:

- Dominant coverage: The data source coverage of one platform appears superior to that of the other platform, such as Altmetric's Twitter, PlumX's Wikipedia.
- Different coverage: The data coverage of the two platforms limitedly overlaps, but each platform has its own unique data source, such as both platform's Blogs.
- Similar coverage: The data coverage of the two platforms remains similar, such as both platform's CiteULike.

Then, the study compared the coverage of altmetrics across the four disciplines including Humanities, Business Management, Chemical Engineering and Medicine, from the perspective of disciplinary differences. We can see that the coverage of altmetrics data for the medical discipline is the highest. Chemical Engineering and Humanities showed similar coverage in some altmetrics data sources, such as Mendeley and Twitter. However, Chemical Engineering holds better Blogs, Facebook, and News coverage than Humanities and Business Management. Chemical Engineering benefits from more discussions on the mainstream social media than Business Management and Humanities. The coverage of the humanities discipline in Wikipedia, though, is better than that of Chemical and Business Management but is the lowest in the News. Compared to the other disciplines, altmetrics coverage of Business Management appears relatively low, especially the lowest among the three social media data sources: Twitter, Facebook, and Blogs. It is assumed that business-related papers attract less attention on mainstream social media. The difference between various disciplines may require more in-depth research and evaluation.

In the correlation analysis, we discussed the internal relevance of altmetric data on each platform and the correlation of the same altmetric data on both platforms. In the discussion of internal relevance, it is suggested that multiple data sources are moderately or weakly correlated, which indicates that a small number of articles, such as trending articles, are widely discussed on different social media platforms. However, the differences in audiences and communication methods generate differences in the impact reflected by varied social media data. Besides, when an academic article is diffused through different social media networks, it may generate a certain extent of academic impact first since scholars are the pioneers in posting and spreading the article at the beginning, and then some public impact may be generated later. Since different media have different levels of public influence, consequently, a certain correlation between various social media platforms exists.

The correlation analysis between the same altmetric data on different platforms has found that the data of Mendeley, Twitter, News, Wikipedia, and CiteULike for the two platforms displayed a strong association. Notably, the Mendeley data of the two platforms appeared extremely correlated ($r=0.97$). Nevertheless, the correlation coefficients between Blogs and Facebook on the two platforms were 0.576 and 0.413, respectively, showing a moderate correlation. This is attributed to the apparent differences in the data collection of the two platforms. Besides, the coverage analysis also found the two data sources have an inevitable difference in the data coverage, as illustrated in Figure 4.3, affecting the correlation coefficient of the two data sources.

Moreover, to discuss whether the overlapping data of the two platforms appear consistent, we employed the non-parametric Wilcoxon Signed Rank test to compare if any significant differences exist between the data of the two platforms. The results signify no significant difference between the CiteULike data of the two platforms, while all the other altmetric data show significant differences between the two platforms. It indicates that the CiteULike data of the two platforms belongs to the same population, and the difference of a small amount of data value is caused by random errors. The differences between other altmetric data

sources could be due to factors including different ways of data counting (definitions), data processing, and different coverage of data sources. Section 4.7 showed the detailed analysis of reasons for heterogeneity between the two platforms. When studying the data of Mendeley readers, we further collected the data from Mendeley.com and compared three platforms. Data consistency has yet to be found. However, we found that Mendeley reader data follows a lognormal distribution, providing a reference for using such data in subsequent statistical analyses.

Finally, based on the results of the analysis, we proposed recommendations on how to use altmetric data sources among the Altmetric and PlumX. We should also address that the current results and analysis are limited to the specific situation of this paper's data sample, so further research on different samples may be needed in future.

Chapter 5 - An evidence-chain-based approach to evaluating scholarly impact

(Part of the content of this chapter has been published in Lin, D., Gong, T., Liu, W. et al. (2020). An entropy-based measure for the evolution of h index research. Scientometrics, 125, 2283–2298. <https://doi.org/10.1007/s11192-020-03712-1>)

5.1 Introduction

The evaluation of scholarly research impact is a hotspot and a challenging issue in the management of science and technology (see Section 2.3). Apart from the qualitative evaluation method in the form of peer reviews, the citation-based scientometrics evaluation methods prevail in the research evaluation (see Section 3.1). However, citation-based evaluation methods suffer from the limitations of indicators. The declined quality of citation data is another issue caused by complicated citing motivations (see Section 1.1). Exploring new methods has been called for in academia (Bornmann & Haunschild 2016).

In the big data era, owing to the rapid circulation and dissemination of information, altmetrics, which is based on social network data (under multi-source data), has risen gradually and aroused much concern in Scientometrics (see Section 3.4 & 3.6). The availability of altmetric data has increased the possibility of using multi-source data in the research evaluation and inspired us to carry out quantitative evaluation approaches with a wider range of data sources. We have studied the altmetric data quality and data source provider in-depth in Chapter 4, which aimed to guide us to apply altmetric data properly in later research. This chapter will discuss an evidence-chain-based quantitative aggregation method for scholarly research impact by using impact disseminating evidence chains with internal logic and supporting data. Both citation and altmetric data will be used. This proposed methodology increases the reliability of the quantitative evaluation output and could better serve science and technology management.

5.2 Research background

With regard to quantitative methods and tools for academic impact evaluation, theories and empirical studies using bibliometric indicators (citation-based) have been studied in most of the literature (see Section 3.2). In respect of academic impact evaluation specific to a research unit (generally a researcher, a research team, a research department, a research institute or a periodical, etc.), quantitative evaluation indicators have evolved from being simple to complex, and from being single to compound (Booyesen, 2002). Traditional citation indicators typically include total cites, highly cited papers, cites per paper and proportion of highly cited papers, which mainly evaluate the academic impact of a research unit within a certain period of time and are the most basic and simplest quantitative evaluation indicators. The *h*-index proposed in 2005 by Hirsch, a professor of physics from the University of California, which takes into account both quantitative and qualitative evaluation, has triggered a new upsurge of research on citation indicators in academia and then promoted wider application of the *h*-index and research on its derivative indexes (see Section 3.2).

Beyond the enthusiasm for studies on citation-related evaluation indicators and methods, the quality of citation data has gradually grabbed the attention of academia. From the perspective of the cause, factors that influence the quality of citation data include non-normative citation conducting and data quality issues resulting from complicated citing motives of citing authors (Bornmann & Daniel, 2008). Brooks (1985) enumerated 26 types of citer motivations and identified the complexity of citer's behaviours, including citing for persuasiveness, positive credit, social consensus, personal bias, citations for flattery, biased citations for logrolling, improper citations to support the interests of a certain academic research group, citations catering to editors, superfluous citations, etc. Liu et al. (2018) studied 337 duplicate publications in 198,518 bibliographic references to documents published from 2001 to 2010 under "Library and Information Science and Digital Libraries" in the CNKI database and analysed the citing motives of researchers in the field of LIS in China. Their results showed that when two papers are of similar quality, the one published in a core journal will be cited more. It reveals that when the quality of papers is effectively controlled, the citing

motivations of researchers clearly shift to respect for authority. Furthermore, Reedijk and Moed (2008) have summarised several practical manipulating methods for journal impact factors, including tactical edition, excessive self-citations, paid citations, citation stacking and so forth.

It is clear that when stimulated by complicated citing motives, especially improper ones, quantitative evaluation of academic impact with citation data alone will probably go against the research evaluation principle of being fair and equitable, scientific and rational, and may have significant negative effects on the growth of researchers and research management. We need to emphasise that such an issue is shared by most of the result-data-only approaches. Because all desirable results (like citations) can be produced by improper actions/processes such as cheating. Thus, to increase the reliability of evaluation, it is essential to evaluate not only the result but also the generating process. For example, for a regular highly cited publication, there should be a series of evidence on high abstract/views, downloads, commentaries or discussions. When such evidential data are small, we may doubt the reliability of citation counts. Consequently, if we can build an evidence based process chain model for the academic impact, we should be able to produce more reliable evaluation results.

Moreover, previous literature and research results (in Chapter 3) showed certain consistency between the citation-based indicators and altmetric data based altmetrics in evaluating academic impact. Part of altmetrics presented a weak positive correlation with citation indicators. This provides an inspiration that these two types of data can complement each other in academic impact evaluation.

Given the above background, we aim to develop an evidence-chain-based aggregation approach that considers both the result of the impact and its generating process. In our evaluation approach, we will apply citation and altmetric data as behaviour evidence and form evidence chains with an internal logical relationship.

5.3 Evidence-chain-based quantitative evaluation method

5.3.1 Evidence chain and cause-effect chain

First of all, we shall discuss what the evidence chain and the cause-effect chain are. From the perspective of judicial criminal investigation, an evidence chain refers to a group of evidence consisting of more than two nodes that are objectively connected and mutually corroborated with respect to the contents, and reflecting or enhancing the probative force of scattered evidence (Chen, 2007). In a broad sense, the evidence that constitutes the evidence chain refers to the elements used to support a conclusion, while in a narrow sense, they refer to elements recognized in criminal, civil and administrative procedures in laws (Bex, 2011). According to Chen (2007), the evidence nodes and connectors are basic components in an evidence chain, and the nodes may be composed of either a single evidence or an evidence cluster pointing to the same object: different evidence nodes are linked in a chain by connectors. The connectors of the judicial evidence chain are generally explicit (such as forensic authentication and weapons), while connectors for nodes of the evidence chain in general reasoning processes may also be implicit such as the formal logic, social laws and subject knowledge. At present, relevant studies on the evidence chain exist mostly in the field of jurisdictive criminal investigation, and the concept of the evidence chain is occasionally mentioned in clinical research (Cook et al., 1997; Mulrow et al., 1997; Jiabo Wang et al., 2015). Using evidence or a chain of evidence approaches in the evaluation of research impact is very rare.

This thesis argues that the causal relationship (Mingers & Standing, 2017), which is a relationship between a specific process earlier in a time series (cause) and a specific process or status later in another time series (effect), is the base to construct an evidence chain. According to the Theory of Critical Realism proposed by Roy Bhaskar (Bhaskar, 1997), the causal relationship exists in three domains from the shallower level to the deeper level. First is the empirical domain, which contains sensory experience. Second is the actual domain, which contains perceivable events and things. Third is the real domain, which contains

determinate structures and mechanisms that humans cannot directly perceive through experience.

This thesis uses the causality view of critical realism to understand causality from multiple perspectives. That is, the true relationship between the cause and the effect exists only in the real domain and cannot be directly perceived through human experience. Specific causes and effects can be connected in the actual and empirical domains through cause-effect chains that take different forms and can be recognized by human experience. The cause-effect chain is a specific form for depicting the cause-effect relationship producing particular effects. It includes a series of specific causes leading to the effect and displays the logical relationship between the causes. In an open system, the number of cause-effect chains is infinite, and even the causes of a single cause-effect chain can be infinitely divisible. Therefore, in practice, people often combine cognitive purposes, cognitive capabilities, cognitive habits and other factors and use a limited number of cause-effect chains and attributions of appropriate granularity for cognition (Dobrusskin, 2016; Wang, 2016). As shown in Figure 5.1, the same abstract result (effect) may be caused by an infinite number of cause-effect chains, such as cause-effect chain 1, cause-effect chain 2 ..., and cause-effect chain n, which have different attribution granularities.

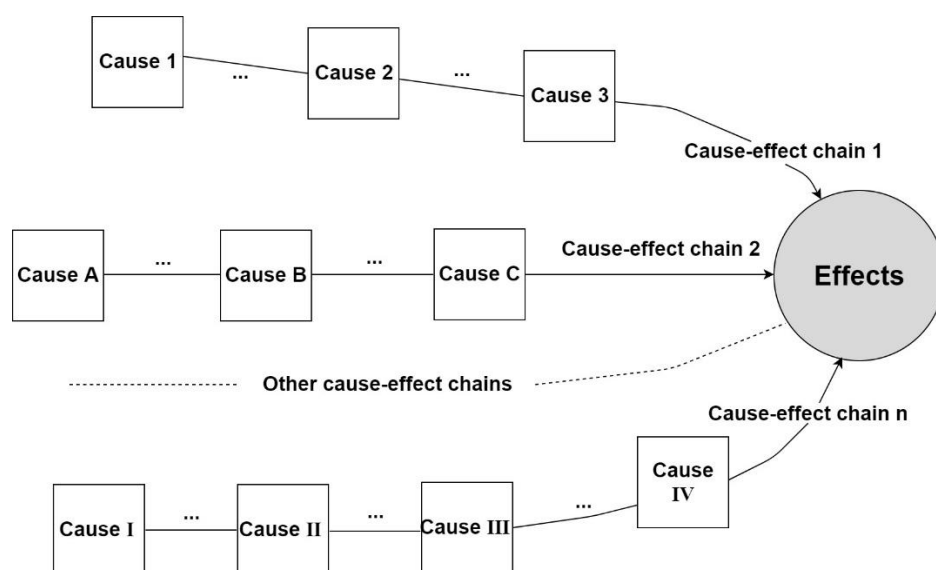


Figure 5.1 The schematic diagram of causality and relationship between cause-effect chains

After clarifying the relationship between causal relationships and cause-effect chains, we then investigate the evidence and the evidence chain. As shown in Figure 5.1, the effectiveness of causes 1, 2 and 3 on the cause-effect chain 1 can be judged based on the probability strength of the corresponding evidence, and the specific correlation between the evidence nodes of each cause constitutes the evidence chain. Therefore, the evidence chain is used to determine the possibility that a particular cause and a particular effect are (or are not) associated with a specific cause-effect chain. The evidence nodes on an evidence chain are not simply listed in the form of a single line but form several evidence clusters in series-parallel connection relations based on many causes of the cause-effect chain. Taking Figure 5.2 as an example, the single-line evidence clusters (a_1, b_1, c_1) surrounding cause 1 can only strengthen the power of judging cause 1, but cannot directly enhance the power of judging cause-effect chain 1. Only when the evidence clusters (a_1, b_1, c_1) , (a_2, b_2) and (a_3, b_3, c_3) can be linked to each other for mutual corroboration, then the evidence chain can be formed to further enhance the overall power of judging cause-effect chain 1.

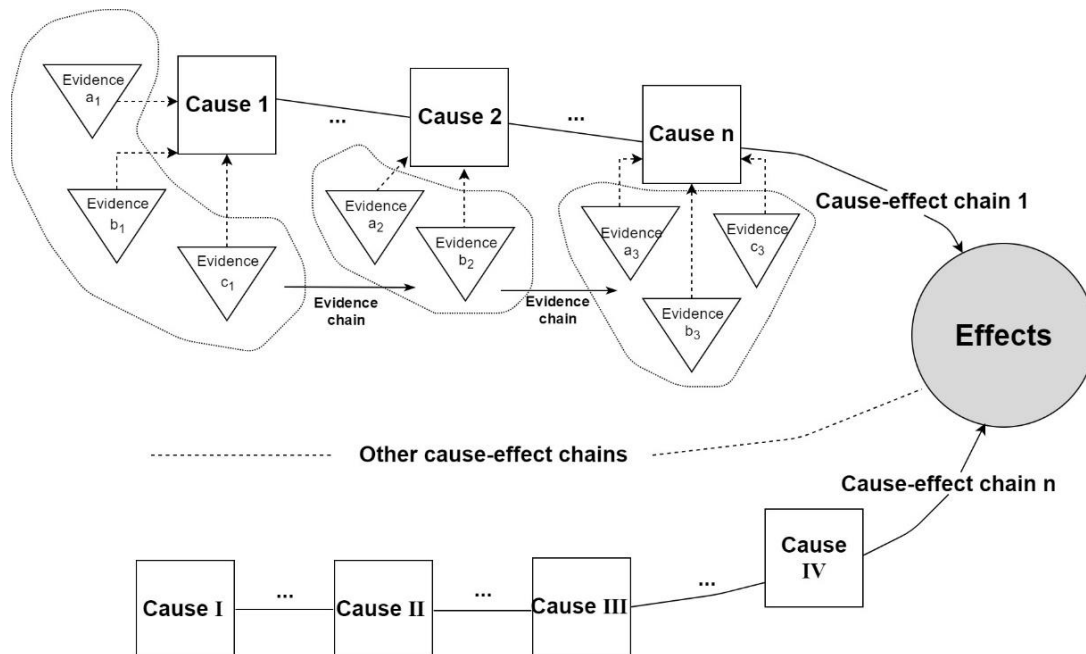


Figure 5.2 The schematic diagram of relationship between cause-effect chains and evidence chains

5.3.2 How to build an evidence chain?

This thesis proposed that an evidence chain should be built in the following three steps:

The first step is to determine the cause-effect chain that to be verified. In an open system, there are countless chains linking a specific cause to a specific effect, but we often can make clear the specific way of connection between a cause and its effect, that is, the cause-effect chain, according to existing natural or social science theories. In the absence of the above-mentioned theories or in case of inadequate consistency or interpretation of existing theories, we may use cognitive tools such as Soft Systems Methodology (Checkland & Poulter, 2010) to build an abstract logic-based cause-effect chain.

The second step is to identify evidence based on the cause-effect chain. Evidence (or groups of evidence) can be organized based on physical, social, and data signs of attribution elements on the causal chain. Generally, if high-weight direct evidence exists, we may select to support a specific attribution element with a single piece of evidence; if only indirect evidence exists, we often increase the weight of an attribution element with a group of indirect evidence.

The third step is to link evidence (or groups of evidence) to form an evidence chain. Here, we have to inspect the relationship between different evidence (or groups of evidence) to guarantee no conflict or irrelevance between the evidence (or groups of evidence). Such inspection can be carried out according to existing consensus theories or practices. In the meantime, process tracing or data mining can be used to verify the degree to which different evidence (or groups of evidence) support each other.

5.3.3 Primary evaluation models based on the evidence chain

The characteristics of the evidence chain can be summarized as follows: due to the interdependent nature of nonlinear evidence nodes, the evidence chain can generate a higher level of credibility than merely adding up the individual pieces of evidence. In other words,

the whole is greater than the sum of its parts. Therefore, in this study, the issue of evaluating the academic impact of the evidence chain can be transformed into an estimation of the overall probability of occurrence of the evidence chain for the academic impact, so as to further estimate the overall probability of occurrence of the cause-effect chain of academic impact. Estimating the overall probability of occurrence of the cause-effect chain in a systematic manner is a complex statistical problem. Below, we first discuss one basic estimation method.

We assume that there are only two causes, namely, cause A and cause B, on a cause-effect chain; $P(A)$ and $P(B)$ represent the probability of occurrence of cause A and cause B, respectively; $P(AB)$ represents the probability of simultaneous occurrence of cause A and cause B. Generally speaking, the product of probabilities of occurrence of each cause can be used to estimate the probability of occurrence of a particular cause-effect chain. However, because there is a logical relationship between cause A and cause B, the estimation of the total probability (i.e. $P(AB)$) of occurrence of the cause-effect chain must be obtained through a synthesis based on the Bayesian method rather than a simple probability product. Therefore,

$$P(AB) = P(A)P(B/A) \quad (5.1)$$

Where, $P(B|A)$ represents the probability of occurrence of cause B in the case where cause A occurs. It should be noted that $P(B|A)$ refers to the conditional probability under cause A, so $P(B|A) \leq P(B) \leq 1$. Therefore,

$$P(AB) = P(A)P(B)^\alpha, \alpha \geq 1 \quad (5.2)$$

This suggests that the power of $P(AB)$ can be estimated based on the following variables:

$$P(A)^\beta P(B)^r, \beta, r \geq 0, \beta + r = 1 \quad (5.3)$$

Because the cause-effect chain is theoretically unobservable, it is necessary to estimate likelihood through formula (5.3) based on evidence data in the actual calculations. Therefore,

$L(P(AB)) = \left(L(P(A))\right)^\beta \left(L(P(B))\right)^r$. Where L represents the estimation of corresponding probability likelihood based on evidence data. If there are N causes, including causes A, B, C...N, on a certain cause-effect chain, the overall likelihood L of occurrence of the cause-effect chain can be estimated based on the following variables:

$$L = P(A)^\alpha P(B)^\beta \dots P(N)^\mu$$

$$\text{where } 0 \leq \alpha \leq 1 \dots \dots 0 \leq \mu \leq 1, \text{ and } \alpha + \beta + \dots \mu = 1 \quad (5.4)$$

The weighting factors need to be further estimated for different application scenarios, which is not discussed in detail here. A specific application of evidence chain based on the publication citation is discussed below.

5.4 Empirical study - A case study of the impact evaluation of Journal A and Journal B

Journal A and journal B are top journal publications with international influence in the discipline of management science and operations research. Although the impact factor ($JIF_{2018} = 4.311$) of journal A is always significantly higher than that of journal B ($JIF_{2018} = 3.544$), most experts in relevant fields recognize that the quality of journal B is higher than that of journal A. In this empirical study, we evaluated the academic impact of journal A and journal B based on the evidence chain that to explain the views of scientific community.

5.4.1 Construct the cause-effect chain and evidence chain for academic impact

As outlined in Section 5.3.2, before identifying an evidence chain, we must first explore the cause-effect chains of these two journals' academic impact. Generally, we use a journal's number of citations to express its academic impact. Therefore, the journal's citation number is the "effect". Prior to follow-up investigations or systematic research, we cannot recognize the exact purpose of the citation by an author as the causes may vary. In other words, it is difficult to identify which specific "cause" leads to that "effect". Weinstock (1971) has summarized 15 specific citing motivations. Different citing motivations (causes) may lead to the same "effect" of citation; thus, different "cause-effect chains" can be constructed. For example, causes such as inappropriate mutual citation by associates and friends may lead to this effect as the citation; a quid-pro-quo favour for editors also leads to this "effect".

We take Weinstock's citing motivations as examples to draw Figure 5.3 and assume the existence of 15 cause-effect chains. While if a scholar read a piece of research work by any chance, understood well the knowledge involved in the paper, then disseminated this research work and cited this work in later writings, such "cause" based on the chain of "read-understood-disseminated-cited" also leads to the "effect" of citation, we assume it is the 16th

cause-effect chains in Figure 5.3. Obviously, the cause-effect relationship in Chain 16 is more in line with the normal logic behind a citation, which is a consensus in the academic community regarding this logic (Fry et al., 2009). We have good reasons to believe that the number of citations reflecting the academic impact would be more authentic and reliable if they were produced through this chain.

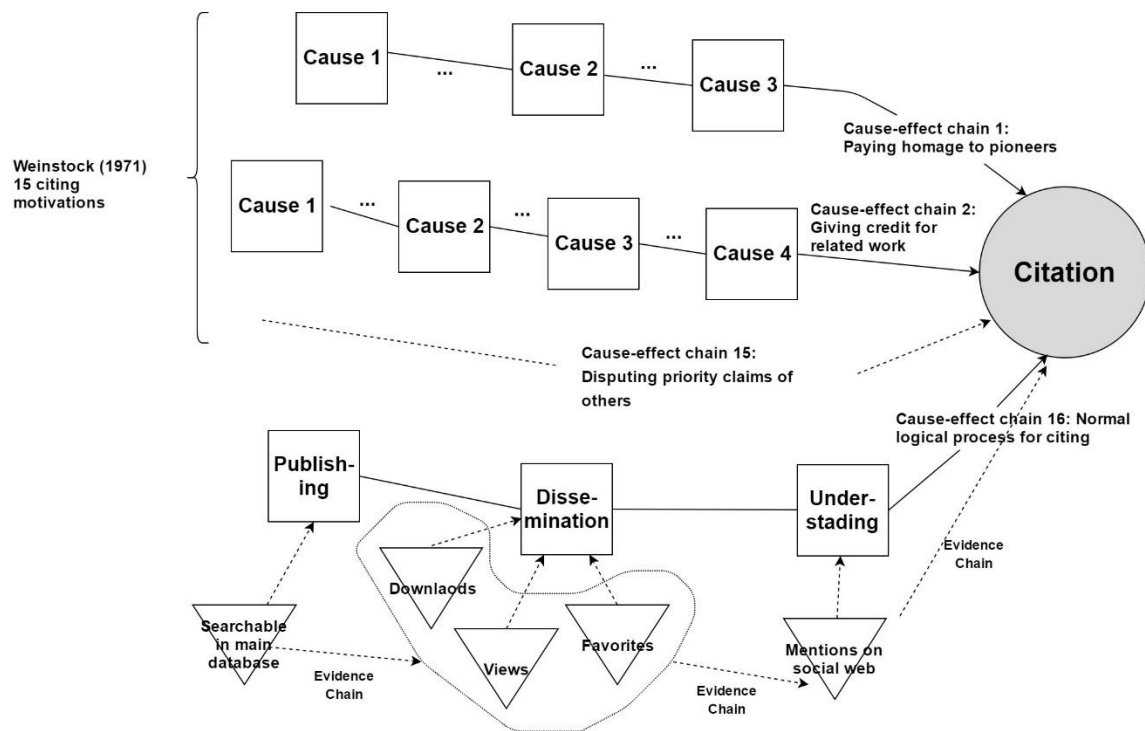


Figure 5.3 The schematic diagram of cause-effect chains and evidence chains for academic impact

The next step is to explore evidence and evidence chains to support the chosen cause-effect chain, the 16th cause-effect chain in Figure 5.3. Benefiting from the technological development in the age of big data and the increasing trend of online communication among scientific researchers, the evidence supports the “read-understood-disseminated-cited” cause-effect chain could be traced from a series of behaviour data such as the number of views, downloads, reads, comments, citations and other manifestations of research work. The above evidence data can be linked according to the internal logic of their mutual corroboration to build up an evidence chain for the academic impact. Then, we can use this

evidence chain to judge and verify the likelihood of the occurrence of a specific “read-understood-disseminated-cited” cause-effect chain.

Therefore, the overall probability of generating a citation result by a specific causal chain can be estimated through the collected evidence data (in accordance with relevant evidence chains), calculations of probabilities, and tests of parameters. Accordingly, we could quantitatively evaluate academic impact based on the built-up evidence chains.

Additionally, if one piece of research work has not been read, downloaded, marked as a favourite, or discussed before and suddenly been cited in a large number without any earlier sign, we would have reasons to doubt the legitimacy of the cause-effect chain leading to the citations. The contribution of the number of citations generated by such an undesirable chain should be discounted to reflect a reliable academic impact. Notably, “Sleeping Beauty” publications (van Raan and Winnink, 2019), which are recognised as “delayed recognition” literature, are exceptional and not included in this situation.

5.4.2 Evidence data collecting and processing

The objects of this empirical study are journal A and journal B, respectively. We initially assumed the citations of these two journals are produced through the normal “read-understood-disseminated-cited” cause-effect chain. If “number of readers” and “number of comments” are regarded as legitimate evidence (or evidence clusters) for the “read-understood-disseminated” sub-chain, then “readers-comments-citations” constitutes an evidence chain for our favourable cause-effect chain. Therefore, we first need to collect the relevant data corresponding to the evidence chain, including the number of readers, comments, and citations for each article in observing journals.

- **Number of readers:** This means the number of readers for the research output. A research paper could be viewed or read by thousands of readers. However, a completed reader number is difficult to record and almost impossible to obtain. Most scholars like to save the interested research papers using reference management tools to facilitate their future reading or subsequent use, such as citing. The reader number

of a specific piece of research work recorded by the reference management software could become evidence proving the solid number of readers for the research to a certain extent. Currently, Mendeley is the most widely used online reference management tool for users to manage and share scholarly materials in the academic community (see Section 3.5.3). The scholarly output in Mendeley is uploaded by users themselves into the “Mendeley library” for editing and managing. Many studies have proved that the Mendeley reader count significantly correlates to the number of citations (see Section 3.6.1). Therefore, this case study will collect Mendeley reader count as the evidence data for the number of readers of targeted publications.

- **Number of comments:** during the dissemination of research output, readers may spontaneously disseminate, comment, or forward articles on social networks or online media for different purposes. Marks of sharing, commenting, forwarding, and mentioning constitute the evidence to prove that such a piece of research work has been disseminated and is concerned, as well as the secondary evidence proving the contents were understood. In practice, these data evidences are small in numbers and parallel to each other; hence, we can combine them to form an evidence cluster. We used the sum of the numbers of discussions and mentions on Twitter, Facebook, News, Blogs, Wikipedia, Google+ and F1000Prime as the evidence data.
- **Number of citations:** we will use the number of citations provided by Web of Science, which is the most popular citation database, as the evidence data (Ezema, 2017). As the world’s most authoritative indexing tool for scientific and technical literature, Web of Science covers more than 8,700 core academic journals with the most significant impact in various research fields such as natural science, engineering technology and biomedicine. It is highly accurate, normative and authoritative (Repiso et al., 2019b).

Regarding the databases for data collection, the mainstream altmetrics platforms that provide altmetric data include Altmetric.com (Altmetric), Plum Analytics, PLOS ALMs,

ImpactStory and so on (see Section 3.5.2). Considering the slow operating speed and unstable login for ImpactStory, and the fact that PLOS only provides Article-level metrics for its own journals, we mainly consider collecting altmetric data from the two most used platforms, Altmetric and PlumX. In Chapter 4, we systematically discussed the quality of altmetric data on Altmetric and PlumX. According to our findings in Chapter 4, we will collect comment data from Altmetric, including the data sources of Facebook, Twitter, Google+, F1000Prime, News, Blogs and Wikipedia. Mendeley reader data will be collected from the Mendeley platform, and citation data will be collected from the Web of Science.

After analysing and filtering the sources of evidence data, the next stage is data collection and data cleaning. We initially collected the basic bibliographic information and citation count of articles in Journal A and Journal B for 2015 from the Web of Science. Then, we used the DOI number of each paper to gain Mendeley reader data from Mendeley and altmetric data from Altmetric. The consolidated data information of the journals A and B is shown in Table 5.1.

Table 5.1 Data information of Journal A and B

	Journal A	Journal B
<i>Number of journal papers</i>	117	173
<i>Number of readers</i>	3488	6451
<i>Number of comments</i>	40	663
<i>Average citation count</i>	13.32	8.06
<i>Total citation count</i>	1559	1395

5.4.3 Evaluation results and analysis

After the above steps, we obtained the data of each evidence node on the “readers-comments-citations” evidence chain for the academic impact of Journal A and Journal B. According to formulas (5.3) and (5.4), we need to estimate the likelihood of the occurrence of the cause-effect chain with normal citations in order to evaluate the academic impact of the journals. We applied two scenarios for estimating the likelihood of the occurrence of the citation cause-effect chain for journals A and B. The scenario 1 estimated the likelihood

according to the number of readers, comments and citations of a single paper, and then calculated the average value to gain the average likelihood of occurrence of the citation cause-effect chain for the journal, which is expressed in formula (5.5) as follows:

$$L = \sum_{i=1}^N (r_i^\alpha * d_i^\beta * c_i^\mu) / N$$

(5.5)

where $\alpha \geq 0, \beta \geq 0, \mu \geq 0$, and $\alpha + \beta + \mu = 1$

In formula (5.5), $i=(1,2, \dots, N)$ represents the i th paper published in the journal, r_i represents the number of readers of the i th paper, d_i represents the number of comments of the i th paper, and c_i represents the number of citations of the i th paper.

The scenario 2 totalled the numbers of readers, comments and citations of all papers in one journal to obtain the numbers of readers, comments and citations of the journal, and then calculated the overall likelihood of occurrence of the citation cause-effect chain for the journal, which is expressed in formula (5.6) as follows:

$$L = \left(\sum_{i=1}^N r_i \right)^\alpha * \left(\sum_{i=1}^N d_i \right)^\beta * \left(\sum_{i=1}^N c_i \right)^\mu$$

(5.6)

where $\alpha \geq 0, \beta \geq 0, \mu \geq 0$, and $\alpha + \beta + \mu = 1$

In formula (5.6), $i=(1,2, \dots, N)$ represents the i th paper published in the journal, r_i represents the number of readers of the i th paper, d_i represents the number of comments of the i th paper, and c_i represents the number of citations of the i th paper.

We call the above L as the likelihood value of academic impact based on the evidence chain. The parameters α, β, μ in formulas (5.5) and (5.6) need to be further estimated for different application situations, and this study only selected the most representative parameters to compare the calculation results of probability for the two journals (refer to Table 5.2 and Table 5.3). In the meantime, in order to further compare the evaluation results of the academic impact of the two journals based on the evidence chain, we standardized the

calculation results by calculating the ratio of the academic impact of the two journals. The results are shown in Tables 5.2 and 5.3.

Table 5.2 Academic impact measurement results based on evidence chain for Journal A and Journal B (Scenario 1)

	L1	L2	L3	L4	L5	L6	L7	L8
Weights	$\alpha=0.2$; $\beta=0.6$; $\mu=0.2$;	$\alpha=0.6$; $\beta=0.2$; $\mu=0.2$;	$\alpha=0.2$; $\beta=0.4$; $\mu=0.4$;	$\alpha=0.4$; $\beta=0.2$; $\mu=0.4$;	$\alpha=0.1$; $\beta=0.3$; $\mu=0.6$;	$\alpha=0.3$; $\beta=0.1$; $\mu=0.6$;	$\alpha=0.05$; $\beta=0.15$; $\mu=0.8$;	$\alpha=0.15$; $\beta=0.05$; $\mu=0.8$;
Likelihood of occurrence of Journal A (L_A)	0.74%	2.76%	1.10%	2.20%	1.29%	2.64%	1.88%	2.73%
Likelihood of occurrence of Journal B (L_B)	4.92%	12.38%	5.14%	8.38%	4.49%	7.50%	4.62%	6.05%
L_B / L_A	6.65	4.49	4.66	3.82	3.48	2.84	2.46	2.22

Table 5.3 Academic impact measurement results based on evidence chain for Journal A and Journal B (Scenario 2)

	L1	L2	L3	L4	L5	L6	L7	L8
Weights	$\alpha=0.2$; $\beta=0.6$; $\mu=0.2$;	$\alpha=0.6$; $\beta=0.2$; $\mu=0.2$;	$\alpha=0.2$; $\beta=0.4$; $\mu=0.4$;	$\alpha=0.4$; $\beta=0.2$; $\mu=0.4$;	$\alpha=0.1$; $\beta=0.3$; $\mu=0.6$;	$\alpha=0.3$; $\beta=0.1$; $\mu=0.6$;	$\alpha=0.05$; $\beta=0.15$; $\mu=0.8$;	$\alpha=0.15$; $\beta=0.05$; $\mu=0.8$;
Likelihood of occurrence of Journal A (L_A)	2.03%	12.15%	4.23%	10.34%	5.63%	13.76%	9.37%	14.65%
Likelihood of occurrence of Journal B (L_B)	12.13%	30.13%	14.07%	22.18%	13.01%	20.50%	13.47%	16.91%
L_B / L_A	5.96	2.48	3.33	2.14	2.31	1.49	1.44	1.15

From the results shown in Tables 5.2 and 5.3, it is observed that the two scenarios presented the same trend: Journal B is more likely to generate the effect of citations through our observed cause-effect chain, “read-understood-disseminated-cited”, by showing a much higher likelihood of occurrence on our evidence chains. In Scenario 1, the likelihood ratio for individual papers in Journal B and Journal A shows the former has approximately two to six times over the latter occurred through our evidence chain when applying different parameters. In Scenario 2, the likelihood ratio between the two journals has a lower scale

compared to the individual paper when applying the same test parameters. However, Journal B still has a higher likelihood of occurrence than Journal A.

From the statistical point of view, Journal B has a much higher likelihood of occurrence of citations from the path as we recognized that the chain of “read-understood-disseminated-cited” compares to Journal A. Therefore, we can conclude that the citation results of Journal B are more reliable. It could explain the scientific community’s common views that Journal B has a higher academic impact than Journal A if the two are compared.

5.5 Limitations of the study

It should be pointed out that the evidence-chain-based impact evaluation method depends on screening and processing massive data, identifying evidence and improving an evidence chain, which is not emphasised by this chapter. As an emerging impact indicator, altmetrics has ambiguities and overlaps in data collection and statistics. Altmetric data must still be used with caution. Additionally, the empirical study in this chapter probes into the simplest case, with certain limitations on parameter testing and a limited sample size. However, these problems are expected to be technically solved with progress on open data so that the evaluation method can be further improved.

5.6 Summary

With the rapid development of scientometrics, it is possible to utilise altmetric data to overcome shortcomings of the current research evaluation process, such as the results-driven perspective only. This chapter presented an innovative, evidence-chain-based evaluation method that effectively integrates all types of information across various stages of academic impact generation, thereby improving the reliability of final outcomes. This method breaks the limitation of unstructured, fragmented evidence in research evaluation. Complementing diverse pieces of evidence in the evidence chain increases the completeness and reliability of evaluation information and the accuracy and objectivity of evaluation results. The research findings involved in this chapter are highlighted below:

- This chapter conducted a discrimination of three core concepts in the evaluation, namely causality, cause-effect chain and evidence chain, and their relations from the perspective of the causality of critical realism (C.R.), and then built a logic framework for impact evaluation. Based on this framework, we believe that the academic impact, as a result (effect), may have multiple causes, representing a typical case of a “multiple causes to one effect” situation. Thus, academic impact evaluation should focus on discriminating and justifying the compliance and morality of the specific causal chain that results in an effect, in addition to the final results. The evidence chain built based on a specific cause-effect chain is an effective tool for such discrimination and judgment. This chapter has proposed the basic steps and methods to construct an evidence chain.
- Regarding the selection of evaluation indications and data, this chapter proposed building an evidence chain based on the specific cause-effect chain that generates the academic impact and includes both altmetric and classic scientometric indicators that reflect the characteristics of academic impact formation in the evaluation system. It is a combined use of bibliometric and altmetric data in the impact evaluation and a new type of application of altmetric data in the field. Our research also pioneered using multiple data sources in the quantitative evaluation of academic impact.
- Concerning quantitative evaluation measurement, as non-linear evidence on an evidence chain can support each other to increase the weight of evidence, this chapter takes an innovative perspective that turns academic impact evaluation into an estimation of the overall likelihood of the occurrence of the academic impact evidence chain to further calculated the possibility of the occurrence of the cause-effect chain of academic impact.
- In the empirical research, this chapter expounded on the application of the evidence-chain-based evaluation method by evaluating the academic impact of journal A and journal B. We have preliminarily studied how to build an evidence chain for specific research output from publication to impact generation and presented how to carry out an evidence-chain-based quantitative evaluating approach to evaluate academic

impact on this basis. The results from the empirical study in this chapter better match the scientific community's knowledge and highlight the method's advantages.

Chapter 6 - Evidence chain of impact of research publications

6.1 Introduction

As discussed in previous chapters, the research impact study significantly improves the research management level and promotes research innovation. In current studies on quantifying research impact, the evaluation system developed based on the number of citations is the chief focus in the scientometric field. Bibliometrics, one of the leading research methods for analysing the impact of research output, has evaluation indexes that cover the Journal Impact Factor, *h*-index, and many other citation-based indicators for measuring the quality of academic papers, the rank of journals, and the level of researchers (see Section 3.2). Citation-based research impact evaluation actually stresses academic influence.

However, research impacts of output and activities are not limited to academics but are oriented to the whole society. Benneworth and Olmos-Peñuela (2024) suggested that social impact goes beyond academic research as it encompasses the impacts on education, society, culture or economy. Like academic impact, social impact requires proof rather than assumptions. Tahamtan and Bornmann (2020) explained the connotation of the social impact of research as the impact, benefits and changes that scientific research brings to areas beyond academia, such as the economy, society, culture, public policy or services, health, living environment or quality of life. Bornmann et al. (2019) pointed out that the term “impact” in scientometrics seems to undergo a taxonomic change where impact is no longer defined as the impact on science alone but on all sectors of society (e.g. economics, culture, or politics). Several existing international social impact assessment models evaluate the impact of research from multiple dimensions, such as economy, culture, and environment. UK’s REF, however, evaluates social impact by a separate case study (REF, 2021). The Research Quality Framework proposed by Australia defines research excellence in terms of social contribution (Lacey, Coates and Herington, 2020).

A comprehensive research impact involving academic influence and broader impacts (economy, society, culture, public policy or services, health, and more) is challenging to measure as a whole. However, the diversity of today's data types makes it possible to use a wide range of data and sources in quantitative research evaluation. For example, altmetrics provides a new perspective on viewing research impact through scientific information usage (see Section 3.6). Researchers' actions can be recorded online, such as searching, downloading, and saving a research output. We can identify academic behaviours digitally. Therefore, obtaining the evidential data in the process of impact generating for the research output is possible, further inspiring us to construct an evidence-chain-based impact assessment method.

In Chapter 5, we recognised citation as the "effect", citation motivations as multiple "causes", and the citing process as "cause-effect chains". We analysed the evidence chains in the specific cause-effect chain, "read-understood-disseminated-cited," and validated that results from evidence-chain-based evaluation are more in line with opinions from peer reviews. The next step is constructing the cause-effect and evidence chains for comprehensive research impact.

Clearly, a comprehensive impact of research output is compounded by multiple cause-effect chains that generate impact. Clarifying the cause-effect chains that pose impacts and the logical relationships between the causes in each chain can help us evaluate the impact of research output more reliably. We believe that one should clearly define the stages of research output, from knowledge-generating to impact posing, and then evaluate its impact (including various dimensions of impact) by tracing the evidence from the process of impact generation. This is the first time in the literature that evidence from research impact generation has been utilized for impact evaluation. Below, we will develop an evaluation method of research publications for research impact based on the cause-effect and evidence chains.

6.2 The cause-effect chains of impact of research publications

The study on the research impact of publications should be based on combining qualitative and quantitative analysis, and the cause-effect chains of impact are the keys to analysing impact when using our new evaluation approaches. It is essential to identify the causal chains in the impact generation process first. Then, by illustrating the impact generation process, we may find a clear and specific way to describe those causes and effects. Therefore, we can identify evidence based on the cause-effect chain with the help of accessible and reliable databases.

It is argued that the concept of research impact is often broken into stages that align with the lifecycle of knowledge development, including the stages of knowledge generation, meaning the creation of new insights or discoveries through research activities, experiments, studies, and theoretical development; the knowledge dissemination or spreading, which sharing the generated knowledge with a broader audience; the knowledge application or regeneration, which involving adapting or integrating the knowledge into practical solutions, innovations, or further research that regenerates the cycle (Evans et al., 2014).

The research impact generation process can be seen as a system problem. It can be addressed by redesigning the systems for key activities through discussion and deliberation among the primary stakeholders, ensuring their interests are aligned and balanced (Wang et al., 2015). Here, we adopted the stakeholder identification theory derived from the soft system methodology to construct the stages and process of impact generation for research publications. The methodology starts by establishing the system's top-level objectives through stakeholder discussions, and then the objectives are systematically broken down into the essential key activities required to achieve them. At each stage, stakeholders relevant to the specific activities and results are identified (Wang et al., 2015).

The generation of research impact for publications has to be traced back to its roots, as it is widely present in the process of research knowledge generating, disseminating, understanding, and transforming into new knowledge or applications. We take the research knowledge life cycle as the basic concept to decompose the stages of the research impact process into five stages: (research) knowledge generation (S1), knowledge publishing (S2), the first wave of knowledge dissemination (S3), knowledge understanding and extensive dissemination (S4), and knowledge transformation and regeneration (S5). The research impact of publications exists throughout the whole process and is clearly reflected in the results of the knowledge transformation and regeneration stage (S5). In Figure 6.1 below, we presented this five-stage schematic diagram and labelled these five stages with S1 to S5. The detailed analyses of each stage were developed in the later sections.

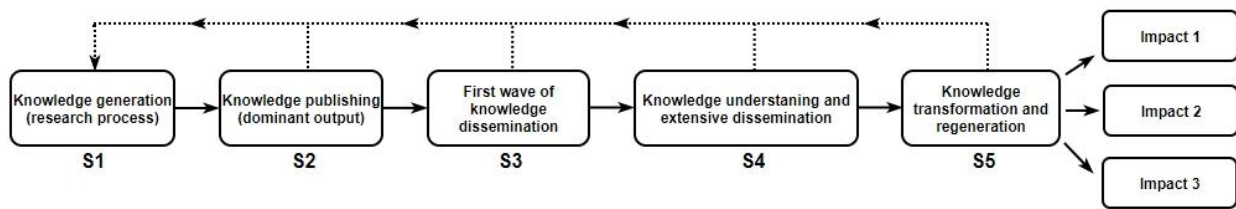


Figure 6.1 The schematic diagram on the stages of generation of research impact

In Figure 6.1, the stages can jump from front to rear; for instance, a research paper has just been published in S2, and an author from the same research team could cite this paper in a new research output in S5. The stage could not jump from rear to front due to not violating the causal relationship on the timeline. It must be emphasised that the research knowledge can be fed back during the dissemination process. Thus, we use dotted lines to describe the flow of scientific research knowledge. Since feedback occurs after the positive dissemination of research impact, only the positive dissemination is calculated in subsequent analyses. Besides, the data on scientific research knowledge feedback is too challenging to obtain.

In the following sections, we use key stakeholders, key actions (and processes), and key results to reflect the cause-effect relationship in the analysis for each stage based on the organisational stakeholder identification methodology (Wang et al., 2015). The study of the impact of research publications needs to consider both academic and non-academic levels, and taking non-academic impact into account is the key to extending the evaluation system to a broader scope. The next step is to analyse the path of generating research impact starting from knowledge generation, construct the internal logic of each process, and compile the corresponding measurement indexes.

6.2.1 Stage 1 - Knowledge generation (S1)

The generation of knowledge goes through two essential stages: the generation of implicit knowledge and the transformation of implicit knowledge into explicit knowledge (Poulin-Dubois & Rakison, 1999). According to Suzuki (2017), explicit knowledge is that can be fully expressed by human beings in a certain symbolic system (the most typical one is language, but it also covers mathematical formulas, various kinds of diagrams, braille, sign language, semaphore and other symbolic forms). Implicit knowledge, as opposed to explicit knowledge, is the knowledge that we understand but cannot be described in a verbal way. This kind of knowledge originates from people's observation and thinking about the world and is developed into explicit knowledge through a series of actionable processes. Let us take the research publications we want to produce as an example to analyse the internal logic schematic of the knowledge generation stage, as shown in Figure 6.2 (the illustration of the entire process of impact generation for publications, see Appendix 1).

Here, we define two processes: the process of generating implicit knowledge and the process of generating explicit knowledge. The common consensus is that the stakeholders for knowledge generation mainly have researchers, students, research teams, research departments (or R&D), research institutes, research funds, countries and so on. Such subjects are classified as key stakeholders in the analysis of this thesis, and we lay stress on the research knowledge that we want to produce and only discuss the main key stakeholders, including the research team, researchers, general scholars, and students.

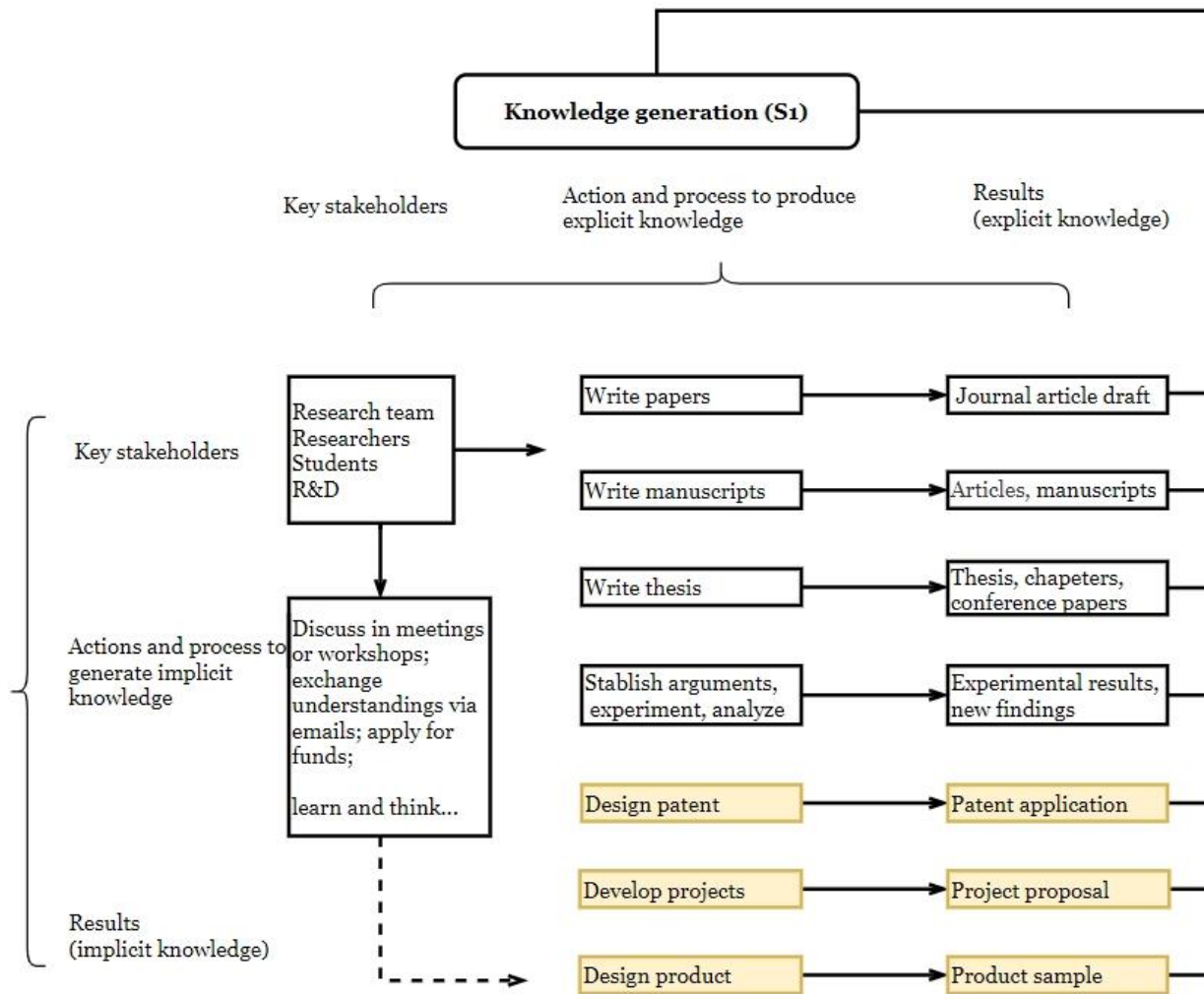


Figure 6.2 The schematic diagram of the process of knowledge generation

In the process of implicit knowledge and explicit knowledge generation, the key stakeholders' observation, reflection, and mutual communication produce implicit knowledge (Berkes, 2009). When they express the implicit knowledge and form the manuscript, it is the explicit process of the knowledge. Researchers make knowledge explicit by establishing arguments and writing papers, creating drafts and preprints of articles, and students summarize research output by writing a thesis, etc. Some experimental results and new findings that produced by key stakeholders can also be formed into posters or slides to discuss with peers at conferences. In addition, the research team may be involved in more work before entering the stage of

knowledge generation, including applying for grants and carrying out primary investigations, and the results of these actions support the production of explicit knowledge.

6.2.2 Stage 2 - Knowledge publishing (S2)

Knowledge publishing refers to the process of transforming tacit and explicit knowledge into public formats such as articles, books, or digital content, facilitating its sharing and accessibility (Singh & Singh, 2016). Knowledge publishing is a key act for the effective dissemination of knowledge (Neavill, 1975). The published knowledge should be presented in a generally recognized format so that the knowledge can be understood and disseminated. Knowledge publishing is a key medium for the creator of knowledge to disseminate his intellectual work to other audiences. Under the Internet era, online forms of publishing are more popular, which expands knowledge distribution channels. Presenting knowledge online will help people acquire new knowledge. Need to note, the published results of knowledge are often used as outcomes and measurement to evaluate the research process, however the research output is not the end of the value of the research but a key indicator for impact.

In the stage of knowledge publishing (S2), the scope of key stakeholders who are exposed to the explicit knowledge has extended from those related to the original research team to those who are exposed to it in the process of knowledge publishing, such as editors and reviewers of journals, conference organizers and more, as shown in Figure 6.3. The increase in the key stakeholders means that the number of people exposed to this research knowledge is increasing, which is also the embodiment of dissemination and the germination of potential influence. When draft papers get approved by the editor's review, they can finally be published in the journal. Conference papers can also be published in conference proceedings after being reviewed and approved by the reviewers in the academic conference committee. If a submitted paper is requested to be revised, the author or the research team will further modify it and then try to publish it afterwards. Alternatively, the authors could post their papers on preprint platforms like arXiv (physical sciences), bioRxiv (biology), and SocArXiv (social sciences) according to disciplines (Polka & Vale, 2015). The preprints uploaded to the server can be published soon after the review. It must be noted that knowledge is irreversible

after being published, and changes to some ideas require rewriting the paper. The published knowledge becomes a physical object that people can read and understand.

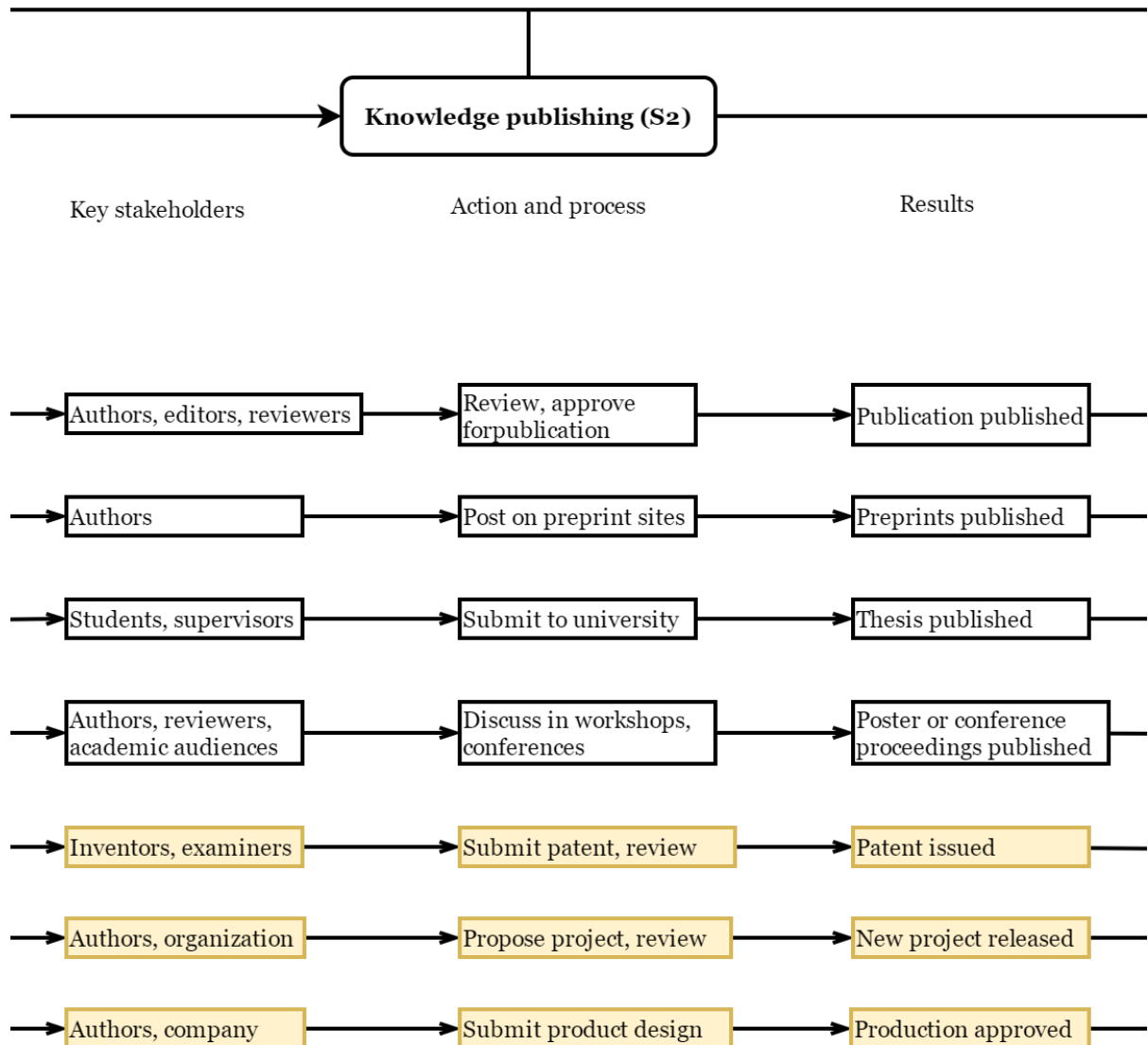


Figure 6.3 The schematic diagram of the process of knowledge publishing

6.2.3 Stage 3 - First wave of knowledge dissemination (S3)

Graham et al. (2006) define the term knowledge dissemination as “the spreading of knowledge or research, such as is done in scientific journals and at scientific conferences”. Salvi (2020) stressed that knowledge dissemination is the transfer of knowledge within and across communication settings, and the expectation of knowledge dissemination is that the

knowledge receiver will use knowledge to change practices or viewpoints or for intellectual growth. It means that the purpose of knowledge dissemination is to serve new knowledge regeneration. Matthews (2012) divided the knowledge dissemination network into physical and intangible dissemination networks at individual and societal levels. Individuals have multiple identities of knowledge creator, knowledge receiver and knowledge disseminator in the path of research impact generation. The order in which people receive knowledge should be clearly distinguished. Generally, people in academia receive knowledge one step ahead of people in society, and those who receive knowledge first are the main force for further dissemination of knowledge. The first effective dissemination of knowledge after its publishing is the key process in which the scope of research impact is manifested. The receivers of knowledge undergo the first wave of increase at this stage, and at the same time, it is the basis for the continued dissemination of knowledge. The schematic diagram of this stage is drawn in Figure 6.4.

The primary receivers of the first wave of knowledge dissemination are academic readers, such as those who download and read the research output in the first place after its publication. These academic readers may include scholars, researchers, and students. There may also be some field experts who are first readers of the research output, such as researchers in the R&D department, technicians, managers, and other highly educated readers. If the scientific outputs are published from preprint platforms, university thesis depository or other channels, academic readers like researchers and students may become the first wave of readers. For conference papers, conference participants and spectators could be the receivers of first-wave dissemination. The result of the first-wave transmission is that everyone understands and gets the implicit knowledge. Some stakeholders may have no interest in the content or understand the knowledge involved and, therefore, have no desire to continue the dissemination. The dissemination process and the scope of the first wave of knowledge dissemination can initially reflect the scope of research content involved and thus influence the size of the scope of scientific research impact.

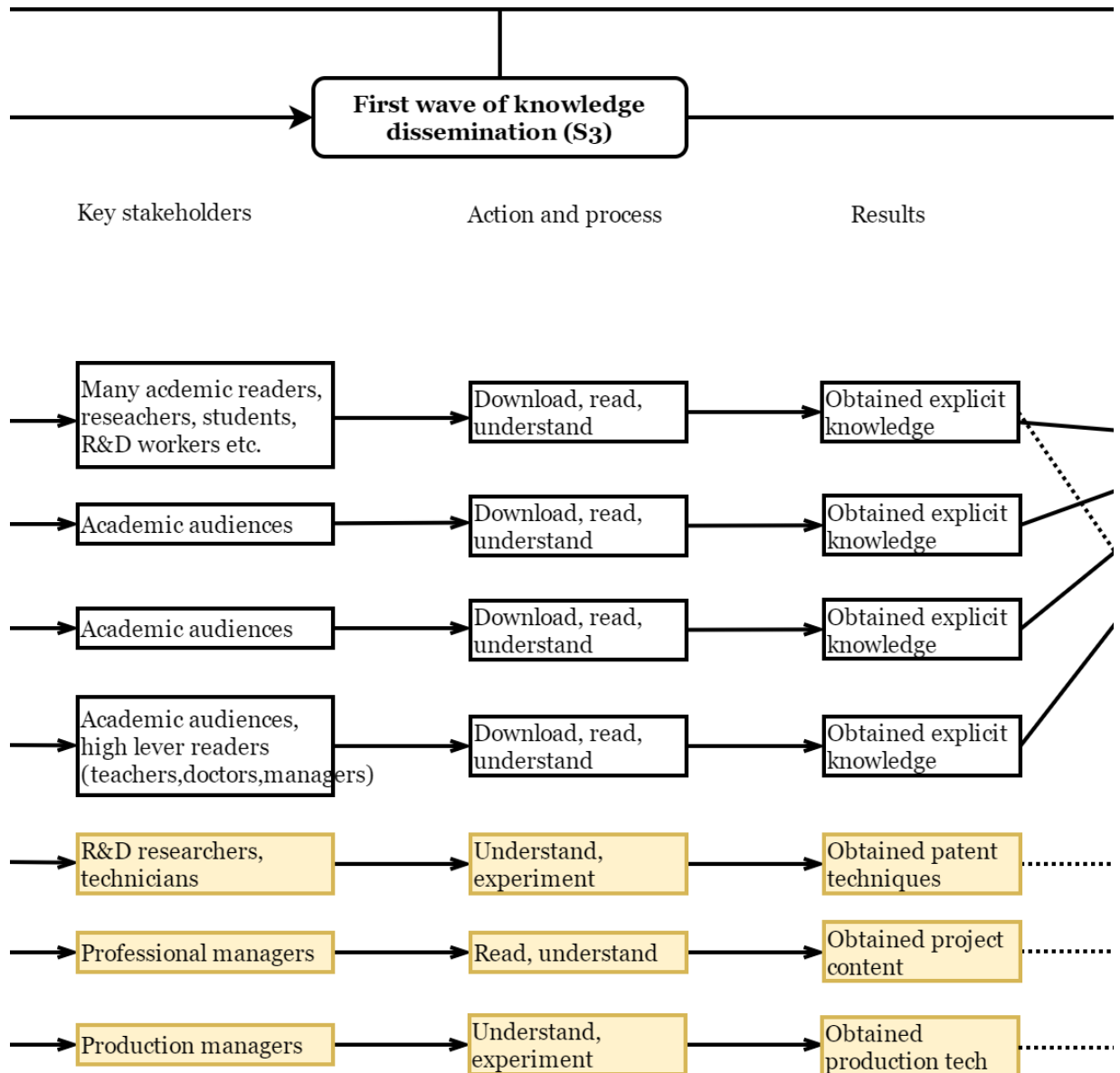


Figure 6.4 The schematic diagram of the first wave of knowledge dissemination

6.2.4 Stage 4 - Knowledge understanding and extensive dissemination (S4)

After knowledge is disseminated through the first wave, it is understood by stakeholders while being continued to be disseminated and forms the stage of knowledge understanding and extensive dissemination, as shown in Figure 6.5. After stakeholders understand the knowledge, they will develop their own opinions, such as agreeing, disagreeing and commenting. Stakeholders may spread their own opinions when further spreading the existing knowledge. For example, when a teacher introduces an article to students, he or she

will talk about why he or she recommends the article. Usually, scholarly professionals would like to comment on the publications when they share these research outputs with peers. Comments or opinions upon the publication show an understanding of the knowledge.

The receiving group in the first wave of knowledge dissemination lays the foundation for knowledge re-dissemination. After understanding the content of the knowledge, the audience in the first wave of dissemination may carry out recommended dissemination or may have already generated new views and directly start the stage of knowledge transformation and regeneration (S5). In the course of understanding and disseminating, the number of key stakeholders involved in this process is increasing. Receivers of dissemination further spread publications after reading and understanding them, and then gradually resulting in extensive dissemination. The scope of knowledge dissemination is increasingly expanded, not only within the academic community but also in other aspects of society in the extensive dissemination stage. Besides, some stakeholders will stop the dissemination at this stage, and some will jump into the knowledge transformation and regeneration stage.

Key stakeholders in S4 include academic readers and various societal readers. Academic readers play an important role in the prophase of extensive dissemination of scientific knowledge because of their own expertise. Researchers may directly pass the research output to other researchers, colleagues, or grant applicants to play a role in generating new research projects. People close to the academic community, such as teachers, students, general scholars, domain experts, and service providers of the academic community, are also early receivers in the dissemination process.

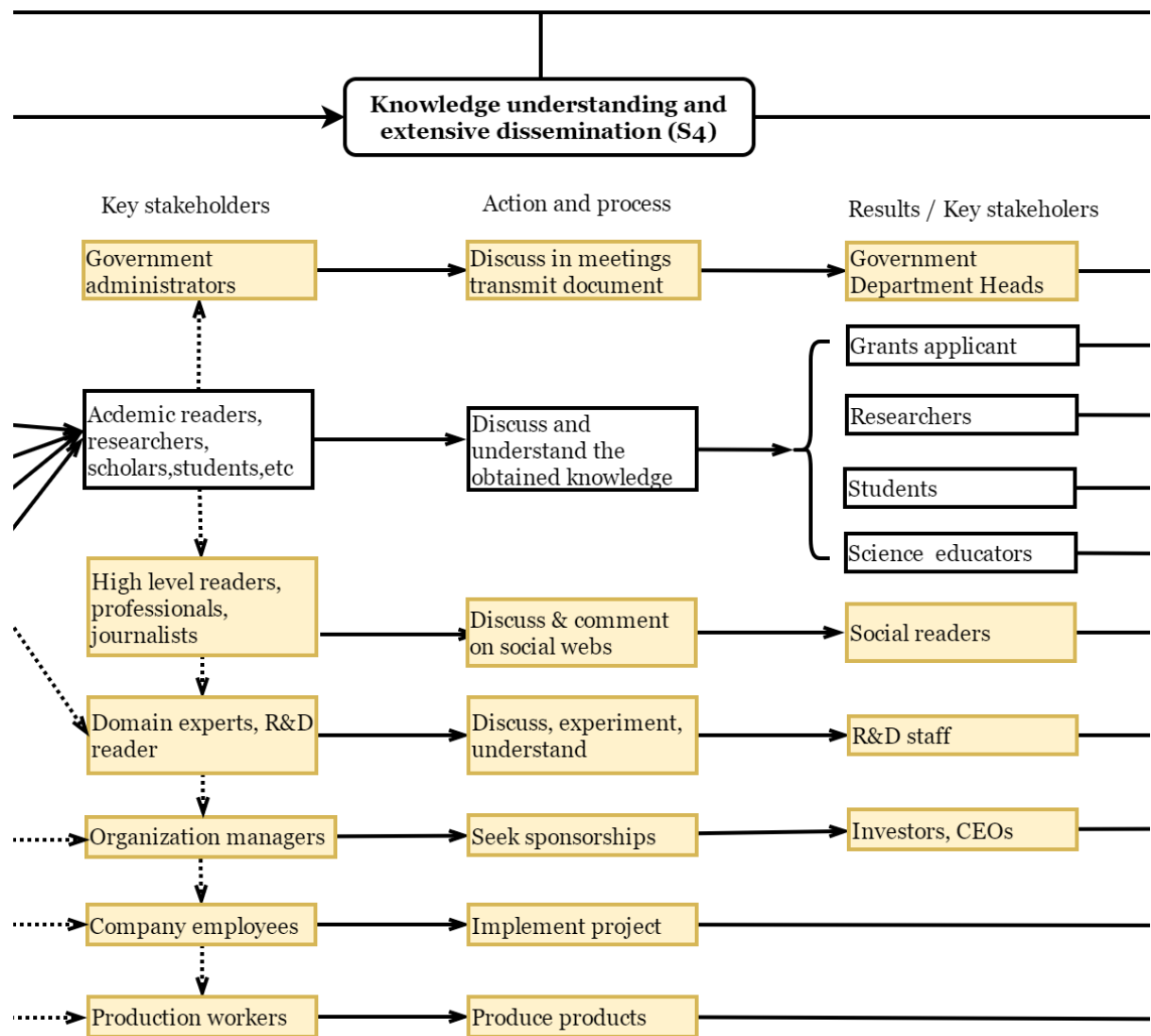


Figure 6.5 The schematic diagram of process of knowledge understanding and extensive dissemination

With an understanding of the knowledge in research output, more readers continue to spread the knowledge. Domain experts may transmit knowledge to specific work areas, such as R&D departments and laboratories in companies, to impact the development of new technologies, patent applications, and more. Company personnel may, in turn, pass on knowledge to investors, CEOs, project managers, and product managers, thus influencing the company's operations. A small number of societal readers who are close to the research team, such as friends and family members of the team, may become the early receivers of knowledge

dissemination. Most social readers become receivers of knowledge dissemination after academic readers are fully convinced of the validity of scientific knowledge. Research output related to specific industries will be further disseminated to the related fields, such as medical research content will be spread to doctors, management research content will be transmitted to management service personnel, and policy research outcomes will be relayed to policymakers and government personnel.

As the impact of the dissemination continues to grow, more social readers will be generated. For example, journalists are also important catalysts for disseminating research knowledge to the community. Bear in mind that on the path of knowledge dissemination involving patents, projects, and products (see Figure 6.2 to 6.5), the scope of further dissemination will not be significantly expanded due to the narrower scope of specific patents. Nevertheless, it is easy to enter the stage of knowledge transformation and regeneration (S5) faster in the form of the application of the knowledge. The results of broader impact, such as societal impact, take a longer time to brew to reflect the significant influence of research output at some point, so the stage of knowledge understanding and extensive dissemination (S4) is an important part of the overall impact of research.

6.2.5 Stage 5 - Knowledge transformation and regeneration (S5)

When knowledge is continuously disseminated, all stakeholders, as illustrated in Figure 6.6, generate some own ideas as implicit knowledge and then transform the implicit knowledge into new explicit knowledge or applications. For instance, researchers quote obtained scientific knowledge in new papers or books to achieve knowledge reproduction. They may cite the obtained knowledge in grant applications for applying funds for new research. Other scholars, such as teachers, transform the acquired knowledge into teaching materials, and students may use acquired knowledge to complete their thesis or coursework. Some science services peoples, such as librarians, can optimize academic materials after obtaining statistical knowledge from the latest research outcomes, and they can manage academic resources by understanding new methods of management or data applications. Domain experts and R&D staff can use the knowledge gained to start new technology research, and new outputs may

include patents and papers. The company managers can carry out new projects and launch new products by understanding new technologies, thus making the company more profitable. Company investors and CEOs can set up new strategic goals and advance toward new development directions through the understanding of obtained scientific knowledge, thus influencing the company's future development.

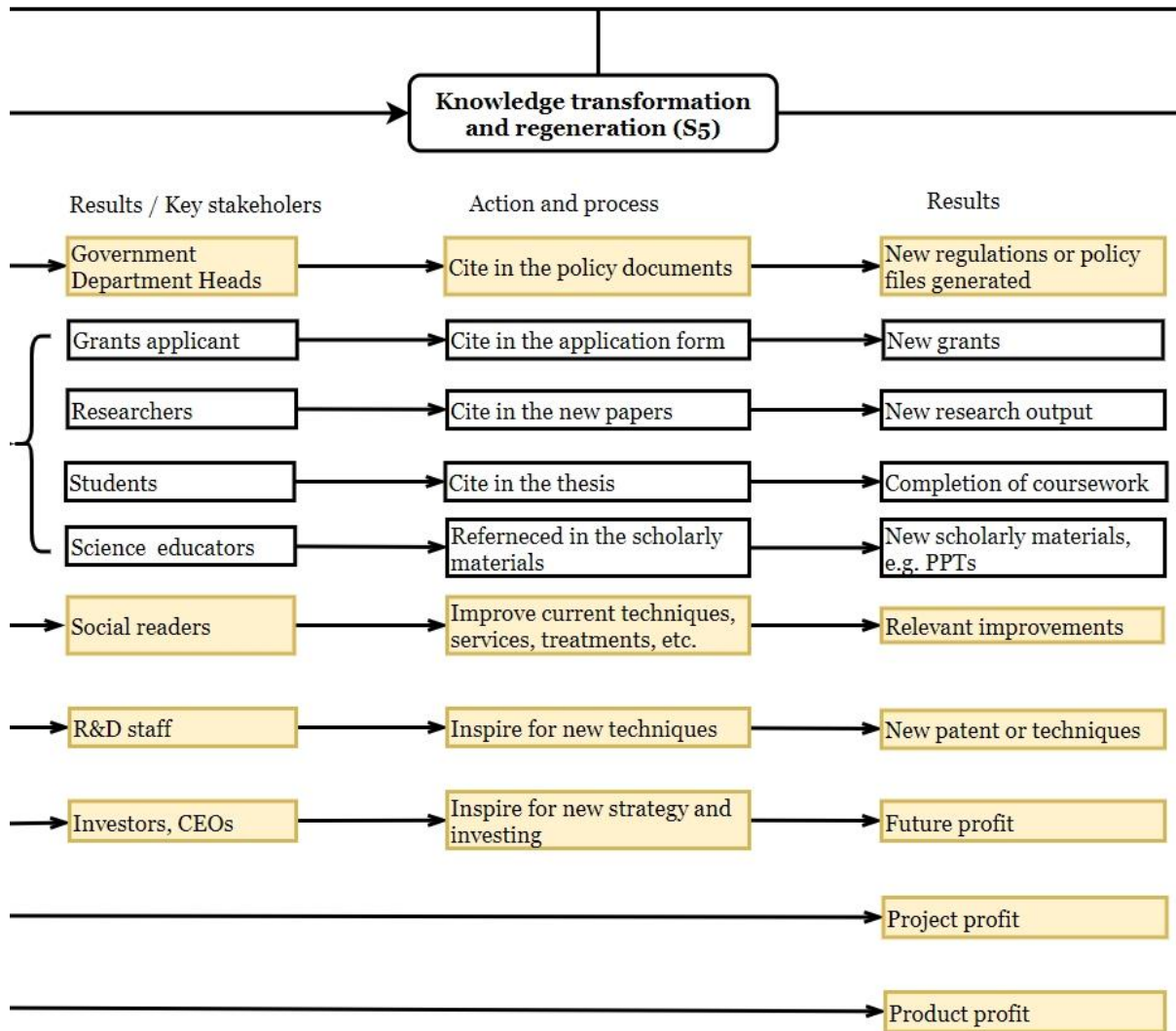


Figure 6.6 The schematic diagram of the process of knowledge transformation and regeneration

Regarding social readership, the transformation of knowledge can vary from industry to industry. For example, doctors will improve the quality or method of treatment based on the latest research output; government workers can modify the policy documents or develop new

ones based on new issues presented in scientific research. Other social service workers may improve the quality of public services based on new scientific research techniques. General scholars, or high-level readers, can transform obscure research knowledge into popular science articles and, therefore, may be conducive to disseminating knowledge to all sectors of society. When research knowledge is widely disseminated at the societal level, more stakeholders will emerge to transform and regenerate the knowledge in different fields.

Key results in the stage of knowledge transformation and regeneration (S5) are evident proofs to reflect the research impact, as shown in Figure 6.6. The number of publication citations is one key result that can best reflect the original knowledge transformation and regeneration in the academic context. Such citations could include journal, book, and student paper citations. Scientific knowledge transferred to the company and the industrial sector can inspire patent research in the company's R&D department, and then patent citations are an obvious result. In addition, the profit results that scientific research knowledge brings to the company by inspiring new research, projects and products may not be easy to distinguish from the daily profit results. However, if the product uses a patent and the patent has cited previous research output, then it can still indirectly reflect the economic impact brought by the original research output. Moreover, key results like citations in policy documents and mentions in news, reports, or public blogs embody the impact of research knowledge on society. The citations in Wikipedia also reflect the public's understanding of knowledge and the social impact generated.

The ultimate impact of research output is expressed through the realization of its application and value, which is a comprehensive impact on all aspect as a whole. The key results in the S5 cross-reflected a comprehensive impact of research output on various aspects of society. We summarised the following aspects of influence according to our process of generation of impact based on cause-effect chains: generation of new knowledge, educational improvement, economic wealth, innovation and entrepreneurship, sustainable development, public service improvement, health and well-being, government governance and social

awareness. To generalize the above aspects can better understand and analyse the composition of the comprehensive impact of research out.

To sum up, we have distinguished and defined the key actions and results in the process of generation of research impact, which exists in the five stages from research knowledge generation to knowledge transformation and regeneration. There is a causal relationship in terms of the timeline between the five sub-processes. However, it should be noted that knowledge can also enter the transformation and regeneration stage (S5) directly without publishing (S2) or external dissemination (S3 & S4). For example, a researcher in the same research team could skip the current research but start an extended topic directly. Alternatively, research knowledge could be transformed and regenerated at the first wave of dissemination (S3). The stages can jump from front to rear, but not from rear to front, and cannot violate the causal relationship on the timeline. In summary of the above analysis and sub-processes, the whole process of research publications, from knowledge generation to posing impact, is presented below in Figure 6.7 (a large print version is available in Appendix 1).

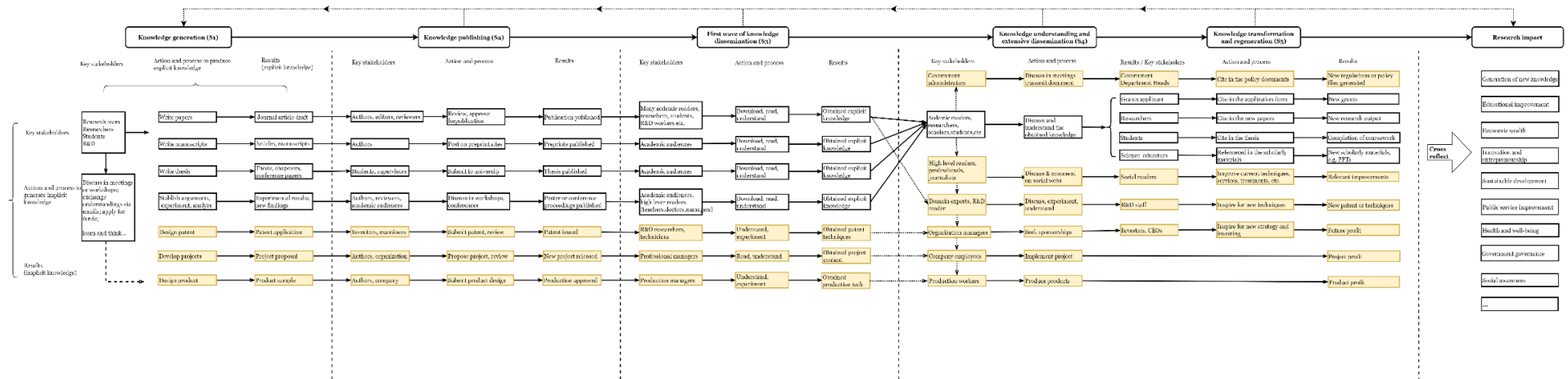


Figure 6.7 The comprehensive schematic diagram of the generation of research impact for publications

6.3 The evidence data of measurement in impact evaluation

Through the above theoretical discussion, the quantitative evaluation of the impact of publications based on the evidence chain requires abundant evidence (data) on actions and results. The evidence data should follow logical relationships on the cause-effect chains of the impact generation process. The collection of sufficient evidence relies on the availability of accessible data. Today, the scholarly community uses online platforms for research work and communications. The traces left online recorded their academic activities and interactions, which form many unique academic behaviour data (see Chapter 4). The increased diversity of data sources, such as altmetrics, provides opportunities to trace such evidence data that reflect the impact related to the research outputs. Therefore, extracting traceable data is important in constructing the evaluation measurements.

6.3.1 Measurement of impact generation process

We aim to summarise the measurement data to reflect the actions, results and impact in each stage of the comprehensive paths of research impact generation, as developed in Section 6.2. All stages of the impact generation process could contain rich data, but not all the data are collectable. Some actions, such as writing a thesis, are in the knowledge generation stage, but it is challenging to form visible data. Our mission is to extract and specify the traceable data that can potentially be collected. Table 6.1 below summarises some example indicators derived from the comprehensive process of impact generation for research publications from stage 1 to stage 5.

Table 6.1 Measurements of the impact generation process

	Key stakeholders	Key action	Key results	Measurement data	Indicators (example)
S1	Authors	Apply grant for research (output)	Grant application	Grant data	The number of supporting grants; the amount of supporting grants;
S2	Authors	Submit research output for publishing	Research output published	Publication data	Published journal papers, Preprints, conference proceedings, theses, books, and patents;

S3	Academic audiences	Read the research output	Research output is viewed	View data	The number of views of the research output across displayed platforms;
	Academic audiences	Download the research output	Research output is downloaded	Download data	The download number of the research output;
S4	Academic readers	Save the research output to reference management software	Research output is added to reference management software	Reader data	The number of times the research output is saved to Mendeley, CiteULike or other reference management software;
	Various readers	Mention the research output in the media or various online platforms.	Research output is mentioned in the media or various online platforms.	Mention data	The number of times the research output is mentioned in the media, such as news, newspapers, or websites such as Wikipedia, blogs, forums, and more;
	Various readers	Discuss or comment on the research output on social media	Research output is discussed or commented on social media platforms	Discussion data	The number of times the research output is liked, discussed, or commented on the social media platforms such as X, Facebook, Reddit and more;
S5	Various readers	Cite the research output in new literature	Generation of new knowledge	Publication citation	The number of citations of the research output in scholarly literature, such as journal papers, books, essays, and other scientific documents.
	Various readers	Cite the research output in new patents	Generation of new patents	Patent citation	The number of citations of the research output in new patents.
	Various readers	Cite the research output in new clinical trials	Generation of new clinical trials	Clinical trial citation	The number of citations of the research output in clinical trials.
	Various readers	Cite the research output in new policy documents	Generation of new policy documents	Policy document citation	The number of citations of the research output in policy documents.

Various readers	Use research output in business, design, project, product, various applications or techniques	Generation of new forms of output	Other referring data	The number of times the research output is used in wider applications, such as business meetings, design, project, product, new treatments, techniques, and more.
-----------------	---	-----------------------------------	----------------------	---

First, we need to clarify what measurement data is. Sebastian-Coleman (2012) states that measurement data are “facts and statistics collected together for reference or analysis.” Measurement data could include attribute data to count the number of facts, such as 160 downloads. It is clarified that the purpose of measurement data is to increase our understanding of an object or situation by measuring that object or situation (Sebastian-Coleman, 2012). At each stage of the research impact generation process, measurement data mainly reflect the number of key actions taken by key stakeholders or the number of the key results from their actions.

In the stage of knowledge generation (S1), most data that reflects key actions or key results are invisible. A specific data that needs to be discussed is the grant data. Here, the grant data means the supporting grant to aid the production of the research output. Grants were applied by authors or the research team to support producing the research output. When a grant has been granted from the fund body, it means the proposed research has promising research value. Therefore, grant data is a result data that exists in the stage of knowledge generation (S1), with the key stakeholders being the authors or the research team. Additionally, the grant may indicate an extra value recognised by investors so that the grant means that a particular impact has been generated. At present, grant data becomes publicly available once the research output is published. This is because publications acknowledge the specific funds used, allowing us to count the supporting grant data. While gathering grant data directly from the funding body may be challenging due to data privacy, we can still access it from the published output (Bordons et al., 2024).

In the stage of knowledge publishing (S2), publication data could include published journal papers, preprints, conference proceedings, theses, books, and patents. These published research outputs are the results of knowledge publishing. The stage of the first wave of knowledge dissemination (S3) mainly contains measurement data on viewing and downloading the research output. Example indicators could be the download number of the research output, number of abstract views, full-text views, clicks, etc. Some platforms count views and downloads of papers by themselves. For example, ScienceDirect has publications with full-text views and download statistics (Yom, 2023). The challenge is that there is rarely a platform that can count all view or download times for all published publications.

The stage of knowledge understanding and extensive dissemination (S4) involves numerous indicators. As the research output's readership increases, the knowledge is spread and understood on various platforms. These include reference management software like Mendeley and CiteULike, social media networks like X(Twitter), Facebook, and Reddit, online websites like blogs, Wikipedia, Google, and Quora, various blogs, peer review forums like PubPeer and Publons, and more (Kassab et al., 2020; Liu, Jones & Wang, 2023). The above examples are not exhaustive but only list specific common indicators. As data diversification develops, more data could become available to reflect the dissemination and understanding of knowledge on different platforms. At this stage, the more diverse the data for the research output, the more extensive the knowledge is disseminated. It should be noted that some data sources only emerged in recent years, so their coverage and statistical methods need to be examined. The quality of the data should be verified before using it.

In the knowledge transformation and regeneration stage (S5), a new form of output containing the understood knowledge was conducted. Key stakeholders in this stage transform the knowledge into new forms, and the types could be various—for example, new research papers, essays, ideas, patents, products, and more. The data that can be used to trace the source of new knowledge is reference data, which records the source knowledge. Therefore, the traceable measurement data in this stage may include publication citations,

patent citations, clinical trial citations, and policy document citations. Different stakeholders can also transform knowledge into ideas in design, projects, products, or other social applications or techniques, but such data may be difficult to obtain unless the source knowledge is referenced.

6.3.2 Measurement of different dimensions of impact

The impact of research outputs should be extensive and comprehensive. We have summarized the different dimensions of impacts commonly used in Figure 6.7, but this list is not exhaustive. Besides, many aspects of the impact are also cross-reflected. It also needs to be noted that impacts could be directly or indirectly influenced. We may find data to support the direct impact related to the research output, but those indirect impacts would be difficult to distinguish. Table 6.2 illustrates a few examples of indicators for different dimensions of impact. However, whether these data can be formulated and collected is another issue that depends on data development in this industry.

Table 6.2 Examples of impact indicators

Dimensions	Indicators (example)
Generation of new knowledge	Amount of new knowledge generated, which includes academic output and non-academic output;
Educational improvement	Amount of applications in teaching PPTs, textbooks, and other school documents;
Economic wealth	Profit from new patents, projects, or products;
Innovation and entrepreneurship	Number of new ideas, start-ups, strategy change;
Sustainable development	The number of transformational enterprises; the number of environmental-friendly products and policies;
Public service improvement	Development of social public service provisions;
Health and well-being	Number of new medicines, treatments or services; Number of improved standards in the level of medical care;
Government governance	Number of policies formulated;
Social awareness	Amount of Knowledge diffused and increased level of civic awareness;

6.4 Summary

Along with the development of the Internet era, when a large amount of data becomes available, evaluation methods that are more adapted to the times will be developed. The scientometric evaluation method based on the number of citations is increasingly unable to evaluate the comprehensive impact of research output in an all-round way. When considering the impact of scientific research, the focus should be both the academic community and other aspects of society in a broader way. The comprehensive impact is the next focus of research evaluation, as many countries consider the evaluation of impact to be an essential basis for allocating science and technology resources. To develop the evidence-chain-based methodology on comprehensive research impact for publications, we had the following achievements:

- In this chapter, we proposed a research logic based on the impact generation cause-effect chains, which puts together the academic, social, and economic impacts of the research output as a study of the comprehensive impact posed by scientific research. We identified five logical stages of generating impact for research publications: knowledge generation (S1), knowledge publishing (S2), the first wave of knowledge dissemination (S3), knowledge understanding and extensive dissemination (S4), and knowledge transformation and regeneration into new outputs (S5). The original research output's extensive impact arises across the different paths of the whole process.
- We also clearly defined the key stakeholders, key actions, and key results of each stage and then extracted quantitative indicators to help aggregate the comprehensive impact of the scientific research. The more indicators, the better. However, we need to note that the indicators were involved in various types of data and data sources, and not all kinds of data are collectable or of good quality for use.
- We have illustrated and summarized the path of research impact generation along the logic of research output from knowledge generation to impact generation. Our cause-

effect paths of the impact provide a clear and logical perspective on impact, which breaks through the original limitation of only considering the output's results data in the impact evaluation. Using the data collected at each stage to evaluate the impact could solve the attribution and reliability problem and make evidence data causally connected.

- Our evidence-chain-based impact evaluation method was further developed with a conceptual model of a comprehensive impact generation process. Using multi-source data as a quantitative method for research evaluation is a new way of thinking. This perspective provided new insight into impact evaluation, with data objectivity and logical accuracy.

The limitation of this chapter's study is that our impact-generation process model only summarised the most common conditions; more impact-generating paths exist but have yet to be demonstrated. Besides, we cannot collect all the path information due to the current data tracing capability. The expected practical impact calculation could only be based on the available measurable data. Furthermore, the final quantitative evaluation results rely on a large amount of data screening and processing, and no authoritative methods are available to assign weights and quantify specific indicators. Using big data obviously provides a better way of thinking for research evaluation. However, the cost of data collection and establishing a database with all chain data are still hurdles to a more comprehensive impact evaluation.

In the future, as data collection and processing technology develops, these problems are expected to be solved at the technical level, and the method of evaluating the comprehensive impact of scientific research based on the evidence chain will be further perfected. In the next chapter, we will conduct a real case study using the currently available data to evaluate the impact of research publications.

Chapter 7 - A case study of comprehensive impact evaluation for research publications

7.1 Introduction

Chapter 6 delved into the intricate cause-effect chains of scientific research knowledge from generation to posing impact. Using the evidence data on the cause-effect chains, we can construct impact evaluation models and aggregate the assessment results. This chapter will collect journal papers from five major category groups to evaluate their comprehensive impact based on evidence data collected in the impact generation process. We developed primary evaluation models in Chapter 5, and this chapter will further develop comprehensive impact evaluation models utilising the structural equation modelling approach. The structural models for different disciplines will be built for impact evaluation on different category groups. At the end of the chapter, we will illustrate some ranking results of the impact of publications using our evidence-chain-based evaluation methods.

7.2 Database screening and data quality verification

Our evidence-chain-based quantitative evaluation method requires substantial evidence data extracted from the impact generation process. The data shall be of good quality and accessibility. Therefore, ensuring good data quality is the precondition for using these data to construct impact evaluation models. The data quality largely depends on its database's coverage and counting methods. Choosing the appropriate database for data collection needs to be intensively investigated and discussed. Moreover, obtaining all evidence data across the impact generation process is extremely challenging. We pursue as much as we possibly can to collect the data within the process of impact generating, yet this depends on the availability and accessibility of the data. Section 6.3 has addressed a range of measurement data and potential indicators. Below, we will discuss the corresponding data and database across different stages of the impact generation process.

7.2.1 Measurement data in S1

In the stage of knowledge generation (S1), most measurement data are invisible except grant data, as discussed in Section 6.3.1. Potential indicators for grants linked to the research output may include the number of supporting grants or the amount of supporting grants for the research output. To select the best database of grant data of publications for our case study, in the following sections, we will explore the most used worldwide grant databases, their data forms, data counting mechanism, coverage of funding sources, and use sample data to conduct the comparative analysis when necessary.

7.2.1.1 Research grant and database

Research grant varies by the fund's types and disciplines, and different countries have different funding institutes. For example, the United States has the National Institutes of Health (NIH) and National Science Foundation (NSF) for national fund application and management (Berg & Ashurst, 2019). UK has Cancer Research, Innovate UK, Economic and Social Research Council (ESRC), Medical Research Council (MRC), Arts and Humanities Research Council (AHRC), Natural Environment Research Council (NERC), Biotechnology and Biological Sciences Research Council (BBSRC) and many other fund bodies (Cooksey, 2006). Other famous national research funds include the National Natural Science Foundation of China, the Swiss National Science Foundation (SNSF), etc.

Each fund body has its own funding disciplines and a range of fund supporting methods. For example, the National Institutes of Health (NIH) is a leading global provider of research funding in life science (Berg & Ashurst, 2019). NIH's database, Research Portfolio Online Reporting Tools (RePORTER), offers a user-friendly interface for accessing NIH-funded research projects, publications, and patents. It enables users to search funded projects by keywords, investigator names, or institutions. Similarly, the National Science Foundation Award Search provides a seamless experience for exploring NSF-funded research projects across various disciplines, including science, engineering, and social sciences. Users can search for awards by keywords, institutions, or program areas (Award Search, 2024). Another user-centric grant database in the US is the grants.gov. It serves as a central portal for

searching and applying for federal grants in the United States, aggregating grant opportunities from over 1,000 federal grant-making agencies (Grants.gov, 2024). The Funding & Tenders Portal, as the central gateway for accessing funding opportunities and tender notices offered by the European Union, covers a wide range of research and innovation programs, including Horizon Europe (EU Funding & Tenders Online Manual, 2021). PubMed has the funding information whilst mainly serving life sciences and biomedical topics (Bordons et al., 2024). OpenAire, which harvests open science systems to create datasets of publications from EU-funded projects, is the largest funding database in the EU. The limitation is that OpenAire only records European Commission-supported research projects (Manghi et al., 2019).

It is important to note that the aforementioned grant databases, while valuable resources, have their limitations. They primarily focus on scientific research projects they have invested in and are bound by certain regional boundaries, making them less comprehensive for our case study. Moreover, the grant data we aim to collect should be linked to the resulting published research papers, a connection that is often missing (Boyack, 2009). This understanding of the databases' limitations is crucial for our research approach.

7.2.1.2 Bibliographic database with grant data

Through literature review, comprehensive funding data involving multidisciplinary and worldwide publications is available in commercial bibliographic databases, including the Web of Science (WoS), Scopus, and Dimensions (Bordons et al., 2024; Liu, Tang and Hu, 2020; Mugabushaka, 2020). The WoS started to provide funding information in 2008, Scopus in 2017, and Dimensions in 2018 (Mugabushaka, 2020). As clarified by Bordons et al. (2024), the quality of the funding data collected by a database largely depends on the sources used and the database's ability to extract original text information correctly. Therefore, carefully studying the grant data in these databases is essential to guide our data collection. The following paragraphs will compare each database's funding data forms, data collecting methods, and coverage of funding sources, then use sample data to conduct a comparative analysis.

As learnt from the WoS database and its official websites, WoS provides funding organizations, the preferred funding names, the grant's titles and numbers, and funding acknowledgements for each publication when its fund information is available. The funding data provided by WoS is mainly collected from publications themselves (acknowledgements) and supplemented with grant repositories such as Medline or Researchfish (Clarivate Support, 2023). Starting from November 2020, WoS integrated funding sources into a range of grant repositories with more than 1200 funding agencies in its core collection database (Clarivate Support, 2023).

Regarding the data collection methods in Scopus, Scopus uses natural language processing and machine-learning techniques to identify grant information in funding acknowledgment sections of the documents (Baas. et al, 2020; Bordons et al., 2024). The latest Content Coverage Guide of Scopus updated March 2023 on its official websites exposed that funding information is recognised if the funding body is included in the database of Open Funder Registry (OFR) sponsored by Crossref²³. Scopus presents the funding details of publications include Funding Sponsor (i.e., the National Natural Science Foundation of China), Funding Number (i.e., 31161130357), Funding Acronym (i.e., NSFC) and funding text (full text of acknowledgements).

Dimension provides comprehensive supporting grant information, including the titles of grants, grant numbers, funding amount, funding periods, original grant description, grant abstract and keywords, original grant link, program name, investigators, resulting publications, resulting patents, resulting clinical trials, and more once the relevant information is available²⁴ (Dimensions, 2022). Dimensions link publications and research life cycle data more entirely than any similar bibliometric database to date, including supporting grant information (Herzog, Hook & Konkiel, 2020). According to Hook, Porter, and Herzog (2018), Dimensions adopt a hybrid approach to linking publications and relevant funding data. First, a machine-learning technique is used to capture existing funding data, such as PubMed, which

²³ <https://www.elsevier.com/products/scopus/content>

²⁴ <https://dimensions.freshdesk.com/support/solutions/articles/23000018851-what-information-does-dimensions-have-for-each-funded-grant->

includes grant information associated with the resulting papers. Secondly, grant numbers and codes contributed by ÜberResearch (a data science company focused on establishing the global grants database and initiating the dimensions platform²⁵) are applied to search through funding acknowledgement context in the paper as permitted by publisher relationships. Additionally, when grant identifiers are not explicit, Dimensions identifies funding attributions and matches them with the existing funder to aid identifications. Therefore, through machine learning to analyse the datasets and entity-assisted identification, Dimensions identified over 32 million links between publications and funding bodies (Hook, Porter & Herzog, 2018).

From the above literature, we can see that WoS collects funding information through publications' funding acknowledgment texts, and identifies funding agencies through integrated funding sources. Since Scopus' funding data sources heavily rely on the Crossref funder database, Scopus' funding data quality shall be limited by coverage. A comparative case study on the accuracy of the funding information in Scopus carried out by Liu (2020) states that 68% of the funded papers have inaccurate funding data. Liu (2020) also compared the funding data extracted from WoS, found that 8% of papers are inaccurate, and remarks that the funding information collected by WoS is better than that of Scopus. WoS extracting funding details from publications, where they often coexist with other forms of acknowledgements, adds complexity to the extraction process, potentially resulting in some loss of information through automated identification (Grassano et al., 2017). Prior investigations in the WoS have suggested that the fund information loss could affect approximately 12% of research papers (Bordons et al., 2024). Besides, WoS's data extraction method, which relies solely on funding acknowledgement (FA) text, may have issues with insufficient information or mistakes made by authors.

7.2.1.3 Sample test

A small data sample is used to investigate the quality of grant data between WoS, Scopus, and Dimensions in order to compare these databases' grant data in depth. Five hundred papers with funding information were randomly selected and matched the corresponding

²⁵ <https://www.digital-science.com/blog/2016/10/uberresearch-announces-dimensions-publishers/>

grant data from WoS, Scopus, and Dimensions via DOI. The descriptive statistics are presented in Table 7.1, which shows that Dimensions has the highest maximum and mean number, meaning its funding data has the highest count on average. The data correlations among the three databases are analysed, as shown in Table 7.2, and the results show that all three databases' grant data are significantly correlated. According to the data sample comparison results, the grant data in Dimensions has the best quality. Therefore, we will use Dimensions for grant data collection.

Table 7.1 Descriptive statistics of grant data sample

	N	Minimum	Maximum	Mean	Std. Deviation
WoS	500	0	61	3.80	9.689
Scopus	500	0	68	4.55	9.544
Dimensions	500	0	93	6.45	13.180

Table 7.2 Correlations of grant data sample

	WoS	Scopus	Dimensions
WoS	1	.973**	.943**
Scopus	.973**	1	.891**
Dimensions	.943**	.891**	1

** . Correlation is significant at the 0.01 level (2-tailed).

7.2.2 Measurement data in S2

In the process of generating and disseminating research impact, the knowledge publishing stage (S2) could include all kinds of published content (see Section 6.2.2). Since published journal papers are evaluation subjects in this case study, we need to find other indicators in S2 associated with the published journal papers. Below are detailed research outcomes.

7.2.2.1 Conference proceedings

We explored publication data on conference proceedings and found the main issue is its linkage to published journal articles. If a full conference paper has already been published in the conference proceedings, republishing it might only be possible if it is an updated version of the original paper. It is suggested that at least 30% of new content should be added to the journal version, and the journal version must cite and discuss the previous conference paper

and summarize the new material (Offutt, 2016). Additionally, the paper's title should be changed to avoid having two published papers with the same title in the public domain, which can be confusing and unethical. Conference papers and journal articles are two different types of published documents, and therefore, linkage data is rare. Although a link between a conference proceeding and a journal article can be discovered through overlaps of content and authors, it is challenging to match without errors due to changes in titles, abstract and sometimes authors. More importantly, such available databases are deficient.

7.2.2.2 Preprints

Publication data on preprints can deliver the demand in our case study because preprints can have the same title and same content except for corrections after editors' review.

Preprint platforms provide researchers with the opportunity to quickly share their research findings on an open-access public server. There are numerous preprint repositories available online, each with its own unique focus (Directory of Open Access Preprint Repositories, 2024). Some are subject-specific, while others cover multiple disciplines. It is important for users to explore various platforms to find the most recent research on a specific topic. Besides, preprints can be found through non-preprint resources such as PubMed and Google Scholar (Kirkham et al., 2020). Table 7.3 illustrated ten examples of preprint repositories that are commonly used (Clarivate, 2023; Elsevier, 2024; Kirkham et al., 2020; Lin, 2018 May 31) and summarised their main characteristics and the sizes of the repositories as of 2023 (Directory of Open Access Preprint Repositories, 2024).

Through database investigations, it is verified that arXiv and bioRxiv provide "related DOI" to record the links to published publications afterwards. In arXiv's submitting system, the authors could add the information of the final published article, including DOI and journal references, to their preprint papers (arXiv, 2024). Therefore, these two preprint servers have information for linked published papers with preprints. However, we aim to use one database that covers preprint repositories as much as possible. The commonly used bibliographic databases that contain preprint publications include WoS, Scopus, and Dimensions.

Accordingly, the following paragraphs compare three databases with their coverage and data count records. Examples of preprint repositories.

Table 7.3 Examples of leading preprint repositories

No.	Name	Discipline	Description	Size	Created
1	arXiv	Multidisciplinary	Main focus on mathematics, physics, astronomy, computer science, quantitative biology, statistics, quantitative finance and economics	2,250,000	1991
2	bioRxiv	Biology	Open access preprint repository for the biological sciences	194,000	2013
3	medRxiv	Medicine	Open access preprint archive for medicine	42,000	2019
4	ChemRxiv	Chemistry	Open access preprint archive for chemistry	13,865	2017
5	PeerJ PrePrints	Biology, medicine	Initially developed as preprint server for PeerJ submissions. Closed Sep 30th, 2019	5,086	2013
6	Preprints.org	Multidisciplinary	Preprint server for interdisciplinary research governed by MDPI	36,623	2016
7	Social Science Research Network (SSRN)	Multidisciplinary	Aggregates over 30 preprint servers (Preprints with The Lancet, Cell Sneak Peek, etc.) and covers over 55 disciplines. Initially funded by SSRN in 1994, bought in 2016 by Elsevier.	950,733	1994
8	Nature Precedings	Multidisciplinary	Preprint repository of scholarly work in the fields of biomedical sciences, chemistry, and earth sciences	3,423	2006
9	Research Square	Multidisciplinary	Preprint server plus editing (commercial)	223,000	2018
10	TechRxiv	Engineering	Preprints in Technology Research	1,043	2020

The Preprint Citation Index in Web of Science (WoS) is a multidisciplinary collection of preprints from a wide range of leading preprint repositories, including arXiv, bioRxiv, medRxiv, chemRxiv, Preprints.org, and over 20 repositories to be added in the future (Clarivate, 2023). The five repositories covered by WoS have been listed in Table 7.3. These repositories contribute over 2.5 million preprint records, and each server has a varied start date for data collection.

One of our main tasks is to discover the linkage data between preprint documents and related peer-reviewed journal articles. It is observed that the Preprint Citation Index in WoS is a parallel database to all the other WoS collections. We carefully checked the preprint index pages but failed to find direct links between preprints and resulting journal articles. WoS adds authors of preprints to its author records in the WoS Core Collection, together with ORCID Number, so that we can find both preprints and peer-reviewed publications through author pages. However, such linkage makes it difficult for us to prepare data downloads. On the other hand, Preprint Citation Index also records citation counts from other preprints and all documents collected by WoS. Consequently, WoS may focus on document collection (preprints and journal articles) rather than connections between them.

Scopus has 1.8 million preprint records as of June 2023 from seven preprint servers, which include arXiv, bioRxiv, ChemRxiv, medRxiv, SSRN, TechRxiv, Research Square (McCullough, 2023, May 10). Scopus also supports viewing peer-reviewed articles related to the preprints but is subject to the documents involved in the Scopus database coverage. We must be aware that preprints in Scopus are collected from 2017 onwards, and there is no data for earlier years (Elsevier, 2024). This will limit the data coverage of preprints in Scopus and not favour our case study's target publication set, as our targeting publications should have been published before 2015 to observe their broader impact.

Dimensions indexes preprints from arXiv, bioRxiv, ChemRxiv, SSRN Electronic Journal, PeerJ Preprints, Nature Precedings, Research Square, EGU sphere and others (Dimensions, 2019 November 19). We have listed most of the repositories mentioned in Dimensions in Table 7.3, including arXiv, bioRxiv, ChemRxiv, SSRN Electronic Journal, PeerJ Preprints and Nature Precedings. With over 4.2 million preprint records, Dimensions has a much higher number of preprints than WoS and Scopus. According to our database investigations, a valuable feature of Dimensions for our case study is that it links preprints to the resulting peer-reviewed journal articles where possible. Each preprint page presents information such as authors,

abstract, references, and the resulting publication. The links to original preprint documents are also added to the journal article publication page in Dimensions.

Based on the above analysis, only the Dimensions platform will likely provide the most suitable data for this case study, so we need to verify the coverage of preprints further on the Dimensions platform. We want to study the impact of documents published between 2013 and 2014; thus, the preprint platform needs data from 2013 onwards. According to Table 7.3, arXiv, bioRxiv, SSRN Electronic Journal, PeerJ Preprints and Nature Precedings are covered by Dimensions and have data from at least 2013. Therefore, we selected a set of random preprints (n=500) from each repository and then matched them with the documents in Dimensions. Our sample selection and comparing results are summarized in Table 7.4.

Table 7.4 Matching results of preprint repositories in Dimensions

Repository name	Documents	Matched	Matching %
arXiv	500	500	100%
bioRxiv	500	500	100%
SSRN Electronic Journal	500	475	95%
PeerJ Preprints	500	500	100%
Nature Precedings	500	500	100%

It was found that all randomly selected preprints from arXiv, bioRxiv, PeerJ Preprints, and Nature Precedings were covered by Dimensions. Hence, the matching rate of these preprint repositories is 100% in this sample, respectively. The investigating method is DOI matching. All preprints from the above repositories have valid DOIs. For Nature Precedings, different preprint types include articles, posters, and presentations, all covered with Dimensions in our tests. It is noted that Nature Precedings only has preprint data from 2006 to 2012.

For preprints from the SSRN server, 25 out of 500 papers failed to locate in Dimensions, so the matching rate of SSRN is 95% under this data sample. However, we believe Dimensions' actual coverage is better than this result. SSRN preprints published after 2014 have valid DOIs, and all these DOIs were successfully located in Dimensions, even the preprints written in non-

English languages. Whilst preprints published before 2014 only have SSRN abstract numbers rather than DOIs. Therefore, we changed the matching method to search for titles and author names and found that 25 documents failed to reveal a relevant result in Dimensions. Searching titles and authors' names are less accurate than the DOIs. Besides, Dimensions reported search errors when titles involved unidentified punctuation marks, such as question marks. Dimensions might have covered the missing preprints, but we failed to locate them in this test due to punctuation marks or typing mistakes. As a result, the actual coverage of Dimensions from SSRN might be better than that of this sample test.

To conclude, Dimensions has an outstanding number of preprint records and a favourable linkage feature that suits our research objective. Our experimental matching test verified that the preprint coverage in Dimensions is satisfactory. Consequently, we will use Dimensions' preprint data in our case study.

7.2.3 Measurement data in S3

After research knowledge has been published, the stage of the first wave of knowledge dissemination (S3) starts. At this stage, key stakeholders will view and download publications, and measurement indicators could include the number of downloads of the research output, abstract views, full-text views, clicks, and more, as discussed in Section 6.3.1.

Several aggregate databases, such as ScienceDirect, provide their own counting of download times of the publications (Yom, 2023). The limitation of download times in ScienceDirect is that they only cover publications under its database, most of which are published by Elsevier. PlumX has metrics on full-text and abstract views, but their coverage is limited. There used to have views data from EBSCO, the largest information services provider to libraries of many countries, but the latest information showed that PlumX has stopped cooperating with EBSCO. We tested a small sample with 100 articles in PlumX, and only seven articles have usage data, which is 7% coverage based on this test. Apparently, PlumX has no satisfied views data.

The Scopus database provides a “views count” of each publication from 2021 as an article metric (Gibson, 2021, November 8). According to Scopus websites, the Views Count metric is aligned with its counterpart in SciVal and is updated every two weeks. It shows the number of views/visits a document has received in the current year, alongside the count for the previous year and the last 10 years. This is equivalent to saying that the metric is counted from 2022 onwards, which cannot be applied to articles published before 2022. Consequently, Scopus’ views count is not applicable if we target evaluating publications published before 2022.

Another point of view is to track the download count from major publishers. For example, Springer provides download count as “article accesses”, which include the number of times full-text or PDF versions are accessed directly from the journal website or SpringerLink. This includes actions such as viewing HTML, LookInside, PDF, or Epub files (Pinkowitz, 2012). Elsevier, however, does not display download counts on its ScienceDirect platform, although authors can view the download statistics of their own papers via the Mendeley platform, owned by Elsevier (Schlögl et al., 2014). Elsevier addressed that recording the views of publications has privacy issues that cannot be opened to the public as a public resource (Moed and Halevi, 2016). Therefore, public access to download data for other authors’ papers remains unavailable.

After extensive research and analysis, it has become evident that a comprehensive bibliographic database capable of providing view or download metrics for publications remains unavailable. While some publishers offer robust data on views and downloads, these metrics are limited exclusively to their published articles. Although Scopus has recently introduced view metrics, the statistics started in 2021, resulting in an absence of historical data from prior years. Besides, such metrics are confined to content indexed within the Scopus database. Consequently, it can be concluded that the necessary data for S3 is not currently accessible for use in our case study.

7.2.4 Measurement data in S4

In the stage of knowledge understanding and extensive dissemination (S4), many indicators can be found to reflect this stage, especially altmetrics. For instance, we could collect the number of times the research output is saved to reference management software, the number of times the research output is discussed on various media or social networks, and more.

Altmetrics are available from several altmetric data providers. In Chapter 4, we studied and investigated the altmetric data quality and database in depth, and we drew conclusions on how to use altmetric data within different platforms. Therefore, we will collect the relevant data for our case study according to the findings in Chapter 4.

7.2.5 Measurement data in S5

In the knowledge transformation and regeneration stage (S5), new forms of output containing the understood research knowledge were conducted. Evidence data at this stage reflects the explicit impact of the research. Various forms of citations are the most appropriate evidence data at this stage, such as publication citations, patent citations, policy document citations, and clinical trial citations. There are other forms of new output that contain the understood research knowledge, such as new designs in a project or new inventions in a product. However, such data is very challenging to obtain. Below, we will discuss the best database suitable for collecting different forms of available citation data.

7.2.5.1 Publication citations

The most widely used publication citation databases in bibliometric analysis include the Web of Science, Scopus, Dimensions, Crossref, Google Scholar, and Microsoft Academic. Numerous studies have carried out large-scale comparisons among these citation databases. Therefore, we will summarize the current literature's comparison results to aid in choosing and using these databases. We will also study each database's citation mechanism to validate the reliability of the data.

(1) Web of Science

The Web of Science Core Collection database is a selective citation index of scholarly publications involving journals, books, conference proceedings, and data compilations. The selectivity of indexing coverage that focuses on significant international journals was decided by its creator, Eugene Garfield (1925–2017), who wrote it in Garfield's Law of Concentration (Garfield, 1971; Garfield, 1972). Web of Science (WoS) database is formed by two parts: *Web of Science Core Collection* and *Web of Science Platform*. The WoS Platform citation indexing uses WoS core collection as the standard data set. If a journal gets indexed by its core collection database, all articles in that journal are covered in the WoS Platform (Clarivate, 2023, August 10).

Some scholars argue that the Web of Science has a lower coverage of journals in many disciplines and only includes a few journals from non-English-speaking countries (Basson et al., 2022). However, the Web of Science safeguards the quality and reliability of citation indicators through a rigorous selection process on its core collection database. As most literature agrees, the Web of Science is the most selective and authoritative citation database for use today (Birkle et al., 2020; Schnell, 2017; Singh et al., 2021; Visser et al., 2021).

(2) Scopus

Scopus, owned by Elsevier, is a comprehensive abstract and citation database covering various scholarly contents. Scopus's indexing coverage is decided by a Content Selection and Advisory Board (CSAB), which also involves continuous re-evaluations (Singh et al., 2021). Scopus employs sophisticated indexing algorithms to ensure accuracy and consistency in citation data. To uphold its status as a top-tier data provider and advance quality standards, Scopus implemented internal review procedures to consistently oversee specific areas of quality emphasis, including processing efficiency, profile integrity, and the thoroughness and precision of source data. These measures enable the content team to detect emerging data patterns promptly and track advancements in critical endeavours to enhance quality (Baas et al., 2020).

Visser et al. (2021) produced a large-scale comparison of bibliographic data sources, using Scopus as the baseline to compare document coverage, completeness, and accuracy of citation links with other four major databases including Web of Science, Dimensions, Crossref and Microsoft Academic. The conclusion is Scopus and WoS have certain documents that are not covered, respectively, and nearly all journal articles indexed in WoS are also indexed in Scopus.

(3) Dimensions

Orduña-Malea and Delgado-López-Cózar (2018) analyzed Dimensions' coverage and found its annual growth of scientific literature exceeds that of Scopus and Web of Science Core Collection, though it slightly lags behind Google Scholar. Citation counts in Dimensions are slightly lower than Scopus, but strong correlations exist between the two at the author, document, and journal levels. Basson et al. (2022) highlight Dimensions as an emerging platform that combines the advanced functionalities of traditional bibliographic databases (e.g., metrics, search filters, sorting) with the strengths of Google Scholar (e.g., broad

(4) Microsoft Academic and OpenAlex

Microsoft Academic is also one of the primary citation databases mentioned in several works of literature. Established in February 2016, it is a free public web search engine for academic publications and scientific documents. According to Thelwall (2018d), Microsoft Academic utilized semantic search capabilities and was powered by Microsoft's artificial intelligence technology, allowing it to understand the intent behind search queries and the relationships between topics, resulting in more relevant search results than traditional keyword-based searches. Microsoft Academic has good coverage of scholarly content, as it is studied by Visser et al. (2021) through large-scale data analysis, and the conclusion is that Microsoft Academic has the most inclusive coverage of the scientific literature among WoS, Scopus, Dimensions, and Crossref.

The sad news is that Microsoft shut down its Microsoft Academic websites and APIs in 2022 (Chawla, 2022). OpenAlex, another comprehensive database providing rich information on

scholarly publications and authors, announced that it would be an alternative to Microsoft Academic as it merged all of Microsoft Academic's metadata.

OpenAlex aims to provide researchers with easy and free access to the search for scholarly work, empowering them to delve into the intricacies of patterns, trends, and interconnections within academia (Priem, Piwowar & Orr, 2022). It contains metadata about scholarly literature, including authors, titles, institutional affiliations, citations, concepts, and funders. According to its official website²⁶, OpenAlex beta launched in January 2022. The data is accessible through both an API and a complete database snapshot, with the beta version of the web application for data exploration and search functionality being introduced in October 2023. Jason Priem, the co-founder of the OurResearch team that runs OpenAlex, presents that the database collects data from multiple sources, including Microsoft Academic's existing records, academic publishers, arXiv, ORCID, CrossRef, and institutional repositories (Priem, Piwowar & Orr, 2022). It also integrated data from PubMed and PubMed Central. By combining data from diverse sources, OpenAlex ensures comprehensive coverage of scholarly publications across different disciplines. A recent large-scale study comparing OpenAlex with subscription databases, including Web of Science and Scopus, revealed that OpenAlex demonstrates similar average source reference numbers and internal coverage to both Web of Science and Scopus (Culbert et al., 2024).

OpenAlex has the potential to be used widely in the scientific community due to its openness and ease of use. However, the quality of its citation data requires further study. This database is still being developed, and research on its data completeness is limited. We do not currently choose this database because there are other stable and reliable databases to choose from, such as WoS, Scopus, and Dimensions. However, OpenAlex is a very promising indexing database that will be useful in the near future.

²⁶ <https://help.openalex.org/about-us>

(5) Google scholar

Google Scholar's citation count algorithm attempts to provide a comprehensive view of the impact of scholarly work. Google Scholar employs an inclusive and automated approach, indexing virtually any academic document its crawlers discover and can access online, even those restricted behind paywalls through arrangements with publishers (Martín-Martín et al., 2021). This approach results in a broader range of scholarly content included in its citation database. However, Google Scholar's citations may not always be accurate or up-to-date, as they rely on automated indexing and sometimes include non-scholarly sources and self-citations (Harzing & van der Wal, 2008).

Google Scholar lacks restrictive standards on cited documents, which may cause some non-scholarly content to be included in its citation count. Consequently, its citation data cannot significantly represent academic impact but is mixed with other aspects of impact. Therefore, we do not intend to use Google Scholar's citation data in the current case study.

To sum up, as each database has its own advantages and disadvantages, we decided to use WoS to collect publication citations to ensure the recognition and reliability of the data.

7.2.5.2 Patent citations

(1) Worldwide patent offices and databases

The discussion of patent citation databases shall start from the worldwide patent offices. Each country typically has its own patent office responsible for granting patents within its authority. The "Big Five" patent offices, also known as the "IP5" offices, are the most prominent in the world in terms of the number of patent applications they receive and grant, which include US Patent and Trademark Office (USPTO), European Patent Office (EPO), Japan Patent Office (JPO), Korean Intellectual Property Office (KIPO) and National Intellectual Property Administration in China (CNIPA) (Lee et al., 2021).

Most Patent offices have their own patent databases. For instance, PatentScope is WIPO's online database, offering access to over 4 million international Patent Cooperation Treaty

(PCT) applications and patent documents around the world (WIPO, 2024). Espacenet is a free patent search tool provided by the European Patent Office. It offers access to millions of patent documents worldwide, including those from the EPO and other patent offices (Espacenet, 2024). Furthermore, the USPTO database provides access to patents issued in the United States, including full-text documents and images, and the JPO database provides access to patents granted in Japan (Lee et al., 2021). The limitation of the above patent databases is their regional coverage.

(2) Comprehensive database for patents

Access to a comprehensive database is easy for users when searching for patent information. In our case study, a patent database that is as complete as possible to cover all patent offices is crucial. After deep investigations, the leading databases for comprehensive patent information are summarised in Table 7.5, along with the number of patent records and data coverage. The detailed analysis of each database and discussions on whether it suits our case study are drawn below.

Table 7.5 Leading database for comprehensive patent data

NO.	Database name	Patent amount	Coverage
1	Google Patents	Over 120 million patent publications as of 2024	Provides access to patent information from over 100 patent offices around the world.
2	The Lens	Over 153 million global patent records	Encompasses over 95 jurisdictions of global patent documents. It also excels in integrating patent data with scholarly publications.
3	Dimensions	over 140 million patents	Covering patent data from over 100 patent offices worldwide
4	Derwent Patents Citation Index™ (DPCI)	162 million backward citations, 182 million forward citations and 39 million literature citations	Covers patent-issuing authorities from 32 countries.
5	MATrixware REsearch Collection (MAREC).	19 million patent applications and granted patents	Covers major patent office in the world such as EPO, USPTO, JPO.

Google Patents is a widely used free-access platform for searching and analysing patent data. It offers patent keyword search, full-text patent documents, and patent family information²⁷. The database contains more than 120 million patent publications from over 100 patent offices globally, along with various technical documents and books indexed in Google Scholar and Google Books²⁸. Google Patents is an excellent starting point for patent searches due to its broadness and user-friendly interface. However, it lacks index patent citations or comprehensive analytical tools to help with in-depth bibliometric analysis (Kousha & Thelwall, 2017). Although Google Patents embed Google Scholar for technical document search, which makes prior art searching easier and enhances the features of patent information search, it does not separate its cited scholarly work from patent documents (Krabbe, Sampson and Wetherbee, 2017). It is disappointing that Google Scholar and Google Patents could not match the links between research work and patent citations.

As Google Patents does not support automatic API searches, Kousha and Thelwall (2017) introduced a semi-automatic indirect method via Bing searches to extract and filter patent citations from Google. They collected data by applying research papers' titles and first author surnames to crawl from the Google Patents websites. Furthermore, Younes and Rassenfosse (2024) applied BigQuery language to source the raw data from the Google Patents Public Data to construct their replicable patent indicators study. Therefore, using the Google Patents database is theoretically feasible but requires text-mining techniques and handling data processing errors. If we cannot accurately match the research papers with cited patents by text mining in Google Patents' full-text documents, our data quality still needs to be questioned. We found the Lens database a great alternative, which will be introduced in the following sections.

The Lens database is an open global cyberinfrastructure to make the innovation system more efficient, fair, transparent, and inclusive (About the Lens, 2024). It is free to use, provides

²⁷ https://support.google.com/faqs/answer/7049475?hl=zh-Hans&ref_topic=6390989&sjid=6893726050779330086-EU

²⁸ <https://support.google.com/faqs/answer/7049585#>

patent information and citation analysis, and links patents to scholarly research. The Lens enables users to share, annotate, and embed document collections, aggregations, and analyses, creating an open map of knowledge-driven innovation. This approach aims to rejuvenate the patent system's function as an educational resource, inspiring and informing entrepreneurs, citizens, and policymakers (Krabbe et al., 2017).

According to the Lens's official websites, after over 20 years of development, the Lens has evolved into a comprehensive database with over 268 million scholarly works, 153 million global patent records, and more than 489 million biological sequences, all with unprecedentedly rich metadata (About the Lens, 2024). It encompasses over 95 jurisdictions of global patent documents, such as European Patent Office's (EPO) DocDB bibliographic records from 1907, USPTO Applications since 2001, USPTO Grants since 1976, USPTO Assignments, European Patent Office (EPO) Grants from 1980, WIPO PCT Applications from 1978, and the full text of Australian patents from IP Australia (Saheb & Saheb, 2020). Furthermore, Lens' scholarly datasets integrate with Microsoft Academic, CrossRef, PubMed and OpenAlex and support the search of over 272 million scholarly works (About the Lens, 2024).

Apart from the outstanding coverage of patents and scholarly works stated above, the great advantage of the Lens database is its integration of both scholarly and patent literature. It allows researchers not only to explore patents or academic records but also to easily assess the impact of scholarly works on patents. Many studies in the literature used the lens data for patent analysis (Firdaus et al., 2024; Penfold, 2020; Saheb & Saheb, 2020; Sarjidan et al., 2023), which proves its suitability and popularity in academics and research. We will apply a data sample to verify the data quality of Lens in the next section.

Another database that integrates scholarly publications and patent citation data is Dimensions. Dimensions is committed to providing a wide range of data throughout the research process, so they also attach great importance to patent citation data. We reviewed

this database in Section 3.5.2 as an altmetrics provider; here, we focus on the quality of its patent data. Dimensions' patent data is collected and provided by IFI Claims, a patent service company. IFI Claims was fully acquired by Digital Science in 2021. According to Dimensions (2022, October 28), the data source from IFI Claims includes bibliographic data from DOCDB (the European Patent Office's master documentation database covering patent data from over 100 countries), Legal status from INPADOC (an event database produced by the European Patent Office and covering over 50 international patent authorities), and bibliographic and full-text data from majority national patent offices in the world. By enlarging patent data in the life sciences and chemistry category, Dimensions has extended its patent data to over 140 million documents from over 100 patent offices worldwide (Dimensions, 2022 May 11). We will also apply sample data to verify the patent data quality in Dimensions in the next section.

Furthermore, the Derwent Patents Citation Index™ (DPCI), developed by Clarivate, also provides comprehensive patent citation information with cited scholarly work (Clarivate, 2024). DPCI is a subscribed database that offers both backward and forward citations of patents. Backward citations refer to prior patents and literature that a given patent cites, while forward citations are subsequent patents that cite the given patent. As stated on its website, DPCI has 162 million backward citations, 182 million forward citations and 39 million literature citations. Although DPCI's database is enlarging, currently, it only covers patent-issuing authorities from 32 countries (Clarivate, 2024).

Since DPCI and WoS are the same company, we have searched for connections between the two databases. The WoS does not directly index patents in its citation databases, so we cannot search the document's relevant patent citations from the research paper side. However, if we have a patent number available, WoS allows searches for citations to patent documents using the "Cited Reference" search feature (Clarivate, 2022, May 25). It works by entering the patent number in the Cited Title field. We used a few patent numbers that had been randomly selected and tried over WoS, and the results showed poor patent information

presentation. Considering that the data coverage of DPCI is limited and there is no way to access scholarly work's patent citations from the WoS side, we will not consider using DPCI or WoS for patent data gathering.

Lastly, the MATrixware REsearch Collection (MAREC) is another standardised patent data in XML format to flavour the purposes of research. It consisted of around 19 million patent applications and granted patents in English, German and French for a period of 32 years (1976-2007) (Chakraborty et al., 2020). The dataset covers specific countries and major patent offices, such as the USPTO (United States Patent and Trademark Office), EPO (European Patent Office), JPO (Japan Patent Office), and others. If we want to use this database, text processing will be needed to match patent and cited research work, as MAREC's data are stored in a unified file in XML format. Another issue is that not all records of cited articles in patents provide DOIs. If not using DOI matching, an alternative method could be using author and title for matching, but this will easily cause errors and be time-consuming. Therefore, we will not choose MAREC's patent data in our case study.

(3) Sample test for Lens and Dimensions

The last section's discussions showed that Lens and Dimensions are the most promising patent databases, with extensive coverage and high quality. Therefore, we applied a data sample to test their database for further verification. We randomly selected 1000 articles from the WoS database, collected their DOIs, and then downloaded associated patent citation data from Lens.org and Dimensions.ai, respectively. These 1000 articles can be found in both databases, so our comparison focuses on patent number records.

The sample test results (see Table 7.6 & 7.7) showed that the patent counts in Lens are higher than in Dimensions, and data between the two databases are significantly correlated. Clearly, Lens has the best patent citation records. Therefore, we will collect patent data from the lens.

Table 7.6 Descriptive statistics of patent citations for Lens and Dimensions

	Lens Patent Citations	Dimensions Patent Citations
Mean	2.43	1.68
N	1000	1000
Std. Deviation	6.409	5.119

Table 7.7 Correlations of patent citations for Lens and Dimensions

			Lens Patent Citations	Dimensions Patent Citations
Spearman's rho	Lens Patent Citations	Correlation Coefficient	1.000	.863**
		Sig. (2-tailed)	.	.000
		N	1000	1000
	Dimensions Patent Citations	Correlation Coefficient	.863**	1.000
		Sig. (2-tailed)	.000	.
		N	1000	1000

** . Correlation is significant at the 0.01 level (2-tailed).

7.2.5.3 Clinical trial citations

Clinical trials are research investigations in which people volunteer to test new treatments, interventions, or tests as a means to prevent, detect, treat, or manage various diseases or medical conditions (World Health Organization, 2024). Typical investigations include examining how people respond to a new intervention and what side effects might occur. In short, clinical trials are a type of research that studies new tests and treatments and evaluates their effects on human health outcomes.

Clinical trial citations provide evidence for the efficacy and safety of medical treatments and interventions, which is fundamental for evidence-based medicine (Subbiah, 2023). Van Eck et al. (2013) alleged that the citation impact of clinical intervention research may be significantly underestimated when compared to basic and diagnostic research. It is believed that citation analysis is an effective tool for research performance assessment in the medical sciences, while clinical research citations may have more impact than popular bibliometric indicators in medical sciences studies. The gradually emerging clinical trial data in recent years may provide more opportunities to study the impact of basic medicine on clinical outcomes.

Based on our research, the following databases are highly regarded in terms of the breadth and depth of clinical trial citations:

1. PubMed is one of the most comprehensive biomedical literature databases, maintained by the National Centre for Biotechnology Information (NCBI). It indexes millions of citations from biomedical and life sciences journals, including many clinical trials. It is freely accessible and provides links to full-text content when available.
2. ClinicalTrials.gov is a comprehensive registry and results database of publicly and privately supported clinical studies conducted worldwide. It is managed by the U.S. National Library of Medicine (NLM) at the National Institutes of Health (NIH) and provides detailed information about clinical trial design, sponsors, locations, and outcomes.
3. Embase is produced by Elsevier and is known for its extensive coverage of biomedical literature, including many clinical trials not indexed in PubMed. It covers over 8,500 journals and includes comprehensive indexing of drug and medical device studies. This database requires subscriptions to access.
4. Scopus covers a wide range of scientific disciplines, including clinical trials. The clinical trial citation data is available in PlumX and is embedded in Scopus' metrics.
5. Dimensions aggregates data from various registries and sources, providing extensive coverage of clinical trials worldwide. One advancing feature of Dimensions is its data linkage with publications. Clinical trial records in Dimensions are linked to related publications, making it easy to trace the progression between published articles and trial registration.

PubMed is the most commonly used abstract and reference database for life science and biomedical disciplines. Scopus and Dimensions both claim to have integrated PubMed data and support literature searches by PubMed identifiers. Although ClinicalTrials.gov is the world's largest clinical trial registry, it does not index the cited or referenced publications for registered clinical trials (Huser & Cimino, 2013). Dimensions has clearly stated its clinical trial data coverage on its official websites. Dimensions collects data from international clinical trial

registries globally, including ClinicalTrials.gov (United States), EU-CTR (EU), CHICTR (China), ANZCTR (Australia/New Zealand), and more (Dimensions, 2023 December 5). Embase is not easy to access due to subscriptions. Therefore, we will focus on Scopus and Dimensions for clinical trial citation data.

To further compare the records of clinical trial citations in Dimensions and Scopus (data provided by PlumX), we carried a data sample test with 1000 highly cited medical papers published in 2013 – 2014. These 1000 papers come from more than 15 different journals in the medical discipline. We initially used the Dimensions filter to select the papers, so all the papers have Dimensions data. 17.2% (172 out of 1000) of papers were not found in Scopus, and therefore, we cannot extract missing papers' PlumX metrics through Scopus. Table 7.8 presents the basic statistics of the sample test.

Obviously, Dimensions' clinical trial citation count is higher than PlumX's, and we noticed that Dimensions' data mean is more than twice that of PlumX. Because the data sample is not normally distributed, a non-parametric correlation analysis using Spearman's rank method was carried out. The correlation result shows that the two databases' clinical trial citation data were not significantly correlated. Our sample data test result showed that Dimensions' clinical trial citation data is far better than PlumX. Therefore, we will use Dimensions' clinical trial citation data.

Table 7.8 Descriptive statistics of clinical citations for Dimensions and PlumX

	N	Minimum	Maximum	Mean	Std. Deviation
Dimensions' clinical trial citations	1000	1	10	4.96	2.222
PlumX's clinical trial citations	828	0	15	2.19	2.975

Furthermore, PlumX used to have good clinical trial citation data when we primarily investigated at an earlier time. However, our latest data sample comparison shows that its coverage declined. It is noticed that from November 1, 2020, PlumX has discontinued comprising metrics from EBSCO, a leading global provider of resources for libraries, so PlumX

stops receiving EBSCO Usage and Capture data, as well as clinical citations from DynaMed Plus Topics (Faulkner, 2020 September 3). As a result, PlumX's data sources for clinical trial citations only include PubMed Clinical Guidelines and NICE Clinical Guidelines currently, and the latter is the data from the UK's National Institute for Health and Clinical Excellence (Plum Analytics, 2021 October 13).

Such changes also remind us that the emerging databases are somewhat unstable and may be restricted by the coverage of data sources at any time, affecting the quality of the data they provide. Besides, clinical trial data are mainly from medicine and life science studies, which are not suitable for the comprehensive impact comparison among different disciplines.

7.2.5.4 Policy document citations

Scholars attempted to determine whether policy document citations could be used as an indicator to assess the potential longer-term impacts of research, but there is no consistent conclusion (Newson et al., 2018; Willis et al., 2017; Yu et al., 2023). When research outcomes are mentioned in policy documents, it indicates a meaningful effect of research beyond the academic domain, particularly at the societal level (Bornmann, Haunschild & Marx, 2016). Policy document citation is a part of evidential data within our cause-effect chain of the research impact to trace the integrated and long-term impact of research.

The availability of policy document citations started from being one form of altmetric data source. As we reviewed in Section 3.5.3, early data providers of policy document citation include Altmetric.com and PlumX. Then Dimensions, which appeared in 2018, also covers policy document citation as one indexing source. Overton is a new altmetric database specialising in policy documents, launched in 2019 and proliferating fast. Overton claims to be the world's most extensive searchable index of policy documents, parliamentary transcripts, government guidance, and think tank working papers (Szomszor & Adie, 2022). As of June 2024, Overton covers over 13 million political documents from 43,542 organisations in 2,274 policy sources (Overton, 2024).

Previous experimental studies verified that Overton has a higher coverage of policy document mentions than the Altmetric.com database (Maleki & Holmberg, 2022; Szomszor & Adie, 2022). The Overton has also become a data provider to other platforms. For instance, Sage Policy Profiles is a web-based tool launched by Sage Publishing in 2023, aiming to demonstrate their publications' influence on policy. Their policy document data and citation index were provided by Overton (Chawla, 2023 December 19). Besides, it is noticed that PlumX has started using the Overton data for its policy citation metric since 2021. They believe navigating a global scale of policy documents is challenging, while Overton is doing an excellent job of focusing on policy document collection (Faulkner, 2021 October 13). Hence, the policy document citation in PlumX is Overton's mirror data.

Comparing the three databases, Altmetric.com, Dimensions, and Overton, for their policy citation records is of interest. However, a large-scale data analysis shall be left for future research due to time limitations, and here, we only carry a small data sample comparison to ensure the applicability of our case study.

We randomly selected 1000 articles in the Energy and Fuels category (within Chemistry category group) to form a data sample. Then, we collected policy citation data for the sample articles from three databases: Dimensions, Altmetric.com, and Overton, respectively. Dimensions collected 29 data (out of 1000), accounting for 2.9% of coverage under this data set. Altmetric.com collected 37 data, which accounts for 3.7% of the data coverage, and Overton collected 116, which accounts for 11.6% of data coverage in this data sample. The mean value between Dimensions, Altmetric.com and Overton is 2.03, 1.22 and 4.14, respectively (see Table 7.9). The correlation analysis in Table 7.10 showed that the three databases' policy citation data are moderated correlated. Altmetric.com and Dimensions have a higher correlation with a correlation coefficient of 0.661.

The above data sample extraction and comparison clearly showed that Overton's policy citation data has much wider coverage than Dimensions and Altmetric.com. The data values

in Overton are twice or three times higher than those in Dimensions and Altmetric.com, representing the policy citations collected by Overton, which cover more sources. Therefore, Overton is the most suitable database for collecting policy citation data for our case study.

Table 7.9 Descriptive statistics of policy citations for Dimensions, Altmetric and Overton

	Dimensions' policy citations	Altmetric's policy citations	Overton's policy citations
Mean	2.03	1.22	4.14
N	29	37	116
Std. Deviation	1.523	.479	5.607

Table 7.10 Correlations of policy citations for Dimensions, Altmetric and Overton

			Dimensions' policy citations	Altmetric's policy citations	Overton's policy citations
Spearman's rho	Dimensions' policy citations	Coefficient	1.000	.661**	.390*
		Sig. (2-tailed)	.	.000	.044
		N	29	29	27
	Altmetric's policy citations	Coefficient	.661**	1.000	.423*
		Sig. (2-tailed)	.000	.	.011
		N	29	37	35
	Overton's policy citations	Coefficient	.390*	.423*	1.000
		Sig. (2-tailed)	.044	.011	.
		N	27	35	116

** . Correlation is significant at the 0.01 level (2-tailed).

* . Correlation is significant at the 0.05 level (2-tailed).

Additionally, since PlumX uses Overton's policy data as data sources, PlumX's policy citation could be an alternative for Overton when the latter does not support free download. It is worth to verify the consistency between Overton and PlumX. Therefore, we also collected policy citation data from PlumX for the above data sample. The consistency approach used in Chapter 4 has been carried out, and we found that PlumX and Overton have only 85% consistency on Policy document citation. A few data in PlumX are missing or have less value count than Overton, and we noticed the less counted value is more like the newest citation. It indicated that PlumX's policy citation is delayed in time as mirror data. The updated

conclusion is that we will use Overton as a priority, and PlumX could be an alternative to Overton with awareness of minor data missing caused by time delay.

7.3 Data Collection

Based on the detailed analysis and validation results of relevant databases in the previous section (Section 7.2), we decided to download data reflecting different stages of impact generation from the following databases: (a) Dimensions, (b) Altmetric, (c) Web of Science, (d) Lens.org, and (e) Overton. The database corresponding to the specific data is shown in Table 7.11.

Next, we need to determine the publications that would be evaluated for this case study. According to the WoS category list, we chose five category groups as observing research areas, which are (1) Biology and biochemistry, (2) Chemistry, (3) Clinical Medicine, (4) Environment/Ecology, and (5) Engineering. Each category group contains five single categories randomly selected within the group according to the WoS category list. Consequently, we have 25 chosen categories under the five research areas. All journal papers published in 2013 within these 25 categories were collected. The total number of publications is 306,401.

It should be noted that we originally planned to download journal papers published in 2013 and 2014 in eight category groups. In the actual download process, since the corresponding evidence data needed to be downloaded from five databases respectively and some databases did not support bulk data downloads, we could only crawl data from websites or download through API with a limited flow rate. Completing all planned data downloads was time-consuming and would take more than a year, according to the current speed. Therefore, the observing category groups were reduced to five. The reduction would not affect the evaluation results of current research areas.

In the data download process, we first extract research papers' basic information and citations from the Web of Science. As we obtained all targeting papers' basic information and

their WoS citations, we can use the DOI of each paper to download the rest of the data from the other four databases, including Dimensions, Altmetric, Lens.org, and Overton. Raw datasets contained lots of information, and we only extracted the relevant data as desired. Table 7.11 summarises the final data used after data cleaning. For altmetric data, we have a total of 17 data sources collected. Since some altmetrics have a low volume of coverage in the data, we aggregated one indicator, “other altmetrics”, by combining the altmetrics below 1% volume in our dataset. Therefore, apart from the altmetric data indicators presented in Table 7.11, “other altmetrics” contains the sum of mentions on Reddit, YouTube video, Google+, F1000, Peer review forums, QnA, Weibo, Pinterest, Linkelin and book reviews.

Table 7.11 Data used in the evaluation of impact of research publications

No.	Data	Denotation	Collected from
1	Supporting grants	Number of grants supported to the publication	Dimensions
2	Preprints	Number of preprints linked to the publication	Dimensions
3	Mendeley readers	Number of readers for the publication in Mendeley	Altmetric
4	CitULike readers	Number of readers for the publication in CiteULike	Altmetric
5	X (Twitter) mentions	Number of tweets	Altmetric
6	Facebook mentions	Number of posts on public Facebook Pages	Altmetric
7	News mentions	Number of news mentioned	Altmetric
8	Blogs	Number of blogs mentioned	Altmetric
9	Wikipedia	Number of mentions in Wikipedia	Altmetric
10	Other altmetrics	Total mentions of all the other altmetrics	Altmetric
11	Publication citations	Number of citations of publications	Web of Science
12	Patent citations	Number of citations of patents	Lens.org
13	Policy document citations	Number of citations of policy documents	Overton
14	Clinical trial citations	Number of citations of clinical trials	Dimensions

The dataset contains many zero values as not every research paper has data in different stages of the impact generation process. To detect if any paper has no data, even a citation

number, all indicators are temporarily summed to detect zero-sum values. Research papers with zero-sum values for all indicators were eliminated. The final cleaned dataset has 296,586 publications.

7.4 Methods on impact evaluation modelling

The knowledge impact generation process of one's research (as demonstrated in Figure 6.7) can be modelled by a causal network structure. Each stage of the impact generation cannot be measured directly but can only be reflected through the observed data of the corresponding stage. The Structural Equation Modeling (SEM) is an advanced statistical approach that allows researchers to examine and model the relationships between observed and unobserved variables (Byrne, 2016). It integrates concepts from confirmatory factor analysis, which uncovers underlying factors from observed data, path analysis and multiple regression analysis, and evaluates how one set of variables influences another. The relationships between the variables in SEM could be correlational or bidirectional, isolated or conditionally independent, causal or unidirectional (Madhanagopal and Amrhein, 2019). SEM is adaptive to model relationships in our proposed evidence-chain-based methodology.

IBM SPSS Amos (Analysis of Moment Structures) is a user-friendly software for visual structural equation modelling. With Amos, we can easily create, view, and adjust models graphically using intuitive tools. Then, Amos could translate drawings into a programming interface and perform the necessary calculations (Collier, 2020).

7.4.1 Model establishment

Below, we will use the SEM method to build our models with Amos. The structural relationships of models were established from the conceptual model (demonstrated in Figure 6.7) for the research impact generation process in Chapter 6. Each of our SEM models contains the hypothesised effect of different stages of the impact generation process on the comprehensive research impact of publications. Each stage of the impact generation process, as identified in Section 6.2, has a logical causal relationship in time. Therefore, we have five hypothesised relationships:

H1. The stage of knowledge generation (S1) has a direct effect on the stage of knowledge publishing (S2).

H2. The stage of knowledge publishing (S2) has a direct effect on the stage of knowledge understanding and extensive dissemination (S4).

H3. The stage of knowledge understanding and extensive dissemination (S4) contributes effects to the impact of research papers (Impact).

H4. The stage of knowledge publishing (S2) contributes effects to the impact of research papers (Impact).

H5. The stage of knowledge generation (S1) contributes effects to the impact of research papers (Impact).

We must stress that the ideal conceptual model requires much data to build. In reality, we have limited data for modelling, for instance, the stage of the first wave of research knowledge dissemination (S3) is unavailable in the current model due to the lack of reliable download data. As such, the parameters of our models may have large uncertainty. Our indicators are carefully selected from available and reliable databases to ensure data quality, as discussed in Section 7.2. When more data sources become available and reliable, the uncertainty in parameter estimation can be reduced. Based on the existing data, we built a hypothesised structural equation model, as shown in Figure 7.1.

The model in Figure 7.1 contains two parts: the measurement model and the structural model. In SEM, the measurement model labels the connections between the latent variables and their manifest indicators, and the structural model describes the possible causal relationships between the latent variables (Byrne, 2016). The expression of latent variables and their indicators is explained below.

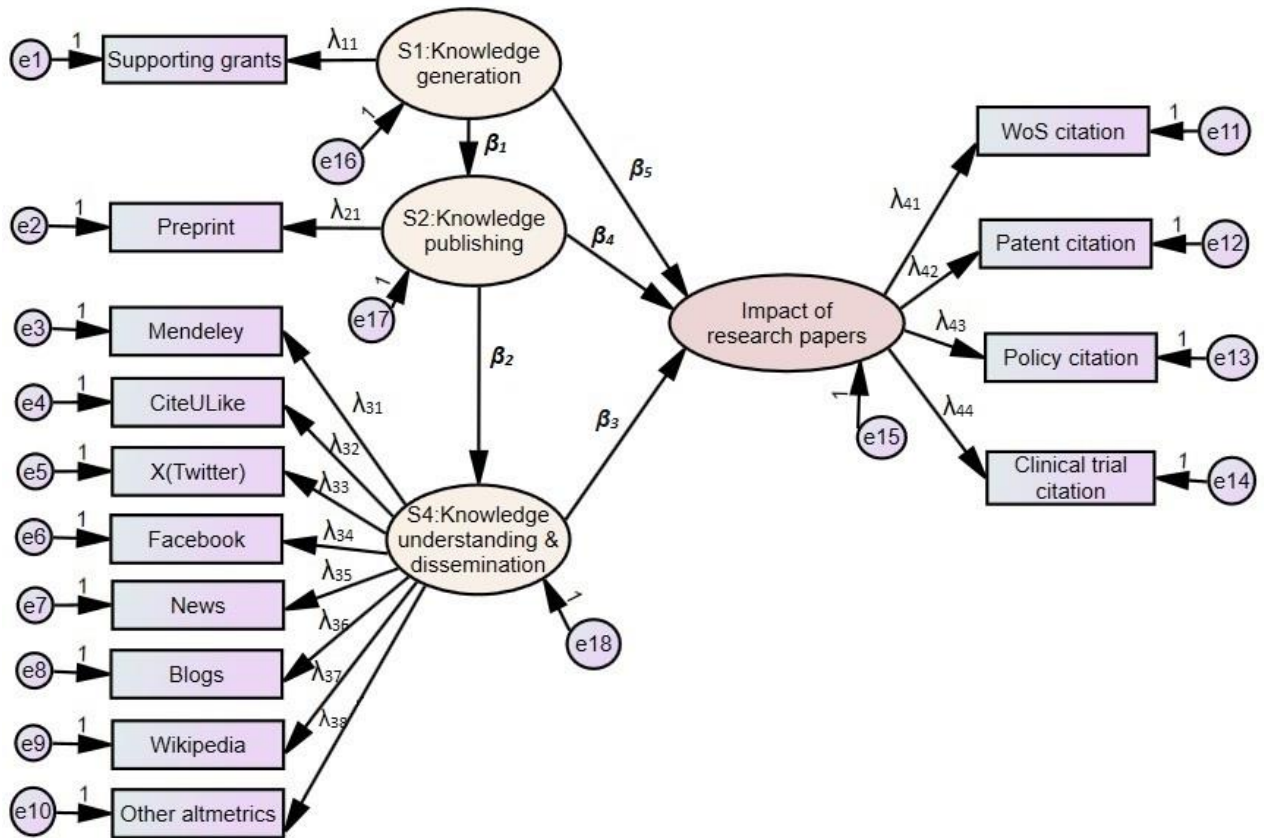


Figure 7.1 The hypothetical model for the impact of research papers

We have four measurement sub-models (also see Figure 7.2). Each sub-model was established according to the currently available and accessible data, with a relationship based on the conceptual model in Chapter 6. The impact of research papers and the different stages of the impact generation process cannot be directly observed, so they are the latent variables, as shown as ellipses in the graph model. The knowledge generation stage is named S1. The supporting grants are the available evidence data we can collect at this stage, so it is the manifest indicator for S1. Similarly, the knowledge publishing stage is S2, and its indicator is preprint data. Although published research papers are the result of S2 in our conceptual model, the published papers are not a meaningful indicator because all documents in this data are published. The stage of knowledge understanding and extensive dissemination is named S4 in the model, and its indicators are available altmetric data. Our data sample contains 17 altmetric data sources. Data sources with a volume higher than 5% are used as indicators for S4, such as Mendeley and Facebook. The other data sources with less than 5%

of valid data volume, such as Reddit and QnA, are combined into one indicator, named “other altmetrics”. The dependent variable, the impact of research papers (hereinafter referred to as “Impact”), has four manifest indicators, including the number of WoS citations, Patent citations, Policy document citations, and Clinical trial citations. These indicators represent different dimensions of impact on research outputs, such as academic impact, economic influence, social impact, and health and well-being impact. All manifest indicators in the model were carefully selected from available and reliable databases to ensure good data quality, which has been discussed in detail in Section 7.2.

Since data prediction will not be perfect, models must include error variables (Byrne, 2016), denoted as e_1 to e_{18} in Figure 7.1. The error terms e_{15} to e_{18} also present the residual effects that affect the dependent variable but are not explicitly included in the model.

In structural models, arrows connect (possible) causal links between variables. Each stage of the impact generation process, S_1 , S_2 , and S_4 have a logical relationship in time. At the same time, S_1 , S_2 , and S_4 can also contribute to the impact. We denote β and λ as the coefficients in the structural model to ease later analysis, as shown in Figure 7.1.

The graphic model in Figure 7.1 can be translated into the following equations:

$$S_2 = \beta_1 S_1 + e_{17}$$

$$S_4 = \beta_2 S_2 + e_{18}$$

$$Impact = \beta_3 S_4 + \beta_4 S_2 + \beta_5 S_1 + e_{15}$$

$$Supporting_grants = \lambda_{11} S_1 + e_1$$

$$Preprint = \lambda_{21} S_2 + e_2$$

$$Mendeley = \lambda_{31} S_4 + e_3$$

$$CiteULike = \lambda_{32} S_4 + e_4$$

$$X(Twitter) = \lambda_{33} S_4 + e_5$$

$$Facebook = \lambda_{34} S_4 + e_6$$

$$News = \lambda_{35} S_4 + e_7$$

$$Blogs = \lambda_{36}S_4 + e_8$$

$$Wikipedia = \lambda_{37}S_4 + e_9$$

$$Other_altmetrics = \lambda_{38}S_4 + e_{10}$$

$$WoS_Citations = \lambda_{41}Impact + e_{11}$$

$$Patent_Citations = \lambda_{42}Impact + e_{12}$$

$$Policy_Citations = \lambda_{43}Impact + e_{13}$$

$$Clinical_trial_citations = \lambda_{44}Impact + e_{14}$$

7.4.2 Model Identification

Our hypothesised structural equation model contains 4 latent variables and 18 error terms. Some assumptions for parameters could be made in the measurement model before estimating the parameters in the structural model. If the path coefficient in the measurement model is positive, the rationality of the model can be further confirmed (Byrne, 2016).

Firstly, we could start adding constraints in measurement models (as seen in Figure 7.2). It is common to set regression coefficients to 1 for error variables. In Amos, this is a default solution unless we have individual needs for regression coefficients of errors (Arbuckle, 2022). The Amos prompted that 5 additional constraints are likely necessary to achieve identifiability in our first running attempt. Therefore, we need to add at least 5 constraints. Reducing the number of parameters to be estimated is also essential to identifying the model (Blunch, 2013). According to Arbuckle (2022), the common way to estimate a parameter is to set a regression weight for at least one path in the measurement model that point from the unobserved variable to its manifest indicators. This weight can be arbitrary but is usually set to 1 as an initial value, which suggests the unobserved variable will have the same scale as the indicator in question. When an unobserved variable has more than one indicator, fixing any of them does not affect the estimates of standardised regression coefficients. Therefore, we set up one indicator with regression coefficients on value 1 as the initial value for each of the measurement sub-models, as shown in Figure 7.2.

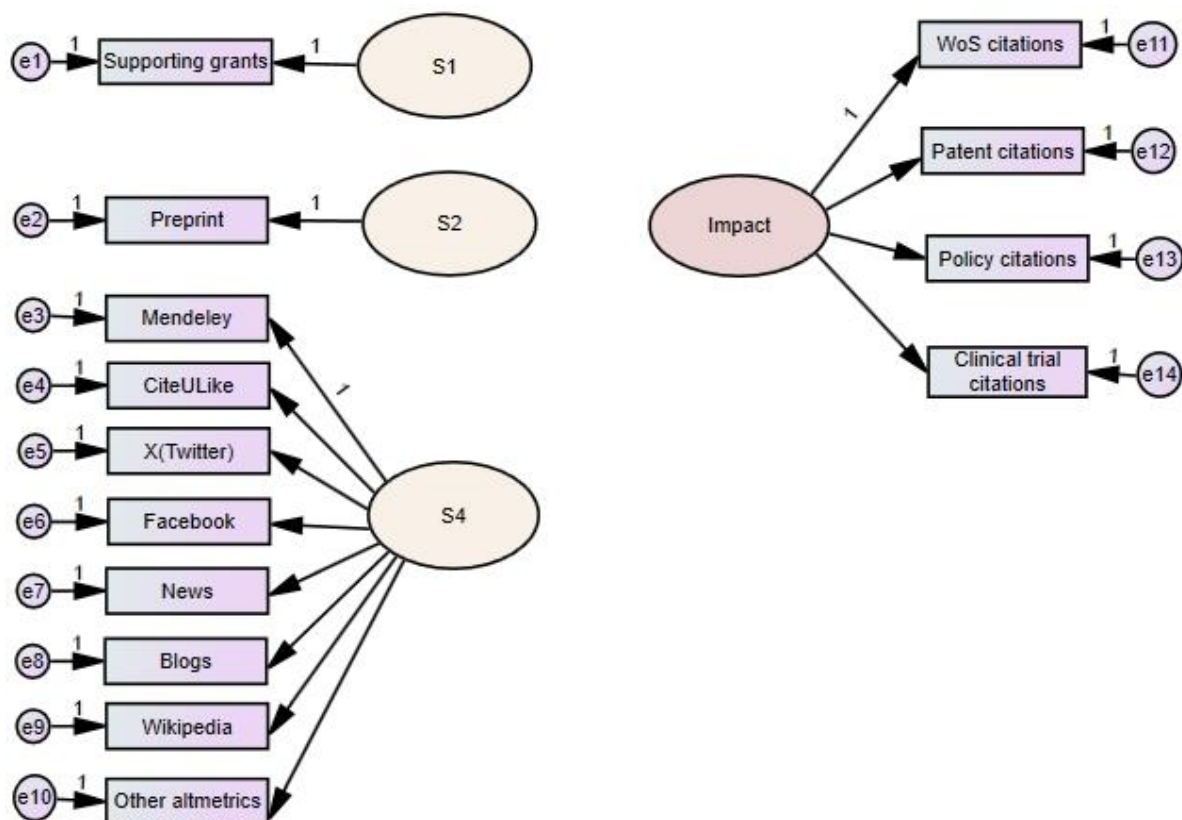


Figure 7.2 Measurement models for the impact of research papers

There is a need for at least one more constraint to be added to achieve identifiability. Blunch (2013) suggested that fixing the latent variable's variance at an arbitrary value, such as 1.00, can also help identify a model. Besides, a parameter can be restricted in three ways: (1) The parameter can be pre-defined to assume a specific value; (2) Two or more parameters can be constrained to have the same unspecific value; (3) The parameter can be omitted from the model, as it is constrained to have a value of 0 (Blunch, 2013). There could be hundreds of ways to fix a regression weight or variance in a model. The more parameters are constrained, the easier the model is to achieve identifiability. However, we will lose the degrees of freedom and model fit if the parameters are over-fixed. A hypothesised model with a low degree of freedom is unattractive, and the model fit is vital to present adaptability. Here, we will apply the Akaike information criterion (AIC) as the model performance measure in selecting the most suitable model.

The AIC is a statistical measure for evaluating how well a model fits the data. AIC is used to compare different possible models and determine which one is the best fit for the data. The formula for AIC is $AIC = 2k - 2\ln(\hat{L})$, where k is the number of estimated parameters in a model and $\hat{L}$ is the log-likelihood estimate (Stoica & Selen, 2004). According to Burnham and Anderson (2004), the AIC penalizes models that use more parameters. This penalty helps prevent overfitting, which is important since adding more parameters to a model typically enhances the fit. The best model in the model pool is the one with the lowest AIC value. Therefore, if a set of models explain the same amount of variation, the one with fewer parameters will have a lower AIC score and will be the better-fit model.

After a series of attempts at changing parameters, we have found when we constrained the variance of e1 and e16 at the same unspecific value and let the Amos find the estimated variances, the identified structural model had the lowest AIC score. If we fix the variance at a specific value, such as 1, the AIC score is still higher than the lowest one found. Many attempts at fixing different regression coefficients or variances were compared. All groups of data were tested. Therefore, we have obtained the primary parameter settings for our structural model. The model has been appropriately identified to assess hypotheses. Using the same initial parameter settings, we can also compare the estimated coefficients across different category groups.

7.5 Model analysis and discussions

7.5.1 Structural model analysis

We would initially explore the structural model results from all the data that combined five category groups. The sample size is 296,586.

The first attempt to run the structural model with all data showed an error message. The variances of error term e12 were negative. The issue might stem from failing to account for existing correlations among some indicators for the same latent variable (Blunch, 2013). These correlations typically arise when there are challenges in designing indicators that are

distinct enough to capture the same concept. One way to solve this issue is by including the necessary correlations. Alternatively, if two correlated indicators are highly similar, we can remove one. The indicator of patent citations is vital and, indeed, cannot be removed. Therefore, it is necessary to add correlations of indicators. The bootstrap is performed to observe the approximate standard errors for every estimate that Amos computes. The modification indices suggested by Amos are considered. It must be noted that modification indices provide many additional connections between variables to improve model fit while not always being feasible and meaningful, so judgment is needed. Subsequently, we carefully introduced the additional correlation between indicators after justifying the rationality to the best of our knowledge. Additionally, since my data sample is relatively large, the iteration limit has been raised from 50 to 500 to avoid invalid results. Figure 7.3 shows the improved structural model and key results.

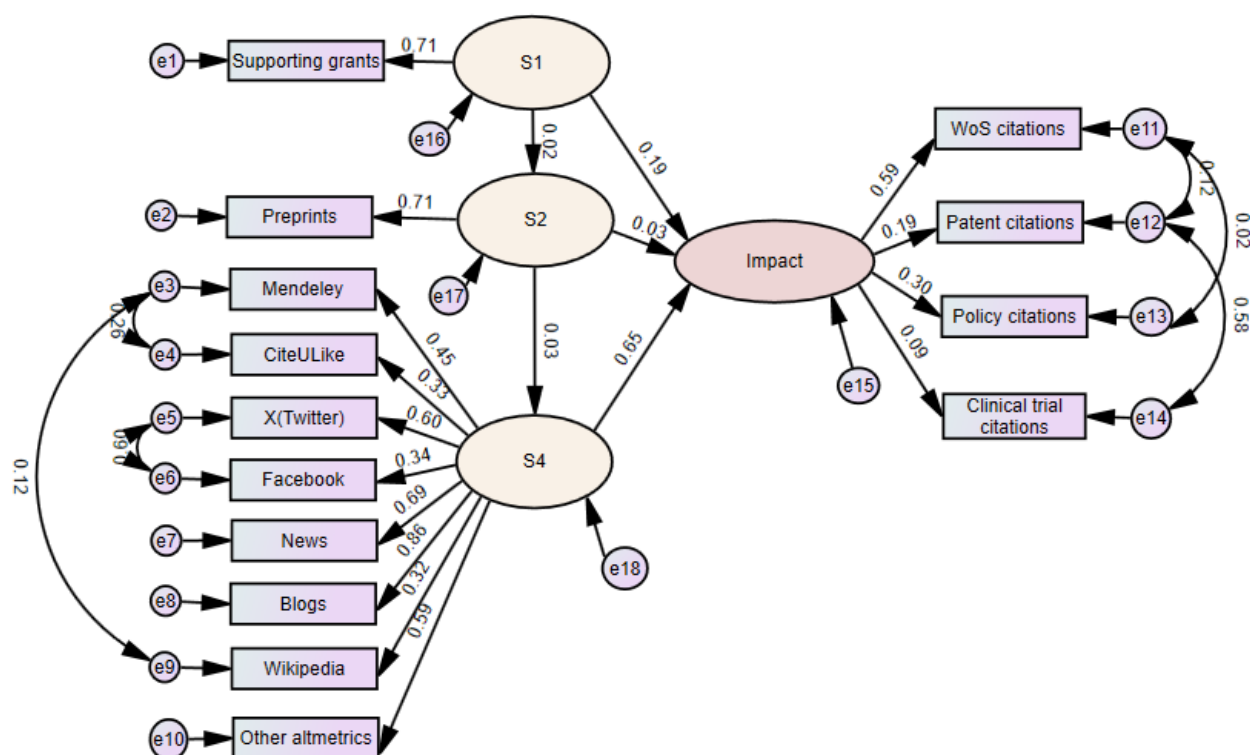


Figure 7.3 The improved model for the impact of research papers

The goodness of fit index (GFI) in the above structural model is 0.908, and the adjusted goodness of fit index (AGFI) is 0.859, which is excellent. GFI > 0.90 suggests a satisfactory fit

(Hair et al., 2010), and AGFI > 0.8 suggest a good fit (Akinyode, 2016). In addition, an adequate-fitting model is accepted if the Amos computed value of the standardised root mean square residual (SRMR) is between 0.05 and 0.08 (Hair et al., 2010). In our model, the standardised RMR is 0.0767, which falls within the acceptable range. Accordingly, the improved model in Figure 7.3 satisfies the goodness of fit criteria to measure the studying constructs.

The Amos provides two types of estimates: unstandardised regression coefficients and standardised regression coefficients. The former refers to the raw coefficients that quantify the relationship between variables in the model without being transformed or scaled by the standard deviations of the variables involved. The latter are scaled to have a unit variance (Arbuckle, 2022). Both regression coefficients reflect the strength and direction of the relationships between variables, and we can use these values to assess the hypothesised relationships. It is essential to observe the squared multiple correlations, which are independent of units of measurement like the standardised estimates. Each dependent variable has an estimated squared multiple correlation. The regression coefficients (β) of hypothesised relationships and squared multiple correlations are criteria for evaluating a structural model (AL-Fadhali, 2024).

Table 7.12 summarises the model's standardised regression coefficients (β), critical ratio (C.R.), p-value (P), and corresponding comments, indicating which hypotheses are supported or not. The *** in Table 7.12 represents a p-value smaller than the smallest value represented by the chosen decimal precision, which, in this case, is < 0.001, as reported in the Amos output (Kline, 2023). The first hypothesised relationship has a positive coefficient of 0.017, a critical ratio of 4.661 (greater than 1.96), and a p-value less than 0.001. It means that the regression weight for S1 in predicting S2 is significantly different from zero at the 0.001 level. Hence, such a hypothesised relationship is accepted. Bear in mind the testing procedures in SEM are 'reverse' to the basic philosophy of traditional statistical testing. According to Blunch (2013), the null hypothesis (H_0) in SEM posits that the model under consideration is valid, whereas

the alternative hypothesis (H_1) suggests otherwise. Therefore, SEM modelling should commence with meaningful models informed by prior research and well-established theoretical frameworks, and the process involves comparing models and evaluating their relative capacity to account for the observed data (Westland, 2015). Similarly, the other hypothesised relationships in Table 7.2 all have p-values < 0.001 and the conclusion is all hypothesised relationships are accepted.

Table 7.12 Standardised regression coefficients and significance level

Hypothesised relationship			Standardised Estimates	Critical Ratio	P-value	Comments
S2	<---	S1	.017	4.661	***	Accepted
S4	<---	S2	.028	9.717	***	Accepted
Impact	<---	S1	.188	48.880	***	Accepted
Impact	<---	S2	.026	6.739	***	Accepted
Impact	<---	S4	.652	160.024	***	Accepted

Note(s): ***p < 0.001; S1 = the stage of knowledge generation; S2 = the stage of knowledge publishing; S4 = the stage of knowledge understanding and extensive dissemination; Impact = the impact of research papers

In evaluating the route coefficient, the highest coefficient is 0.652 for the path of S4 to Impact. Hence, the stage of knowledge understanding and extensive dissemination has the most significant effects on the final impact of research papers. The second largest coefficient is 0.188, which is on the path of S1 to Impact. In other words, the second significant construct that affects the impact of research papers is the stage of knowledge generation. Following is the path of S2 to S4, which has a coefficient of 0.028, much greater than the coefficient on the path of S1 to S2, which is 0.017, proving that there are cumulative effects from an early stage to a later stage. The path of S2 to Impact has a coefficient of 0.026, indicating that the stage of knowledge publishing has a low but significant effect on the ultimate impact of research papers.

Furthermore, the squared multiple correlation of the endogenous latent variable, Impact, is 0.463 in Amos output (see Table 7.13), which means S1, S2, and S4 have accounted for 46.3% of the variance in the impact of research papers. The estimated squared multiple correlations on S1 is 0.000, on S2 is 0.000, and on S4 is 0.001, which is relatively small. According to Kline

(2023), a squared multiple correlation value of 0.25 or higher is often considered substantial in Social Science studies, as effects in this field tend to be more modest due to the complexity of human behaviour. Values of 0.10 – 0.25 are considered moderate, while values below 0.10 are considered small but not necessarily trivial. Therefore, the squared multiple correlation value for the impact of research papers signifies that the constructed factors to interpret impact are convictive. Our structural model has good confirmatory and explanatory power.

Table 7.13 The squared multiple correlations on latent variables

Variable	Estimate
S1	.000
S2	.000
S4	.001
Impact	.463

The above results firmly confirmed that all the hypothesised relationships we built in this chapter (see Section 7.4.1) are accepted. In particular, the stage of knowledge understanding and extensive dissemination (S4) is the most influential factor in the ultimate impact. Such results proved the feasibility and significance of the conceptual model on cause-effect and evidence chains developed in Chapter 6. With sufficient evidence data, we can evaluate the impact from a new perspective, making the evaluation results more diversified, accurate and reliable.

7.5.2 Disciplinary differences

Next, the differences in constructs of the impact of research papers between different category groups were studied. For ease of comparison, each category group's structural model and initial parameters are consistent with the model of all the data in the previous section. The iteration limit is 500 remains. The structural model of each category group, including standardised regression coefficients for the model constructs, was examined. Specifically, fitness indices were evaluated and found to meet the acceptable thresholds, indicating the model's alignment with the data and the accuracy of the hypothesised relationships.

Table 7.14 summarises the construct coefficients and their p-values to observe the significance. In the Biology & Biochemistry group, β_1 is significant at 0.05 level as p-value = .022 < .05, and all the other β s are significant at .001 level since their p-values are less than .001 shown in the Amos output. In the Chemistry group, all coefficients of β s are significant at the .001 level. In the Clinical Medicine group, β_1 is significant at the 0.05 level as p-value = .028 < .05, while β_2 and β_4 are not significant as their p-values > .05. β_3 and β_5 are significant at .001 level in the structural model of Clinical Medicine. For the Environment/Ecology group, all β s are significant as their p-values are less than .05 or .001. In the Engineering group, the p-value of β_2 is .289 > 0.05; thus, the corresponding path is insignificant. β_3 and β_5 are significant at .001 level. The p-value of β_1 is .037 < 0.05, and the p-value of β_4 is .023 < 0.05; therefore, the corresponding paths are significant in the Engineering subject. Additionally, as we analysed in the last section, all β s are significant for the all data.

Table 7.14 Standardised regression coefficients (β) and significance level for different group

Biology & Biochemistry		Chemistry		Clinical Medicine		Environment/Ecology		Engineering		All data	
Estimate	P	Estimate	P	Estimate	P	Estimate	P	Estimate	P	Estimate	P
β_1	.019	.022	***	.008	.028	.021	.014	.018	.037	.017	***
β_2	.088	***	.549	***	.001	.848	***	.007	.289	.028	***
β_3	.666	***	.974	***	.626	***	.490	***	.642	.652	***
β_4	-.045	***	-.223	***	-.002	.816	.019	.012	.029	.026	***
β_5	.206	***	.093	***	.122	***	.164	***	.132	.188	***

The above results showed that two groups, Clinical Medicine and Engineering, have insignificant paths in their structural models. Clinical Medicine has β_2 , the coefficient on path S2 to S4 (see Figure 7.1), and β_4 , which is the coefficient on path S2 to Impact, are insignificant. It seems the publishing stage in Clinical Medicine has no effect on our studied construct. Tracking data found that this subject's preprint data, the indicator of the stage of publishing, only has a few data that account for 0.008% in the Clinical Medicine data sample; therefore, such a variable indeed has little influence, and the results are not surprising. Besides, the last insignificant path is S2 to S4 in the Engineering group, which implies the stage of publishing has no significant impact on the stage of knowledge understanding and extensive

dissemination. Since S1, S2, and S4 are all significant to the impact in Engineering, different stages of the impact generation process still directly affected the final impact of the research.

As shown in Table 7.14, coefficients of β s presented the strength of the path between latent variables. The higher the path coefficient, the more substantial the impact of the latent endogenous variable. All β_3 are the highest coefficients in each model. Therefore, the path of S4 to Impact is the most influential construct. In other words, the stage of knowledge understanding and extensive dissemination is the most significant factor in influencing the final impact, and all the category groups showed the same result.

It is noticeable that the path coefficients of β_4 in the Biology & Biochemistry and Chemistry are negative but significant, which means S2 has a reverse effect on Impact: when S2 decreases, the impact of research papers increases. Such results do not comply with common sense. The possible reason is that preprints in these two subjects do not positively affect the final impact. Research in the relevant literature has found that preprints represent a relatively small fraction of all biology papers (Abdill & Blekhman, 2019). It is argued that the number of preprints in bioRxiv surpassed the 12,000 threshold in early 2016 while less than 900 in the first year when bioRxiv was opened. The chemistry community is in a similar situation, where preprints for chemistry became popular about two decades after physics. Ciriminna and Pagliaro (2017) commented that Chemistry is among the last of the core natural sciences to embrace preprints.

Furthermore, Table 7.15 illustrates standardised regression coefficients of all indicators (λ s) on latent variables in our structural models, including each category group and all-data group. λ_{11} represents the factor loading of supporting grants on the stage of knowledge generation (S1). λ_{21} denotes the factor loading of preprints at the knowledge publishing stage (S2). λ_{31} to λ_{38} demonstrate the standardised regression coefficients of indicators for the stage of knowledge understanding and extensive dissemination (S4). λ_{41} to λ_{44} denote the standardised regression coefficients of four types of impact indicators.

Table 7.15 Standardised regression coefficients (λ) for different category group

	Biology & Biochemistry	Chemistry	Clinical Medicine	Environment /Ecology	Engineering	All data
λ_{11}	.707	.707	.707	.707	.707	.707
λ_{21}	.707	.707	.707	.707	.707	.707
λ_{31}	.519	.935	.507	.409	.265	.447
λ_{32}	.380	.195	.425	.304	.097	.335
λ_{33}	.608	.164	.649	.573	.719	.604
λ_{34}	.603	.182	.378	.433	.574	.336
λ_{35}	.698	.195	.715	.663	.537	.686
λ_{36}	.900	.259	.867	.782	.771	.864
λ_{37}	.306	.243	.384	.099	.459	.319
λ_{38}	.694	.121	.668	.379	.755	.591
λ_{41}	.667	.947	.674	.729	.328	.591
λ_{42}	.316	.135	.197	.228	.060	.188
λ_{43}	.165	.110	.391	.368	.422	.300
λ_{44}	.326	.131	.441	.208	.050	.091

From the results in Table 7.15, we can observe the following findings:

- It can be seen that all λ_{11} among category groups are the same because supporting grants are the sole indicator for S1, and we set the same default variance of errors as initial parameters. A similar situation occurs in the stage of knowledge publishing (S2), and we only have one indicator, preprint data, at this stage. One of our analysis's limitations is insufficient indicators for S1 and S2. In the knowledge generation and publishing stages, few indicators are available in practice, and we have high data quality requirements that restrict some potential data. When more indicators become available, we could reduce the initial parameter settings. Therefore, λ_{21} has the same standardised value for all studying groups.
- The values $\lambda_{31}, \dots, \lambda_{38}$ are the standardised regression coefficients of indicators for the stage of knowledge understanding and extensive dissemination (S4). In Biology & Biochemistry group, the highest factor is Blogs (λ_{36}), and then the following are News (λ_{35}), other altmetrics (λ_{38}), Twitter (λ_{33}) and Facebook (λ_{34}). All these factors are greater than 0.6. It seems that blogging is a popular way to promote research work in

biology. Biologists are very active in spreading new discoveries on social media platforms. In Chemistry, the situation is completely different. Chemists are keen to use Mendeley to save or store their interested research papers, as λ_{31} is the highest factor loading at this stage, which is 0.935. Then, the factor of Blogs is 0.259 and Wikipedia is 0.243. All social media factors are lower than 0.2. In Clinical Medicine, Blogs (λ_{36}) is the highest factor, which is over 0.8, then is the News (λ_{35}), which is 0.715. Other altmetrics (λ_{38}) and Twitter (λ_{33}) loadings are also high in Clinical Medicine, which is over 0.6. The factor loadings of Clinical Medicine are similar to those of Biology. For Engineering and Environment/Ecology subjects, Blogs (λ_{36}) is the highest factor as well, then Twitter (λ_{33}) and News (λ_{35}) are high factors for these two subjects. Mendeley and CiteULike in Engineering are both lower than those in Environment/Ecology, it seems the Engineering scholars are not keen to use reference management software.

- λ_{41} to λ_{44} denote the standardised regression coefficients of four indicators, WoS Citations, Patent citations, Policy document citations, and Clinical trial citations, for the impact of research papers. The four types of citations reflect the different dimensions of the impact. WoS citation (λ_{41}) which normally represented to the academic impact has the highest factor loading in all studying groups except Engineering. In Chemistry, the factor coefficient of WoS citations is exceptionally high at 0.947. Consequently, academic impact is the most significant dimension of impact in the Chemistry group.
- Clinical Medicine and Biology & Biochemistry both have the second highest factor loading on Clinical trial citations (λ_{44}). Obviously, these two subjects have many clinical trial studies. The factor loading of Patent citations (λ_{42}) in Biology & Biochemistry is higher than in other groups, which indicates that patent registration has an important impact on this subject. The environment/Ecology group has the second highest factor loading on Policy citations (λ_{43}), which means the research outcomes in this field may greatly affect policy documents for society.

- Engineering's highest impact factor is Policy document citations (λ_{43}), which is 0.422, and the second highest factor loading is WoS citations, which is 0.328. Loadings on Patent citations (λ_{42}) and Clinical trial citations (λ_{44}) are small in the Engineering group, which is less than 0.1. The squared multiple correlation of Impact in Engineering is 0.431, meaning 43.1% of the variance in the impact has been explained. In other words, 56.9% was not explained. It could suggest that some other areas of impact emphasised in the Engineering group are not included in this study. It might be the application impact, economic impact, or something else that is worth studying further in later research.

7.6 The evidence-chain-based research impact evaluation

The structural model results in previous sections showed that all stages of the impact generation process in our analysis significantly affect the comprehensive impact of research papers, and the stages follow a logical causal relationship in time. Therefore, we can build up a factor analysis model for the comprehensive impact of research papers and calculate the impact score for every publication.

Assume there are p variable ($X_1, X_2, \dots, X_p$) and m underlying factors ($F_1, F_2, \dots, F_m$). X_j is the variable that denoted in latent factors. Then, we have the mathematical model of factor analysis (Yong & Pearce, 2013) is:

$$X_j - \mu_j = l_{j1}F_1 + l_{j2}F_2 + \dots + l_{jm}F_m + e_j$$

Where $j = 1, 2, \dots, p$. (7.1)

In Equation 7.1, μ_j is the mean of X_j ; F_k ($k = 1, 2, \dots, m$) are common factors; l_{jm} ($j = 1, 2, \dots, p; m = 1, 2, \dots, m$) are factor loadings; e_j are error terms that assumed independent of F_k . We need to note that no means or intercepts of variables were estimated in our models, so the factor weights are calculated using the raw observed values. Since each structural model has four latent variables, the factor score weights are produced on every indicator for all the latent variables. Table 7.16 presents the factor score weights estimated by Amos on the structural model with all combined data.

The coefficient values of the indicators in S1 and S2 are minor but substantively affect the value in S4 and the comprehensive impact. Here, we mainly focus on the coefficient of each indicator to the impact. It is noticed that the preprints and blogs have high weight of factor scores. Preprints published before the formal publication, which indicate a longer time of exposure of research knowledge to accumulate impacts. The high impact research papers would started to be acknowledge since the first publishing time, such as preprints.

Table 7.16 Factor score weights for all data

	S1	S2	S4	Impact
Supporting grants	.496	.000	-.261	4.413
Preprints	.098	.500	2.250	19.303
Mendeley	.000	.000	.036	.020
CiteULike	-.003	.000	5.632	3.196
X (Twitter)	.000	.000	.414	.235
Facebook	-.002	.000	-.254	-.144
News	-.002	.000	3.869	2.196
Blogs	-.024	.000	46.331	26.297
Wikipedia	-.002	.000	4.817	2.734
Other altmetrics	-.004	.000	8.428	4.784
WoS citations	.001	.000	.033	.262
Patent citations	.001	.000	.078	.609
Policy document citations	.023	.000	1.509	11.841
Clinical trial citations	.001	.000	.063	.498

Regarding the high coefficient weight for blogs, this is an interesting finding and worth further exploring. Mao et al. (2014) found that relationships among blog users are typically unidirectional, unlike the bidirectional (friendship-based) connections often observed in other social networking platforms. Blogs function as a form of social media, and their impact can be measured based on the effectiveness of media dissemination. Blogs related to academic articles often use language that is easier for the general public to understand, serving as an effective medium for extending complex and obscure scientific knowledge from the academic sphere to society at large, reaching non-academic audiences. Moreover, researchers have noted that blogs discussing academic articles are more likely to reference publications from high-impact journals (Fausto et al., 2012). Shema et al. (2014) were

pioneers in investigating the relationship between journal citations and blog mentions, identifying positive correlations between the two. Additionally, a well-known phenomenon is the surge of blog posts appearing online after Nobel Prize announcements, aiming to explain the significance of the laureates' research to a broader audience. These examples highlight the influential role of blogs in disseminating high-impact academic content.

Using the coefficient values in Table 7.16, we can now calculate the factor score for the latent variable, the impact of research papers, by applying all the factors. The factor model equation for our comprehensive impact of research papers can be written as below:

$$\begin{aligned} X_{impact} = & 4.413F_{Supportin_grants} + 19.303F_{Preprints} + 0.02F_{Mendeley} + 3.196F_{CiteULike} + \\ & 0.235F_X - 0.144F_{Facebook} + 2.196F_{News} + 26.297F_{Blogs} + 2.734F_{Wikipedia} + \\ & 4.784F_{Other_altmetrics} + 0.262F_{WoS_citations} + 0.609F_{Patent_citations} + \\ & 11.841F_{Policy_citations} + 0.498F_{Clinical_trial_citations} \end{aligned} \quad (7.2)$$

Accordingly, every research paper could have an impact score based on its available (evidence) data. We name this score the impact score for research papers. It will be stimulating to find out the most influential research work published in the year 2013. To this end, we analysed and ranked the top 10 highest-impact research papers based on their impact scores across all subject areas. The detailed results are presented in Table 7.17.

The highest impact research paper in 2013 was from the journal JAMA, under the category Medicine, general & internal, within our third category group, the Clinical Medicine group. This paper achieved an impressive impact score of 5211.35. Highlighting its influence, it received 16,739 WoS citations, 39 clinical trial citations, 16 policy document citations, and 178 Twitter mentions. This research work is very influential in every aspect of society. The second-highest impact paper, published in Cell, falls within the “Cell Biology” category, part of the “Biology & Biochemistry” group. This paper garnered 10,168 WoS citations, 35 patent citations, 779 Twitter mentions and 167 News mentions, underscoring its significant societal

and academic impact. Among the top 10 highest-impact papers, eight are within the “Biology & Biochemistry” group, while the remaining two belong to the “Clinical Medicine” group. Notably, all top 10 papers stem from life science disciplines, highlighting their broad-reaching influence across academia and society.

The third-highest impact papers and the sixth, eighth, ninth and tenth-highest impact papers only have about 4 thousand WoS citations or less. However, they have high news, Twitter, Facebook, and other altmetrics mentions, meaning they have extensive societal attention and have brought these papers to very high rankings. In analysing the impact composition, the stage of knowledge understanding and extensive dissemination (S4) emerge as the most significant factors influencing the overall impact in our structural models. Consequently, the indicators associated with this stage generally carry higher weights. This perspective merits further discussion.

Societal influence is a critical dimension of impact, indeed. The current results reflect weights derived from our collected data and our chosen modelling approach. They are reasonable but may be biased. If there is a consensus that academic influence should be prioritised, the weights can be adjusted by modifying the parameters, emphasising academic impact while de-emphasising societal effects. In practical evaluations, such weights could be reasonably adjusted through expert discussions.

We must also remark that the impact components vary significantly across disciplines, as discussed in Section 7.5.2. Therefore, it is essential to account for disciplinary distinctions when comparing the impact of research publications. Comparisons within the same discipline can enhance fairness and reduce bias. For example, this study has established impact structure models for five major category groups, allowing the factor weights for each research area's structure to be easily determined. When evaluating the specific research area or discipline, the structural model of impact for this field must be built first to obtain the most reasonable weights for impact value calculations.

The evidence-chain-based evaluation method can be applied systematically to assess the impact of specific research publications, journals, categories, or research areas. Next, we will carry out an evaluation of the impact of the journal using our proposed approach. It is a further application of the evidence-chain-based evaluation method. Given our familiarity with the study area, we conducted a journal impact ranking focused on the operations research and management science journals. Table 7.18 demonstrates the top 20 highest-impact journals with papers published in 2013.

Table 7.17 Top 10 highest-impact research papers published in 2013

Rank	Impact Score	Title of paper	Journal name	Year	Category	Category Group	Metrics highlights
1	5211.35	World Medical Association Declaration of Helsinki: Ethical Principles for Medical Research Involving Human Subjects	JAMA	2013	Medicine, general & internal	3	Publication citations:16739; Policy document citations:16; Clinical trial citations:39; Twitter:718;
2	4976.05	The Hallmarks of Aging	Cell	2013	Cell Biology	1	Publication citations:10168; Patent citation:35; Twitter:779; News:167
3	3887.82	DNA methylation age of human tissues and cell types	Genome Biology	2013	Biotechnology & Applied Microbiology	1	Publication citations:4866; Patent citation:15 Twitter: 237; News:164; Other altmetrics:186 (involved Google+ 172)
4	3306.14	TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions	Genome Biology	2013	Biotechnology & Applied Microbiology	1	Publication citations:11150; Patent citation:58
5	3170.71	Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal	Science Signaling	2013	Cell Biology	1	Publication citations:11265; Patent citation:41
6	2350.06	Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis	Nature Medicine	2013	Cell Biology	1	Publication citations:3196; Patent citation:42; Twitter: 424; Facebook:91; News:61;
7	2349.99	Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences	Nature Biotechnology	2013	Biotechnology & Applied Microbiology	1	Publication citations:7346; Patent citation:63
8	2204.89	An Introduction to Social Media for Scientists	PLOS Biology	2013	Cell Biology	1	Twitter:1974; Facebook:106; Blogs:47; Other altmetrics:63
9	2072.20	Enough is enough: Stop wasting money on vitamin and mineral supplements.	Annals of Internal Medicine	2013	Medicine, general & internal	3	Twitter:1210; Facebook:499; News:134; Blogs:45; Other altmetrics:66
10	2012.25	Repurposing CRISPR as an RNA-Guided Platform for Sequence-Specific Control of Gene Expression	Cell	2013	Cell Biology	1	Publication citations:3758; Patent citation:690;

Table 7.18 Top 20 highest-impact journals published in 2013 in Operations Research & Management Science

Rank	Impact Score of Journal	Journal name	Year	Number of papers	Total of WoS citations today	JIF2023 rank	AJG2024 rank
1	24.48	Management Science	2013	181	47,274	20	4*
2	12.43	Journal of Operations Management	2013	39	11,113	10	4*
3	9.31	Technovation	2013	53	11,239	2	3
4	7.75	Manufacturing & Service Operations Management	2013	54	6,443	16	3
5	7.27	Decision Support Systems	2013	217	13,405	8	3
6	6.53	Transportation Research Part B Methodological	2013	131	17,027	14	4
7	6.10	Omega	2013	97	13,376	8	3
8	6.04	Safety Science	2013	226	20,058	19	3
9	5.17	Transportation Research Part E Logistics and Transportation Review	2013	104	17,564	5	3
10	5.14	Production and Operations Management	2013	108	12,030	16	4
11	4.55	Operations Research	2013	101	15,236	34	4*
12	4.53	European Journal of Operational Research	2013	507	62,282	13	4
13	4.34	Mathematical Programming Computation	2013	14	1,695	23	N/A
14	4.30	Mathematical Programming	2013	113	12,372	34	4
15	4.26	Reliability Engineering & System Safety	2013	216	33,263	4	3
16	4.18	Systems & Control Letters	2013	151	8,337	38	N/A
17	3.93	International Journal of Production Economics	2013	387	33,962	3	3
18	3.78	Expert Systems with Applications	2013	750	81,232	6	1
19	3.56	Fuzzy Optimization and Decision Making	2013	24	1,449	16	3
20	3.56	Computers & Operations Research	2013	295	16,520	24	3

We applied the factor score weights of each indicator within the Engineering group to calculate the impact scores, as the operations research and management science category falls under the Engineering group according to the WoS category list. These factor score weights were computed using Amos based on the structural model of the impact of the research paper in the engineering group. The results are shown in Table 7.19.

Table 7.19 Factor Score Weights for the Engineering group

	S1	S2	S4	Impact
Supporting grants	.499	.000	-.068	2.116
Preprints	.067	.500	.119	7.515
Mendeley	.000	.000	.008	.004
CiteULike	.000	.000	1.855	.988
X (Twitter)	.000	.000	.841	.448
Facebook	-.002	.000	7.607	4.050
News	-.001	.000	3.320	1.767
Blogs	-.008	.000	30.755	16.373
Wikipedia	-.002	.000	7.639	4.067
Other altmetrics	-.008	.000	30.088	16.019
WoS citations	.000	.000	.009	.053
Patent citations	.000	.000	-.011	-.064
Policy document citations	.035	.001	2.232	13.161
Clinical trial citations	.130	.002	8.289	48.874

Subsequently, the impact score of each publication in this group was calculated. Then, the impact scores of all articles from each journal within the operations research and management science category were aggregated to compute the average value of impact scores for the journal. We named this score as the Journal Impact Score.

The ranking data was selected from papers published in 2013, as we aimed to examine the cumulative impact of research output over a more extended period, such as ten years. All evidence data (indicators) were downloaded in 2024, which represents the impact value as of the data collection date. The strength of our ranking lies in its ability to identify research contributions with enduring influence, providing a more comprehensive measure of impact.

Our journal ranking is based on the cumulative impact of each paper published in the respective journals.

Moreover, the high-ranked journals on our list indicate that the research papers in these journals continued to exert significant influence over the subsequent decade, accumulating substantial impact across academic and societal domains. Some research works exhibit high citation rates shortly after being published but experience diminishing influence over time, reflecting a limited long-term impact. In contrast, research outcomes with groundbreaking significance often require a longer period to accumulate influence and continue to expand their impact across various aspects of society.

The Journal Impact Factor (JIF) remains the most traditional metric in journal rankings. However, since JIF only accounts for publication citations and reflects academic impact alone, it increasingly fails to represent the overall influence of research outcomes. The Academic Journal Guide (AJG) is another widely used journal ranking system, particularly in business and management disciplines. The AJG incorporates qualitative information alongside established quantitative metrics and is well-recognized by scholars in business studies. It classifies journals into five ratings: 4* (world-leading journals), 4 (top journals in the field), 3 (highly regarded journals), 2 (journals publishing original research of an acceptable standard), and 1 (journals publishing research of a recognized standard) (CABS, October 1, 2024).

To validate our ranking results, we compared our journal impact rankings with those from JIF 2023 and AJG 2024, as shown in Table 7.18. The results revealed significant differences between our rankings and those from JIF, but some alignment with AJG. Management Science journal was identified as the highest-impact journal in our rankings, followed by the Journal of Operations Management. This result aligns with AJG's rankings, affirming that these journals are indeed world-leading in terms of influence. It is important to note that the category classification in AJG differs from the WoS category list, so some high-rated journals

in AJG were not included in our list. Since the rank of AJG is more recognised by business studies, our ranking results showed a similar ranking trend to that of AJG.

Furthermore, to seek peers' opinions, a questionnaire with the impact ranks of journals in Operations Research & Management Science (as demonstrated in Table 7.18) was conducted. The questionnaire was answered by various scholars and experts in operations research and management science studies, including professors, lecturers, associate researchers, and PhD students. Their answers significantly revealed similarities with our ranking results, which proved our rank list is more in line with the scientific community's perception of the journal's impact.

It is worth highlighting that the impact score we calculated is based on the full chain data on the causal chains of the impact generation process, which reflects the comprehensive influence of research. High-quality evidence data makes the results of impact traceable and more reliable. The evidence-chain-based evaluation approach represents a new direction for the impact evaluation field. This approach defines research impact as multidimensional, encompassing academic and societal aspects. It is supported by evidence and data that demonstrate its dissemination, covering the entire process of research impact generation. Therefore, our impact aggregation results are more comprehensive, reliable, and better suited to the evolving demands of modern research management.

7.7 Summary

This chapter used real data to aggregate comprehensive impact for research papers and calculated their impact score for impact evaluation. The findings in this chapter are highlighted below.

Firstly, we collected evidence data on the impact generation process for five disciplinary groups. As these data served as indicators for the impact generation process for research papers and would be calculated in the impact aggregation, ensuring data quality is of paramount importance. We systematically screened the potential data available at each stage

of the impact generation process and thoroughly evaluated the databases providing these data. All databases for data collection were carefully selected from the most representative ones within their category, and quality was verified by clarifying the origins of data sources and the breadth and comprehensiveness of the database's publication coverage.

Secondly, we applied the Structural Equation Modeling approach to constructing the impact aggregation model, and the hypothesised paths in the structural model were based on the cause-effect chains of research impact built up in Chapter 6. We proposed five hypothesised relationships in the impact generation process, and the model-fitting results confirmed that all the hypothesised relationships were accepted. That is, each stage of the research impact generation process has a cumulative effect on comprehensive research impact, and there is a logical causal relationship between the stages. The results of the structural models confirmed the feasibility and significance of using cause-effect chains of research impact generation to build impact evaluation.

Additionally, we compared the differences in the structural models of impact across disciplines, where the coefficients of the indicators and impact dimensions differed. We also explored the differences in the regression coefficients across the five disciplinary groups.

Then, we used all factor weights estimated in the structural models of impact to aggregate impact scores for research papers. Based on available evidence data, every research paper could have an impact score. We calculated the highest-impact research papers published in 2013 and found that all top 10 research works were from life science disciplines.

Furthermore, we also evaluated some journals' impact. We chose parts of operations research and management science journals as being more familiar with the research field and ranked the journal impact according to their impact score. Our rank results were approaching the AJG's ranking list, which is well-recognized in business studies. The peer review opinions from experts in operations research and management science studies also agreed with our

ranking results regarding their perception of the journal's impact. Our ranking excelled in identifying research papers with widespread and enduring impact. Some papers experience a decline in citation frequency over time, indicating their limited influence. In contrast, groundbreaking scientific contributions tend to expand their impact constantly over an extended period, resonating across various aspects of society.

Finally, the impact scores we aggregated were based on full-chain data on the cause-effect chains of impact generation, which reflects the comprehensive impact. By incorporating every stage of the impact generation process into the evaluation, this approach captured the influence of research papers on both academic and broader dimensions. The evidence-chain-based impact evaluation approach initiated a novel perspective for assessing impact, resulting in outcomes that are more diverse, accurate, and reliable.

Chapter 8 – Conclusions and future research

8.1 Conclusions

This thesis aimed to develop evidence-based quantitative evaluating methods for research impact to increase the reliability of assessment results and answered the following three research questions:

- 1) How can we identify and select suitable altmetric data for evaluating the impact of research publications?
- 2) How can we aggregate those altmetric and scientometric data based on a cause-effect analysis to develop more reliable metrics for the impact of research publications?
- 3) How can we apply the above approaches in real applications?

To answer the above research questions, Chapters 2 and 3 carried out a comprehensive literature review on research evaluation, impact evaluation, and traditional and alternative metrics for research impact, including their concepts, definitions, features, methodologies, and research status.

Then, Chapter 4 conducted empirical research on two mainstream altmetrics platforms, Altmetric and PlumX, to investigate the altmetric data quality and suggested how to use these platforms' altmetric data in practice. This chapter involved a detailed analysis of these platforms' altmetric data quality, including data coverage, disciplinary differences, correlation analysis, data consistency, and the reliability of the data sources. Findings in this chapter provided a thorough understanding of the altmetric data sources and platforms. The first research question has been answered properly in this chapter.

In Chapter 5, we proposed an evidence-chain-based quantitative evaluating approach for scholarly impact. We first discriminated three core relationships: causality, cause-effect chains, and evidence chains for academic impact evaluation. Then, we suggested a method on how to build an evidence chain in three steps. A primary evaluation model was proposed. The case study was carried out by calculating the likelihood of occurrence of the evidence

chain for the two targeting journals. The results showed that the journal with a significantly higher likelihood of occurrence of the evidence chain (although its impact factor is lower), which corresponds to the favourable cause-effect chain of “read-understood-disseminated-cited”, is the high-impact journal recognised in the scientific community’s common views.

We recognised citation as the “effect”, citation motivations as multiple “causes”, and the citing process as “cause-effect chains” to analyse the evidence chains in the specific cause-effect chain, “read-understood-disseminated-cited,” to evaluate the academic impact in Chapter 5. Then, we developed our research to a further step in Chapter 6, to construct the cause-effect chains of a comprehensive research impact in order to build evidence-chain-based methodology for comprehensive impact evaluation.

Chapter 6 analysed the impact generation cause-effect chains, which put together the academic and broader impacts (economy, society, culture, public policy or services, health, and more) of the research output as a final effect, and illustrated the cause-effect chains of the impact generation process. We identified five logical stages of generating impact for research publications: knowledge generation (S1), knowledge publishing (S2), the first wave of knowledge dissemination (S3), knowledge understanding and extensive dissemination (S4), and knowledge transformation and regeneration into new outputs (S5). The original research output’s extensive impact arises across the different paths of the whole process. The analysis of key stakeholders, key actions, and key results of each stage was carried out and used to extract quantitative evaluation indicators. The more indicators, the better. However, we need to note that the indicators involved various types of data and data sources, and not all kinds of data are collectable or of good quality for use. Using the data collected at each stage to evaluate the impact could solve the attribution and reliability problem and make evidence data causally connected. Both Chapters 5 and 6 answered the second research question from different perspectives.

Chapter 7 used real data to aggregate comprehensive impact for research papers and calculated their impact score for impact evaluation. We collected a range of evidence data as indicators of the impact generation process, including grant data, preprint data, Mendeley reader data, CiteULike data, X(Twitter) mentions, Facebook mentions, News mentions, Blogs mentions, Wikipedia mentions, publication citations from WoS, patent citations from Lens, policy document citations from Overton and clinical trial citations from Dimensions. Five major category group's journal papers' data were gathered. All databases for specific data collection were carefully selected from the most representative ones within their category, and quality was verified by small sample tests, clarifying the origins of data sources and the breadth and comprehensiveness of the database's publication coverage.

The Structural Equation Modeling approach was applied in Chapter 7 to construct the research impact aggregation models. The structural models' fitting results verified the feasibility and significance of the paths we built based on the cause-effect chains of impact developed in Chapter 6. Consequently, we built different structural models for impact aggregation by discipline to better evaluate the research publications. Finally, all observed publications' impact scores were calculated using the factor weights from disciplinary structural models for research impact. We ranked the Top 10 highest-impact research papers published in 2013 and then conducted a rank list of the 20 highest-impact journals in 2013 in the Operations Research & Management Science category. Rank lists were examples. Our proposed methodology, based on cause-effects and evidence chains for impact evaluation, could systematically assess the impact of research publications from different levels and scopes, such as journals, categories, or research areas. The third research question has been answered thoroughly in Chapter 7.

Finally, there are six main contributions to this thesis:

- This thesis introduced an innovative quantitative evaluation method based on evidence-chains to assess the impact of research publications. The proposed approach effectively evaluates academic contributions and identifies journals with more

authentic influence. By addressing the limitations of traditional results-driven evaluation systems, this method delivers more reliable and insightful assessment outcomes.

- We further explored the evaluation methods for a comprehensive impact of research publications using evidence chains. The conceptual model introduces a clear and logical framework for analysing research impact using cause-effect chains, surpassing the previous reliance on output-based data alone. Collecting and utilizing data at each stage of the process resolves attribution and consistency issues, enabling robust and reliable assessments.
- The final case study evaluated the comprehensive impact of research publications. It successfully utilised the conceptual model of the impact generation process to construct impact evaluation models using a structural equation modelling approach and aggregated the impact scores for publications. The newly established quantitative evaluation methodology for research impact could support various purposes of assessing requirements, including the publication and journal's impact within required categories, disciplines, and research areas.
- Our evidence-chain-based evaluation method is a full-chain evaluation that is logically connected with causal chains of the impact generation process. This is the first time in the literature that a full chain of evidence from the research knowledge generation to impact generation is applied to evaluate impact. By incorporating every stage of the impact generation process into the evaluation, this approach can capture the influence of research papers on academic and broader dimensions, which reflects the comprehensive impact. Our method for research evaluation is a new way of thinking and provides an innovative insight into the impact evaluation system.
- The empirical study on altmetrics platforms, Altmetric and PlumX, involved a detailed analysis of altmetric data coverage, disciplinary differences, data consistency, and the reliability of the data sources, providing a thorough understanding of the altmetric data sources and platforms. Findings and suggestions in the study provided a valuable reference for subsequent use of altmetrics in the research.

- Existing literature indicates that Altmetrics has certain limitations as a direct tool for research impact evaluation. These metrics primarily reflect public interest and preferences, making them less suitable for directly measuring impact, such as social impact. However, this thesis uses altmetric data to extract indicators, recognising its significant role in impact measurement. It offers a new perspective on the appropriate and effective use of altmetric data in the research evaluation.

8.2 Research limitations

Complementary to the challenges discussed in each chapter, this thesis has the following research limitations.

First, our research impact generation process only summarised the most common conditions. More impact-generating paths should exist, but they have yet to be demonstrated. Besides, we cannot collect all the path information due to the current data tracing capability.

Second, our evidence-chain-based evaluation method has high requirements for data, especially when evaluating comprehensive impact. The limitation of data is that, on the one hand, there is limited available data on the process of generating impact. On the other hand, measurement data that can reflect broader impacts, especially long-term impact, is limited and not easy to obtain. When data is limited, however, it is also possible to consider building a model on a smaller scale, such as using data from an entire journal.

Third, our proposed methods also rely on the quality of evidence data. Data quality depends on the coverage and counting methods of data providers. The more reliable the data sources, the more accurate the impact evaluation model will be.

When data collection and processing technology develops in the future, these problems are expected to be solved at the technical level, and the evidence-chain-based impact evaluating methods for research outputs will be further perfected.

8.3 Future research work

Deriving from the literature review, findings in this thesis and the research limitations, several promising research areas and potential topics warrant further research.

Firstly, this thesis primarily focused on scientific research publications as the evaluation object to develop quantitative evaluation methods for impact. However, the evaluation objectives can be expanded to research institutions, individual researchers, and more. Such expansions would require modifications in data collection processes and evaluation methodologies, which are worth further study.

Secondly, when abundant data is available, such as click time, download time, or citation time for a publication, one can build a dynamic Bayesian Network to demonstrate the research impact generation process. However, data collection may be more complicated and challenging.

Thirdly, as more scientific research data are opened, different forms of data are gradually emerging. However, the validations of some corresponding databases are lacking. For example, there are few in-depth studies on policy document citation data. Therefore, a large-scale data analysis among different data providers is needed.

References

- Abdill, R. J., & Blekhan, R. (2019). Tracking the popularity and outcomes of all bioRxiv preprints. *Elife*, 8, e45133.
- Abi Younes, G., & de Rassenfosse, G. (2024). Replicable Patent Indicators Using the Google Patents Public Datasets. *Australian Economic Review*, 57(1), 102-113.
- About PlumX Metrics. Available online: <https://plumanalytics.com/learn/about-metrics/mention-metrics/> (accessed on 1 Aug 2023)
- About The Lens. (2024). *Mapping PORTs*. Retrieved June 9, 2024, from <https://about.lens.org/>
- Abramo, G. & D'Angelo, C. (2011). Evaluating research: from informed peer review to bibliometrics. *Scientometrics*, 87(3), 499-514.
- Abramo, G. (2018). Revisiting the scientometric conceptualization of impact and its measurement. *Journal of Informetrics*, 12(3), 590–597. <https://doi.org/10.1016/j.joi.2018.05.001>
- ACUMEN Consortium. (2014). Guidelines for good evaluation practice with the ACUMEN portfolio. In European commission 7th framework program
- Adie, E. (2014). Taking the alternative mainstream. *El Profesional de La Informacion*, 23(4), 349–351. <https://doi.org/10.3145/epi.2014.jul.01>
- Adie, E., & Roe, W. (2013). Altmeter: enriching scholarly content with article-level discussion and metrics. *Learned Publishing*, 26(1), 11–17. <https://doi.org/10.1087/20130103>
- ADLER, R., EWING, J., & TAYLOR, P. (2008). Citation Statistics. IMU-ICIAM-IMS.
- AGARWAL A. and AGARWAL R. (2007): The vulnerability and limitations of impact factor in evaluating quality, *The Royal Society of Medicine*, 100, 354-355.
- Akinyode, B. F. (2016). The Use of Structural Equation Modelling (SEM) in Built Environment Disciplines. *Research on Humanities and Social Sciences*, 6(6), 109-120.
- AKSNES, D. W. (2006). Citation Rates and Perceptions of Scientific Contribution. *Journal of the American Society for Information Science and Technology*, 57, 169-185.
- AL-Fadhali, N. (2024). An AMOS-SEM approach to evaluating stakeholders' influence on construction project delivery performance. *Engineering, Construction and Architectural Management*, 31(2), 638-661.
- Alhoori, H., & Furuta, R. (2014). Do altmetrics follow the crowd or does the crowd follow altmetrics? *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, 375–378. <https://doi.org/10.1109/JCDL.2014.6970193>
- Allen, P. (2016). PlumX's Facebook Altmeterics – Measure Up! Available online: <https://plumanalytics.com/plumx-facebook-altmetrics-measure-up/> (accessed on 1 Aug 2023).
- Allen, P. (2017). Worldwide news coverage in PlumX. Available online: <https://plumanalytics.com/world-wide-news-coverage-plumx/> (accessed on 1 Aug 2023).

- Araujo, A. C., Nascimento, D. P., Gonzalez, G. Z., Maher, C. G., & Costa, L. O. P. (2018). Impact of low back pain clinical trials measured by the altmetric score: Cross-sectional study. *Journal of Medical Internet Research*, 20(4), 1–9. <https://doi.org/10.2196/jmir.9368>
- Arbuckle, J. L. (2022). IBM® SPSS® Amos™ 29 User's Guide. *Amos Development Corporation*.
- Arnold, D. (2009). Integrity under attack: the state of scholarly publishing. *SIAM news*, 42(10), pp.2-3.
- Artusi, R., Verderio, P., & Marubini, E. (2002). Bravais-Pearson and Spearman correlation coefficients: meaning, test of hypothesis and confidence interval. *The International journal of biological markers*, 17(2), 148-151.
- arXiv (2024). *Submission Guidelines*. ArXiv Info. Retrieved June 30, 2024, from <https://info.arxiv.org/help/submit/index.html>
- Aung, H. H., Zheng, H., Erdt, M., Aw, A. S., Sin, S. C. J., & Theng, Y. L. (2019). Investigating familiarity and usage of traditional metrics and altmetrics. *Journal of the Association for Information Science and Technology*, 70(8), 872–887. <https://doi.org/10.1002/asi.24162>
- Award Search (2024). NSF Funding Search website, U.S. National Science Foundation. Retrieved May 10, 2024, from <https://new.nsf.gov/funding>
- BADOR, P. & LAFOUGE, T. (2010). Comparative analysis between impact factor and h-index for pharmacology and psychiatry journals. *Scientometrics*, 84, 65-79.
- Banshal, S. K., Singh, V. K., Kaderye, G., Muhuri, P. K., & Sánchez, B. P. (2018). An altmetric analysis of scholarly articles from India. *Journal of Intelligent and Fuzzy Systems*, 34(5), 3111–3118. <https://doi.org/10.3233/JIFS-169495>
- Banshal, S. K., Singh, V. K., Muhuri, P. K., & Mayr, P. (2019). Disciplinary Variations in Altmetric Coverage of Scholarly Articles. *17th International Conference on Scientometrics and Informetrics, ISSI 2019 - Proceedings*, 2, 1870–1881.
- Barbic, D., Tubman, M., Lam, H., & Barbic, S. (2016). An Analysis of Altmetrics in Emergency Medicine. *Academic Emergency Medicine*, 23(3), 251–265. <https://doi.org/10.1111/acem.12898>
- Bar-Ilan, J., Haustein, S., Milojević, S., Peters, I., & Wolfram, D. (2018). Peer review, bibliometrics and altmetrics - Do we need them all? *Proceedings of the Association for Information Science and Technology*, 55(1), 653–656. <https://doi.org/10.1002/pa2.2018.14505501073>
- Barnes, C. (2015). The Use of Altmetrics as a Tool for Measuring Research Impact. *Australian Academic and Research Libraries*, 46(2), 121–134. <https://doi.org/10.1080/00048623.2014.1003174>
- Barthel, S., Tönnies, S., Köhncke, B., Siehndel, P., & Balke, W. T. (2015). What does Twitter Measure?: Influence of Diverse User Groups in Altmetrics. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, 2015-June*, 119–128. <https://doi.org/10.1145/2756406.2756913>
- Barton, J. H. (2004). Issues posed by a world patent system. *Journal of International Economic Law*, 7(2), 341-357.
- Baas, J., Schotten, M., Plume, A., Côté, G., & Karimi, R. (2020). Scopus as a curated, high-quality bibliometric

- data source for academic research in quantitative science studies. *Quantitative science studies*, 1(1), 377-386.
- Basson, I., Simard, M.-A., Ouangré, Z. A., Sugimoto, C. R., & Larivière, V. (2022). The effect of data sources on the measurement of open access: A comparison of Dimensions and the Web of Science. *PLOS ONE*, 17(3), e0265545. <https://doi.org/10.1371/journal.pone.0265545>
- Bayley, J. E., & Phipps, D. (2019). Building the concept of research impact literacy. *Evidence & Policy*, 15(4), 597-606.
- Benito, M., Gil, P., & Romera, R. (2019). Funding, is it key for standing out in the university rankings?. *Scientometrics*, 121, 771-792.
- Benneworth, P., & Olmos-Peñuela, J. (2024). An openness framework for ex ante evaluation of societal impact of research. *Research Evaluation*, 33, rvac023.
- Bensman, S. J. (2007). Garfield and the impact factor. *Annual review of information science and technology*, 41(1), 93-155.
- Berg, E. J., & Ashurst, J. (2019). Patterns of recent National Institutes of Health (NIH) funding in general surgery: analysis using the NIH RePORTER system. *Cureus*, 11(6).
- Bergstrom, C. (2007) Eigenfactor: Measuring the value and prestige of scholarly journals. *College & Research Libraries News*, 68(5), pp.314-316.
- Bergstrom, C., West, J., & Wiseman, M. (2008). The Eigenfactor™ Metrics. *Journal of Neuroscience*, 28(45), pp.11433-11434.
- Berkes, F. (2009). Evolution of co-management: Role of knowledge generation, bridging organizations and social learning. In *Journal of Environmental Management* (Vol. 90, Issue 5, pp. 1692–1702). Academic Press. <https://doi.org/10.1016/j.jenvman.2008.12.001>
- Bex, F. J. (2011). *Arguments, stories and criminal evidence: A formal hybrid theory* (Vol. 92). Springer Science & Business Media.
- Bhaskar, R. (1997). On the ontological status of ideas. *Journal for the Theory of Social Behaviour*, 27(2-3), 139-147.
- Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, 1(1), 363–376. https://doi.org/10.1162/qss_a_00018
- Blair, R. C., & Higgins, J. J. (1985). Comparison of the power of the paired samples t test to that of Wilcoxon's signed-ranks test under various population shapes. *Psychological Bulletin*, 97(1), 119–128. <https://doi.org/10.1037/0033-2909.97.1.119>
- Blunch, N. J. (2013). Introduction to structural equation modeling using IBM SPSS and AMOS. *London: Sage*.
- BOLLEN J., DE SOMPEL H., SMITH J. and LUCE R. (2005): Toward alternative metrics of journal impact: a comparison of download and citation data, *Information Processing and Management*, 41, 1419-1440
- BOLLEN J., VAN DE SEMPEL H. and HAGBERG E. (2009): A principal component analysis of 39 scientific impact

measures, PLoS one, 4.6, e6022

- Booyesen, F. (2002). An overview and evaluation of composite indices of development. *Social indicators research*, 59(2), 115-151.
- Bordons, M., Morillo, F., & Álvarez-Bornstein, B. (2024). Towards higher standardization of funding acknowledgements in scientific publications: Current status in the case of a national and a European research fellowship scheme. *Research Evaluation*, rvae017.
- Bornmann, L. (2014a). Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. *Journal of Informetrics*, 8(4), 935–950.
<https://doi.org/10.1016/j.joi.2014.09.007>
- Bornmann, L. (2014b). Do altmetrics point to the broader impact of research? An overview of benefits and disadvantages of altmetrics. *Journal of Informetrics*, 8(4), 895–903.
<https://doi.org/10.1016/j.joi.2014.09.005>
- Bornmann, L. (2015). What do altmetrics counts mean? A plea for content analyses. *Journal of the Association of Information Science and Technology*. <https://doi.org/10.6084/M9.FIGSHARE.1506833>
- Bornmann, L. (2018). Field classification of publications in Dimensions: a first case study testing its reliability and validity. *Scientometrics*, 117(1), 637–640. <https://doi.org/10.1007/s11192-018-2855-y>
- Bornmann, L. & Daniel, H. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3), pp.391-392.
- Bornmann, L., & Haunschild, R. (2015). Which people use which scientific papers? An evaluation of data from F1000 and Mendeley. *Journal of Informetrics*, 9(3), 477–487. <https://doi.org/10.1016/j.joi.2015.04.001>
- Bornmann, L., & Haunschild, R. (2016). To what extent does the Leiden Manifesto also apply to altmetrics? A discussion of the manifesto against the background of research into altmetrics. *Online Information Review*.
- Bornmann, L., & Haunschild, R. (2016a). Normalization of Mendeley reader impact on the reader- and paper-side: A comparison of the mean discipline normalized reader score (MDNRS) with the mean normalized reader score (MNRS) and bare reader counts. *Journal of Informetrics*, 10(3), 776–788.
<https://doi.org/10.1016/j.joi.2016.04.015>
- Bornmann, L., & Haunschild, R. (2016b). How to normalize Twitter counts? A first attempt based on journals in the Twitter Index. *Scientometrics*, 107(3), 1405–1422. <https://doi.org/10.1007/s11192-016-1893-6>
- Bornmann, L., & Haunschild, R. (2016c). Citation score normalized by cited references (CSNCR): The introduction of a new citation impact indicator. *Journal of Informetrics*, 10(3), 875-887.
- Bornmann, L., & Haunschild, R. (2018a). Normalization of zero-inflated data: An empirical analysis of a new indicator family and its use with altmetrics data. *Journal of Informetrics*, 12(3), 998–1011.
<https://doi.org/10.1016/j.joi.2018.01.010>
- Bornmann, L., & Haunschild, R. (2018b). Do altmetrics correlate with the quality of papers? A large-scale empirical study based on F1000Prime data. *PLOS ONE*, 13(5), e0197133.
<https://doi.org/10.1371/journal.pone.0197133>

- Bornmann, L., Haunschild, R., & Adams, J. (2019). Do altmetrics assess societal impact in a comparable way to case studies? An empirical test of the convergent validity of altmetrics based on data from the UK research excellence framework (REF). *Journal of Informetrics*, 13(1), 325–340.
<https://doi.org/10.1016/j.joi.2019.01.008>
- Bornmann, L., Haunschild, R. & Marx, W. (2016). Policy documents as sources for measuring societal impact: how often is climate change research mentioned in policy-related documents?. *Scientometrics* **109**, 1477–1495. <https://doi.org/10.1007/s11192-016-2115-y>
- Bornmann, L., Marx, W., Gasparyan, A., Kitas, G. (2012) Diversity, value and limitations of the journal impact factor and alternative metrics. *Rheumatology International*, 32(7), pp.1861-1867
- Bornmann, L., Mutz, R. & Daniel, H. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), pp.830-837.
- Bornmann, L., & Society, M. P. (2014). Measuring the broader impact of research : The potential of altmetrics. *Recuperado De, June*.
- Boyack, K. W. (2009). Linking grants to articles: Characterization of NIH grant information indexed in Medline. In *12th International Conference of the International Society for Scientometrics and Informetrics* (Vol. 2009, pp. 730-41).
- BRAUN, T., GLANZEL, W. & SCHUBERT, A. (2006). A Hirsch-type index for journals. *Scientometrics*, 69, 169-173.
- Brinn, T., Jones, M. J., & Pendlebury, M. (2000). Measuring research quality: Peer review 1, citation indices 0. *Omega*, 28(2), 237-239.
- Brooks, T. A. (1985). Private acts and public objects: An investigation of citer motivations. *Journal of the American Society for Information Science*, 36(4), 223-229.
- Burnham, J. F. (2006). Scopus database: A review. *Biomedical Digital Libraries*, 3(1), 1.
- Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological methods & research*, 33(2), 261-304.
- Byrne, B. M. (2016). Structural Equation Modeling with AMOS: basic concepts, applications, and programming (multivariate applications series). *New York: Taylor & Francis Group*.
- CABS (2024, October 1). *Methodology*. Academic Journal Guide 2024. Retrieved November 20, 2024, from <https://charteredabs.org/academic-journal-guide>
- Carà, P. D., Ciriminna, R., & Pagliaro, M. (2017). Has the time Come for preprints in chemistry?. *ACS omega*, 2(11), 7923-7928.
- Carpenter, T. A., Lagace, N., & Bahnmaier, S. (2016). Developing standards for emerging forms of assessment: The NISO altmetrics initiative. *Serials Librarian*, 70(1–4), 260–265.
<https://doi.org/10.1080/0361526X.2016.1157737>
- Chakraborty, M., Byshkin, M., & Crestani, F. (2020). Patent citation network analysis: A perspective from descriptive statistics and ERGMs. *Plos one*, 15(12), e0241797.

- Champieux, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, 38(5), 1229–1261.
<https://doi.org/10.1017/S014271641700011X>
- Chang, J., Desai, N., & Gosain, A. (2019). Correlation Between Altmetric Score and Citations in Pediatric Surgery Core Journals. *Journal of Surgical Research*, 243(February), 52–58.
<https://doi.org/10.1016/j.jss.2019.05.010>
- Chawla, D. S. (2022). Massive open index of scholarly papers launches. *Nature*.
<https://doi.org/10.1038/d41586-022-00138-y>
- Chawla, D. (2023, December 19). Has your research influenced policy? Use this free tool to check. *Nature*.
<https://doi.org/10.1038/d41586-023-04070-7>
- Checkland, P., & Poulter, J. (2010). Soft systems methodology. In *Systems approaches to managing change: A practical guide* (pp. 191-242). Springer, London.
- Chen, W. (2007). Study on the Evidences Chain. *Journal of National Prosecutors College*, 2007(4), 128-136
- CHEW M., VILLANUEVA E.V. and VAN DER WEYDEN M.B. (2007): Life and times of the impact factor: Retrospective analysis of trends for seven medical journals (1994–2005) and their editors' views, *Journal of the Royal Society of Medicine*, 100, 142–150.
- Clarivate (2022, May 25). *Web of Science: Cited Reference Search for Patents*. Support. Retrieved June 15, 2024, from https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Cited-Reference-Search-for-Patents?language=en_US
- Clarivate (2023). *Preprint Citation Index*. Web of Science Help. Retrieved June 30, 2024, from <https://webofscience.help.clarivate.com/en-us/Content/pre-prints/pre-prints.htm>
- Clarivate (2023, August 10). *Research with confidence: Explore the Web of Science Core Collection database*. Retrieved August 10, 2023, from <https://clarivate.com/academia-government/scientific-and-academic-research/research-discovery-and-referencing/web-of-science/web-of-science-core-collection/>
- Clarivate (2024). *Coverage*. Derwent Patents Citation Index. Retrieved June 15, 2024, from <https://clarivate.com/products/ip-intelligence/ip-data-and-apis/derwent-patents-citation-index/>
- Clarivate Support (2023). *Web of Science Core Collection: Availability of funding data*. Support information provided by Clarivate. Retrieved May 10, 2024, from https://support.clarivate.com/ScientificandAcademicResearch/s/article/Web-of-Science-Core-Collection-Availability-of-funding-data?language=en_US
- Collier, J. (2020). *Applied structural equation modeling using AMOS: Basic to advanced techniques*. Routledge.
- Cook, D. J., Mulrow, C. D., & Haynes, R. B. (1997). Systematic reviews: Synthesis of best evidence for clinical decisions. In *Annals of Internal Medicine* (Vol. 126, Issue 5, pp. 376–380). American College of Physicians.
<https://doi.org/10.7326/0003-4819-126-5-199703010-00006>
- Cooksey, D. S. (2006). A review of UK health research funding.
- Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “altmetrics” correlate with citations? Extensive comparison of

- altmetric indicators with citations from a multidisciplinary perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019. <https://doi.org/10.1002/asi.23309>
- Cronin, B. & Meho, L. (2006). Using the h-index to rank influential information scientists. *Journal of the American Society for Information Science and Technology*, 57, 1275-1278.
- Cronin, B., Snyder, H. W., Rosenbaum, H., Martinson, A., & Callahan, E. (1998). Invoked on the Web. *Journal of the American Society for Information Science*, 49(14), 1319-1328.
- Cronin, B., & Sugimoto, C. R. (2014). Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact. In *MIT Press*. <https://ieeexplore-ieee-org.chain.kent.ac.uk/book/6825200>
- CUELLAR M.J., TAKEDA H. and TRUEX D.P. (2008): The Hirsch Family of Bibliometric Indices as an Improved Measure of IS Academic Journal Impact, 14th America's Conference on Information Systems, Toronto, Ontario Canada.
- Culbert, J., Hobert, A., Jahn, N., Haupka, N., Schmidt, M., Donner, P., & Mayr, P. (2024). Reference Coverage Analysis of OpenAlex compared to Web of Science and Scopus. *arXiv preprint arXiv:2401.16359*.
- Dardas, L. A., Sallam, M., Woodward, A., Sweis, N., Sweis, N., & Sawair, F. A. (2023). Evaluating research impact based on Semantic Scholar highly influential citations, total citations, and altmetric attention scores: the quest for refined measures remains illusive. *Publications*, 11(1), 5.
- Dardas, L. A., Woodward, A., Scott, J., Xu, H., & Sawair, F. A. (2019). Measuring the social impact of nursing research: An insight into altmetrics. *Journal of Advanced Nursing*, 75(7), 1394–1405. <https://doi.org/10.1111/jan.13921>
- Darling, E. S., Shiffman, D., Cote, I., & Drew, J. (2013). *The role of Twitter in the life cycle of a scientific publication* (p. 1:e16v1). PeerJ PrePrints. <https://doi.org/https://doi.org/10.7287/peerj.preprints.16v1>
- DAVIS P.M. (2008): Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts, *Journal of American Society of Information and Scientific Technology*, 59, 2186-2188.
- De Bellis, N. (2009). *Bibliometrics and citation analysis: from the science citation index to cybermetrics*. Scarecrow Press.
- De Filippo, D., & Sanz-Casado, E. (2018). Bibliometric and Altmetric Analysis of Three Social Science Disciplines. *Frontiers in Research Metrics and Analytics*, 3(December), 1–13. <https://doi.org/10.3389/frma.2018.00034>
- Didegah, F., Bowman, T. D., & Holmberg, K. (2018). On the differences between citations and altmetrics: An investigation of factors driving altmetrics versus citations for finnish articles. *Journal of the Association for Information Science and Technology*, 69(6), 832–843. <https://doi.org/10.1002/asi.23934>
- Dimensions (2019, November 19). *Are preprints included in Dimensions?* Support Home - Solutions. Retrieved June 30, 2024, from <https://dimensions.freshdesk.com/support/solutions/articles/23000019330-are-preprints-included-in-dimensions->
- Dimensions (2022, May 11). *Dimensions Life Sciences & Chemistry – expands patent coverage*. News. Retrieved June 15, 2024, from <https://www.dimensions.ai/news/dimensions-life-sciences-chemistry-expands->

patent-coverage/

Dimensions. (2022, Oct 28). *Information about Patent data in Dimensions*. Retrieved June 15, 2024, from <https://dimensions.freshdesk.com/support/solutions/articles/23000012996-information-about-patent-data-in-dimensions>

Dimensions (2023, December 5). *Information about clinical trials data in Dimensions*. Solutions. <https://dimensions.freshdesk.com/support/solutions/articles/23000012995-information-about-clinical-trials-data-in-dimensions>

Directory of Open Access Preprint Repositories (2024). *Repositories*. Retrieved June 30, 2024, from <https://doapr.coar-repositories.org/repositories/>

Dobrusskin, C. (2016). On the Identification of Contradictions Using Cause Effect Chain Analysis. *Procedia CIRP*, 39, 221–224.

Dong, P., Loh, M., & Mondry, A. (2005). The “impact factor” revisited. In *Biomedical Digital Libraries* (Vol. 2, Issue 1, pp. 1–8). BioMed Central. <https://doi.org/10.1186/1742-5581-2-7>

dos Santos Ribas, L. G., Pressey, R. L., Loyola, R., & Bini, L. M. (2020). A global comparative analysis of impact evaluation methods in estimating the effectiveness of protected areas. *Biological Conservation*, 246, 108595.

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

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, 69, 121-129.

Egghe L. and Rousseau R. (2007): An h-index weighted by citation impact, *Information. Processing and Management*, 44, 770-780.

Egghe, L., Fassin, Y., & Rousseau, R. (2019). Equalities between h-type Indices and Definitions of Rational h-type Indicators. *Journal of Data and Information Science*, 4(1), 22-31.

Elmore, S. A. (2018). The Altmetric Attention Score: What Does It Mean and Why Should I Care? In *Toxicologic Pathology* (Vol. 46, Issue 3, pp. 252–255). SAGE Publications Inc. <https://doi.org/10.1177/0192623318758294>

Elsevier (2024). *Scope and coverage policy for preprints*. Scopus - Content Policy and Selection. Retrieved June 30, 2024, from <https://www.elsevier.com/products/scopus/content/content-policy-and-selection?trial=true#6-preprints>

Erdt, M., Nagarajan, A., Sin, S. C. J., & Theng, Y. L. (2016). Altmetrics: an analysis of the state-of-the-art in measuring research impact on social media. *Scientometrics*, 109(2), 1117–1166. <https://doi.org/10.1007/s11192-016-2077-0>

Espacenet. (2024). *Introduction to Espacenet*. Retrieved August 1, 2024, from https://worldwide.espacenet.com/?locale=en_EP&view=intro

Ezema, I. J. (2017). Measuring Research Impact of Library and Information Science Journals : Citation verses Altmetrics. *IFLA Library*, 1–10. <http://library.ifla.org/1779/1/080-ezema-en.pdf>

EU Funding & Tenders Online Manual (2021). Online Manual for EU funding & Tenders Portal, European

- Commission. Retrieved May 10, 2024, from <https://www.horizonteeuropa.es/file-download/download/public/838>
- FALAGAS, M. E., KOURANOS, V. D., ARENCIBIA-JORGE, R. & KARAGEORGOPOULOS, D. E. 2008. Comparison of SCImago journal rank indicator with journal impact factor. *The FASEB journal*, 22, 2623-2628.
- Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *The FASEB Journal*, 22(2), 338-342.
- Fang, Z., & Costas, R. (2018). Studying the Velocity Index for various Altmetric. com data sources. *Proceedings of the 23rd International Conference on Science and Technology Indicators (STI2018)*, 850–856. <https://openaccess.leidenuniv.nl/handle/1887/65278>
- Fang, Z., & Costas, R. (2020). Studying the accumulation velocity of altmetric data tracked by Altmetric.com. *Scientometrics*, 123(2), 1077–1101. <https://doi.org/10.1007/s11192-020-03405-9>
- Faulkner, S. (2020, September 3). *EBSCO metrics in PlumX to be discontinued*. Plum Analytics. Retrieved June 20, 2024, from <https://plumanalytics.com/ebSCO-metrics-in-plumx-to-be-discontinued/>
- Faulkner, S. (2021, October 13). *Measuring Societal Impact with PlumX Metrics and Overton*. Retrieved July 15, 2024, from <https://plumanalytics.com/measuring-societal-impact-with-overton/>
- Fausto, S., Machado, F. A., Bento, L. F. J., Iamarino, A., Nahas, T. R., & Munger, D. S. (2012). Research Blogging: Indexing and Registering the Change in Science 2.0. *PLoS ONE*, 7(12), e50109. <https://doi.org/10.1371/journal.pone.0050109>
- FERSHT, A. (2009). The most influential journals: Impact Factor and Eigenfactor. *Proceedings of the National Academy of Sciences of the U.S.A*, 6883-6884.
- Firdaus, M., Purnomo, A., Rosyidah, E., Young, M. N., Anam, F., Prasetyo, Y. T., & Persada, S. F. (2024). Patent Landscape of Social Media in Education: A Review. *Procedia Computer Science*, 234, 780-786.
- Fong, E. A., & Wilhite, A. W. (2017). Authorship and citation manipulation in academic research. *PloS one*, 12(12), e0187394.
- Fry, J., Oppenheim, C., Creaser, C., Johnson, W., Summers, M., White, S., & Hartley, D. (2009). Communicating knowledge: How and why researchers publish and disseminate their findings. *Supporting paper*, 4.
- Evans, M., Dalkir, K., & Bidian, C. (2014). A holistic view of the knowledge life cycle: the knowledge management cycle (KMC) model. *Electronic Journal of Knowledge Management*, 12(2), pp85-97.
- Galligan, F., & Dias-Correia, S. (2013). Altmetrics: Rethinking the Way We Measure. *Serials Review*, 39(1), 56–61. <https://doi.org/10.1080/00987913.2013.10765486>
- Gamble, J. M., Traynor, R. L., Gruz, A., Mai, P., Dormuth, C. R., & Sketris, I. S. (2020). Measuring the impact of pharmacoepidemiologic research using altmetrics: A case study of a <scp>CNODES</scp> drug-safety article. *Pharmacoepidemiology and Drug Safety*, 29(S1), 93–102. <https://doi.org/10.1002/pds.4401>
- Garbarino, S., & Holland, J. (2009). Quantitative and qualitative methods in impact evaluation and measuring results.
- Garden, M. (2012). Defining blog: A fool's errand or a necessary undertaking. *Journalism: Theory, Practice &*

- Criticism*, 13(4), 483–499. <https://doi.org/10.1177/1464884911421700>
- Garfield, E. (1971). The mystery of transposed journal lists— wherein Bradford’s Law of Scattering is generalized according to Garfield’s Law of Concentration. *Current Contents*, No. 17, 5–6, reprinted in Garfield, E., *Essays of an Information Scientist, 1962–1973*, I (pp. 222–223). ISI Press, Philadelphia.
- Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, 178(4060), 471–479. <https://doi.org/10.1126/science.178.4060.471>
- Geuna, A., & Martin, B. R. (2003). University research evaluation and funding: An international comparison. *Minerva*, 41(4), 277–304. <https://doi.org/10.1023/B:MINE.0000005155.70870.bd>
- Gibson , E. (2021, November 8). *What’s new on the Document details page? A Views Count metric, author previews, and enhanced Reaxys integration!* Scopus. Retrieved October 30, 2023, from <https://blog.scopus.com/posts/what-s-new-on-the-document-details-page-a-views-count-metric-author-previews-and-enhanced>
- Gingras, Y. (2016). *Bibliometrics and Research Evaluation: Uses and Abuses*. In MIT Press. MIT Press.
- Glanzel, W. (2006). On the h-index - A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67, 315–321.
- Glanzel, W. & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information processing & management*, 35(1), pp.31-44
- Glanzel, W. & Meyer, M. (2003). Patents cited in the scientific literature: An exploratory study of reverse citation relations. *Scientometrics*, 58, pp.415-428
- Glänzel, W., Moed, H. F., Schmoch, U., & Thelwall, M. (2019). *Springer Handbook of Science and Technology Indicators*. Springer International Publishing. http://link.springer.com/10.1007/978-3-030-02511-3_14
http://link.springer.com/10.1007/978-3-030-02511-3_14
http://link.springer.com/10.1007/978-3-030-02511-3_14
- Goldbeck-Wood, S. (1999). Evidence on peer review—scientific quality control or smokescreen?. *Bmj*, 318(7175), 44-45.
- González-Pereira, B., Guerrero-Boteb, V., & Moya-Anegón, F. (2010). A new approach to the metric of journals’ scientific prestige: The SJR indicator. *Journal of Informetrics*, 4(3), pp. 379-391
- González-Valiente, C. L., Pacheco-Mendoza, J., & Arencibia-Jorge, R. (2016). A review of altmetrics as an emerging discipline for research evaluation. *Learned Publishing*, 29(4), 229-238.
- Graham, I. D., Logan, J., Harrison, M. B., Straus, S. E., Tetroe, J., Caswell, W., & Robinson, N. (2006). Lost in knowledge translation: time for a map?. *Journal of continuing education in the health professions*, 26(1), 13-24.
- Grant, J. (2006). Measuring the benefits from research. *Policy Resource*.
- Grants.gov (2024). The Grants.gov program management office, Department of Health and Human Services, U.S. government. Retrieved May 10, 2024, from <https://www.grants.gov/search-grants>
- Grassano N., Rotolo D., Hutton J., Lang F., Hopkins M. (2017) ‘Funding Data from Publication

- Acknowledgments: Coverage, Uses, and Limitations', *Journal of the Association for Information Science and Technology*, 68: 999–1017.
- Greene, J. C. (2016). Advancing equity: Cultivating an evaluation habit. *Evaluation for an equitable society*, 49-66.
- Gumpenberger, C., Glänzel, W., & Gorraiz, J. (2016). The ecstasy and the agony of the altmetric score. *Scientometrics*, 108(2), 977–982. <https://doi.org/10.1007/s11192-016-1991-5>
- Guthrie, S., Wamae, W., Diepeveen, S., Wooding, S., & Grant, J. (2013). Measuring Research - A guide to research evaluation frameworks and tools. In *RAND Europe*.
- Hammarfelt, B. (2014). Using altmetrics for assessing research impact in the humanities. *Scientometrics*, 101(2), 1419–1430. <https://doi.org/10.1007/s11192-014-1261-3>
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate Data Analysis: Overview of Multivariate Methods* (Seventh Edition ed.). Pearson Prentice Hall: Upper Saddle River, New Jersey: Pearson Education International.
- Harzing, A.-W., & van der Wal, R. (2008). Google Scholar as a new source for citation analysis. *Ethics in Science and Environmental Politics*, 8(1), 61–73.
- Hassan, S.-U., Imran, M., & Zuccala, A. (2018). Alt-Index: A proposed Index for measuring the Social Activity of Scientific Research. *Proceedings of the 23rd International Conference on Science and Technology Indicators*. <https://hdl.handle.net/1887/64521>
- Haunschild, R., & Bornmann, L. (2016). Normalization of Mendeley reader counts for impact assessment. *Journal of Informetrics*, 10(1), 62–73. <https://doi.org/10.1016/j.joi.2015.11.003>
- Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author keywords and hashtags. *Journal of Informetrics*, 13(2), 695–707. <https://doi.org/10.1016/j.joi.2019.03.008>
- Haustein, S. (2016). Grand challenges in altmetrics: heterogeneity, data quality and dependencies. *Scientometrics*, 108(1), 413–423. <https://doi.org/10.1007/s11192-016-1910-9>
- Haustein, S., Bowman, T. D., & Costas, R. (2016). Interpreting 'Altmetrics': Viewing Acts on Social Media through the Lens of Citation and Social Theories. In C. R. Sugimoto (Ed.), *Theories of Informetrics and Scholarly Communication* (pp. 372–406). De Gruyter. <https://doi.org/10.1515/9783110308464-022>
- Haustein, S., Costas, R., & Larivière, V. (2015). Characterizing Social Media Metrics of Scholarly Papers: The Effect of Document Properties and Collaboration Patterns. *PLOS ONE*, 10(3), 1–21. <https://doi.org/10.1371/journal.pone.0120495>
- Haustein, S., Peters, I., Bar-Ilan, J., Priem, J., Shema, H., & Terliesner, J. (2014). Coverage and adoption of altmetrics sources in the bibliometric community. *Scientometrics*, 101(2), 1145–1163. <https://doi.org/10.1007/s11192-013-1221-3>
- Haustein, S., & Siebenlist, T. (2011). Applying social bookmarking data to evaluate journal usage. *Journal of Informetrics*, 5, 446–457. <https://doi.org/10.1016/j.joi.2011.04.002>

- HEFCE (2009). *Research Excellence Framework -second consultation on the assessment and funding of research* [Online]. Retrieved July 20, 2020, from http://www.hefce.ac.uk/pubs/hefce/2009/09_38/09_38.pdf
- Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative science studies*, 1(1), 387-395.
- Hirsch, Jorge E. (2005). An index to quantify an individual's scientific research output. *PNAS*, 102(46), pp.16569-16572
- Holmberg, K., & Bowman, T. D. (2019). Google Search results as an altmetrics data source? *Proceedings of the 17th International Conference on Scientometrics and Informetrics*, 477–488. <https://www.utu.fi/en/publications/google-search-results-as-an-altmetrics-data-source>
- Holmberg, Kim, & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, 101(2), 1027–1042. <https://doi.org/10.1007/s11192-014-1229-3>
- Hook, D. W., Porter, S. J., & Herzog, C. (2018). Dimensions: building context for search and evaluation. *Frontiers in Research Metrics and Analytics*, 3, 23.
- House of Commons (2011). *Peer review in scientific publications - Eighth Report of Session 2010–12*. House of Commons Science and Technology Committee. <https://publications.parliament.uk/pa/cm201012/cmselect/cmsctech/856/856.pdf>
- Howard, J. (2013). Rise of “altmetrics” revives questions about how to measure impact of research. *The Chronicle of Higher Education*, 7–13. http://chronicle.com/article/Rise-of-Altmetrics-Revives/139557/?cid=wc&utm_source=wc&utm_medium=en
- Htoo, T., & Jin-Cheon, N. (2017). *User Profiling in Altmetrics : the need to enrich altmetric data* (Issues altmetrics17, workshop).
- HTOO, T., & Na, J.-C. (2015). Comparison of Altmetrics across Multiple Disciplines: Psychology, History, and Linguistics. *International Conference of Asian Special Libraries 4th ICoASL 2015, April 22-24*, 1–12.
- Huang, M. H., & Chang, Y. W. (2008). Characteristics of research output in social sciences and humanities: From a research evaluation perspective. *Journal of the American Society for Information Science and Technology*, 59(11), 1819-1828.
- HUANG, Y., XIAOLAN, A. & YUH-SHAN, H. (2008). Use of citation per publication as an indicator to evaluate pentachlorophenol research. *Scientometrics*, 75, 67-80.
- Hug, S., & Brandle, M. (2017). The coverage of Microsoft Academic: Analyzing the publication output of a university. *Scientometrics*, 113(3), 1551–1571. <https://doi.org/10.1007/s11192-017-2535-3>.
- Huser, V., & Cimino, J. J. (2013). Linking ClinicalTrials.gov and PubMed to track results of interventional human clinical trials. *PloS one*, 8(7), e68409.
- Jobmann, A., Hoffmann, C. P., Künne, S., Peters, I., Schmitz, J., & Wollnik-korn, G. (2014). Altmetrics for large, multidisciplinary research groups: Comparison of current tools. *Bibliometrie-Praxis Und Forschung*, 3. <https://doi.org/https://doi.org/10.5283/bpf.205>
- Joshi, P. B., & Pandey, M. (2024). Deception through manipulated citations and references as a growing

- problem in scientific publishing. In *Scientific Publishing Ecosystem: An Author-Editor-Reviewer Axis* (pp. 285-306). Singapore: Springer Nature Singapore.
- Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R-and AR-indices: Complementing the h-index. *Chinese science bulletin*, 52(6), 855-863.
- Jin, Z., Hou, J., & Sun, H. (2019). A study on present status, future and related issues of altmetric data aggregated tools. *Information studies: theory & application*, 042(004), 89-95.
- Kassab, O., Bornmann, L., & Haunschild, R. (2020). Can altmetrics reflect societal impact considerations?: Exploring the potential of altmetrics in the context of a sustainability science research center. *Quantitative Science Studies*, 1(2), 792-809.
- Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, 112(24), 7426–7431. <https://doi.org/10.1073/pnas.1424329112>
- KELLY, C. D. & JENNIONS, M. D. 2006. The h-index and career assessment by numbers. *TRENDS in Ecology and Evolution*, 21, 167-170.
- Kenix, L. J. (2009). Blogs as Alternative. *Journal of Computer-Mediated Communication*, 14(4), 790–822. <https://doi.org/10.1111/j.1083-6101.2009.01471.x>
- Khazragui, H., & Hudson, J. (2015). Measuring the benefits of university research: Impact and the REF in the UK. *Research Evaluation*, 24(1), 51–62. <https://doi.org/10.1093/reseval/rvu028>
- Kirkham, J. J., Penfold, N. C., Murphy, F., Boutron, I., Ioannidis, J. P., Polka, J., & Moher, D. (2020). Systematic examination of preprint platforms for use in the medical and biomedical sciences setting. *BMJ open*, 10(12), e041849.
- Kline, R. B. (2023). *Principles and practice of structural equation modeling*. Guilford publications.
- Kolahi, J., & Khazaei, S. (2016). Altmetric: Top 50 dental articles in 2014. *British Dental Journal*, 220(11), 569–574. <https://doi.org/10.1038/sj.bdj.2016.411>
- Konkiel, S. (2013). Tracking citations and altmetrics for research data: Challenges and opportunities. *Bulletin of the American Society for Information Science and Technology*, 39(6), 27–32. <https://doi.org/10.1002/bult.2013.1720390610>
- Konkiel, S. (2016). Altmetrics: Diversifying the understanding of influential scholarship. *Palgrave Communications*, 2, 1–7. <https://doi.org/10.1057/palcomms.2016.57>
- Konkiel, S., & Scherer, D. (2013). New opportunities for repositories in the age of altmetrics. *Bulletin of the American Society for Information Science and Technology*, 39(4), 22-26.
- KOSMULSKI, M. (2007). MAXPROD - A New Index for assessment of the scientific output of an individual, and a comparison with the h-index. *International Journal of Scientometrics, Informetrics and Bibliometrics*, 11, 13-18.
- Kousha, K., & Thelwall, M. (2017). Patent citation analysis with Google. *Journal of the Association for Information Science and Technology*, 68(1), 48-61.

- Krabbe, E., Sampson, S., & Wetherbee, I. (2017). Patent searching using free search tools. *Intellectual Property Owner's Association*.
- Kraker, P., Lex, E., Gorraiz, J., Gumpenberger, C., & Peters, I. (2015). *Research Data Explored II: the Anatomy and Reception of figshare*. 8010. <http://arxiv.org/abs/1503.01298>
- Kumar, M. (2009). A review of the review process: Manuscript peer-review in biomedical research. *Biology and Medicine*, 1(4), 1–16. <https://doi.org/10.4172/0974-8369.1000047>
- Lacey, J., Coates, R., & Herington, M. (2020). Open science for responsible innovation in Australia: understanding the expectations and priorities of scientists and researchers. *Journal of Responsible Innovation*, 7(3), 427-449.
- Lapinski, S., Piwowar, H., & Priem, J. (2013). Riding the crest of the altmetrics wave: How librarians can help prepare faculty for the next generation of research impact metrics. arXiv preprint arXiv:1305.3328.
- Larivière, V., & Gingras, Y. (2010). The impact factor's Matthew Effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology*, 61(2), 424-427.
- Lee, J. W., Lee, W. K., & Sohn, S. Y. (2021). Patenting trends in biometric technology of the Big Five patent offices. *World Patent Information*, 65, 102040.
- LEEUWEN, T. N. V. & F.MOED, H. 2005. Characteristics of Journal Impact Factors: The effects of uncitedness and citation distribution on the understanding of journal impact factors. *Scientometrics*, 63, 357-371.
- Lemay, M., & Sá, C. (2012). Complexity sciences: towards an alternative approach to understanding the use of academic research. *Evidence & Policy*, 8(4), 473-494.
- Li, X., Thelwall, M., & Giustini, D. (2012). Validating online reference managers for scholarly impact measurement. *Scientometrics*, 91(2), 461–471. <https://doi.org/10.1007/s11192-011-0580-x>
- Lin, J. (2018, May 31). *Preprints growth rate ten times higher than journal articles*. Blog. Retrieved June 30, 2024, from <https://www.crossref.org/blog/preprints-growth-rate-ten-times-higher-than-journal-articles/>
- Lin, J., & Fenner, M. (2013). Altmetrics in Evolution: Defining and Redefining the Ontology of Article-Level Metrics. *Information Standards Quarterly*, 25(2), 20. <https://doi.org/10.3789/isqv25no2.2013.04>
- Liu, C. (2016). Development, Current Situation and Related Issues of Altmetrics Tools. *Library and Information Service*, 2016(5), 87-92.
- Liu, J., & Adie, E. (2013). Five challenges in altmetrics: A toolmaker's perspective. *Bulletin of the American Society for Information Science and Technology*, 39(4), 31–34. <https://doi.org/10.1002/bult.2013.1720390410>
- Liu, L., Jones, B. F., Uzzi, B., & Wang, D. (2023). Data, measurement and empirical methods in the science of science. *Nature human behaviour*, 7(7), 1046-1058.
- Liu, L., & Wang, Q. (2016) Does Altmetrics Apply to Influence Evaluation? —A Meta-analysis of Foreign Scholars' Empirical Studies and Related Controversies. *Information Studies: Theory & Application*, 39(6), 7-14
- Liu, W. (2020). Accuracy of funding information in Scopus: A comparative case study. *Scientometrics*, 124(1),

803-811.

- Liu W., Tang L., Hu G. Y. (2020) 'Funding Information in Web of Science: An Updated Overview', *Scientometrics*, 122: 1509–24.
- Liu, Y., Zhang, Y., Qi, L., & Hui, S. (2018). Knowledge Enlightenment or Authority Adoration: An Empirical Study of Citation Motivation Based on Duplicate Publications (in Chinese), *Library Tribune*, 2018(4), 49-57
- Madhanagopal, B., & Amrhein, J. (2019). Analyzing Structural Causal Models Using the CALIS Procedure. In *SAS Global Forum* (pp. 3240-2019).
- Maleki, A., & Holmberg, K. (2022). Comparing coverage of policy citations to scientific publications in Overton and Altmetric. com: Case study of Finnish research organizations in Social Science. *Informaatiotutkimus*, 41(2–3), 92-96.
- Mao, J., Liu, Y., & Ma, S. (2014). Social Influence Analysis for Micro-Blog User Based on User Behaviour. *Chinese Journal of Computers*, 37(4).
- Manghi, Paolo, et al. (2019). "The OpenAIRE research graph data model." *Zenodo*
- Marginson, S. (2022). What drives global science? The four competing narratives. *Studies in higher education*, 47(8), 1566-1584.
- Martín-Martín, A., Orduna-Malea, E., Ayllón, J. M., & Delgado López-Cózar, E. (2016). *The counting house, measuring those who count. January 2015*, 1–60.
- Mas-Bleda, A., & Thelwall, M. (2016). Can alternative indicators overcome language biases in citation counts? A comparison of Spanish and UK research. *Scientometrics*, 109(3), 2007–2030. <https://doi.org/10.1007/s11192-016-2118-8>
- Mas-Bleda, A., Thelwall, M., Kousha, K., & Aguillo, I. F. (2014). Do highly cited researchers successfully use the social web? *Scientometrics* 2014 101:1, 101(1), 337–356. <https://doi.org/10.1007/S11192-014-1345-0>
- Matthews, M.R. (2012). Mario Bunge, Systematic Philosophy and Science Education: An Introduction. *Science & education*, 21(10), pp.1393-1403.
- Matyushok, V., Vera Krasavina, V., Berezin, A., & Sendra García, J. (2021). The global economy in technological transformation conditions: A review of modern trends. *Economic Research-Ekonomska Istraživanja*, 34(1), 1471-1497.
- McCullough, R. (2023, May 10). *Preprints are now searchable on Scopus!* Retrieved June 30, 2024, from <https://blog.scopus.com/posts/preprints-are-now-searchable-on-scopus>
- McDonald, R. P. (2014). *Factor analysis and related methods*. Psychology Press.
- McGillivray, B., Marongiu, P., Pedrazzini, N., Ribary, M., Wigdorowitz, M., & Zordan, E. (2022). Deep impact: A study on the impact of data papers and datasets in the humanities and social sciences. *Publications*, 10(4), 39.
- McGowan, J. (2017). Measuring most informative titles (declarative titles) as a knowledge translation dissemination tool is possible using altmetrics. *Journal of Clinical Epidemiology*, 85, 12–13. <https://doi.org/10.1016/j.jclinepi.2017.03.009>

- McKenna, H., Daly, J., Davidson, P., Duffield, C., & Jackson, D. (2012). RAE and ERA-spot the difference. *International journal of nursing studies*, 49(4), 375-377.
- Melicherová, M., Ondrišová, M., & Susol, J. (2020). Bibliometrics versus altmetrics: Researchers' attitudes in Slovakia. *Iberoamerican Journal of Science Measurement and Communication*, 1(1), 2.
- Merigó, J. & Yang, J. (2017). A bibliometric analysis of operations research and management science. *Omega*, 73, pp.37-48.
- Meschede, C., & Siebenlist, T. (2018). Cross-metric compatability and inconsistencies of altmetrics. *Scientometrics*, 115(1), 283–297. <https://doi.org/10.1007/s11192-018-2674-1>
- Michalek, A. (2016). NISO Altmetrics Working Group on Data Quality – Plum Analytics Code of Conduct Report. Available online: <https://plumanalytics.com/niso-altmetrics-working-group-on-data-quality/> (accessed on 1 Aug 2023).
- Milat, A. J., Bauman, A. E., & Redman, S. (2015). A narrative review of research impact assessment models and methods. *Health Research Policy and Systems*, 13(1), 18. <https://doi.org/10.1186/s12961-015-0003-1>
- Miles, R. A., & Price, R. (2023). Using Altmetric Data Responsibly: A Guide to Interpretation and Good Practice.
- Mingers, J., & Leydesdorff, L. (2015). A review of theory and practice in scientometrics. *European Journal of Operational Research*, 246(1), 1–19. <https://doi.org/10.1016/j.ejor.2015.04.002>
- Mingers, J., & Standing, C. (2017). Why things happen–Developing the critical realist view of causal mechanisms. *Information and Organization*, 27(3), 171-189.
- Mingers, J. & Xu, F. (2010). The drivers of citations in management science journals, *European Journal of Operational Research*, 205, pp.422-430.
- MOED H.F. (2010). Measuring contextual citation impact of scientific journals, *Journal of Informetrics*, 4, 265–277.
- Moed, H.F. (2012): Does open access publishing increase citation or download rates? *Research Trends*, 28.
- Moed, H. F. (2016). Altmetrics as Traces of the Computerization of the Research Process. In *Theories of Informetrics and Scholarly Communication* (pp. 360–371). <https://doi.org/10.1515/9783110308464-021>
- Moed, H. F., & Halevi, G. (2015). Multidimensional assessment of scholarly research impact. *Journal of the Association for Information Science and Technology*, 66(10), 1988–2002. <https://doi.org/10.1002/asi.23314>
- Moed, H. F., & Halevi, G. (2016). On full text download and citation distributions in scientific-scholarly journals. *Journal of the Association for Information Science and Technology*, 67(2), 412-431.
- Mohammadi, E., & Thelwall, M. (2014). Mendeley readership altmetrics for the social sciences and humanities: Research evaluation and knowledge flows. *Journal of the Association for Information Science and Technology*, 65(8), 1627–1638. <https://doi.org/10.1002/asi.23071>
- Mohammadi, E., Thelwall, M., Haustein, S., & Larivière, V. (2015). Who reads research articles? An altmetrics analysis of Mendeley user categories. *Journal of the Association for Information Science and Technology*, 66(9), 1832–1846. <https://doi.org/10.1002/asi.23286>

- Mohammadi, E., Thelwall, M., & Kousha, K. (2016). Can Mendeley Bookmarks Reflect Readership? A Survey of User Motivations. *Journal of the Association for Information Science and Technology*, 67(5), 1198–1209. <https://doi.org/10.1002/asi.23477>
- Mohammadi, E., Thelwall, M., Kwasny, M., & Holmes, K. L. (2018). Academic information on Twitter: A user survey. *PLOS ONE*, 13(5), e0197265. <https://doi.org/10.1371/journal.pone.0197265>
- Mohd Sarjidan, M. A., Abu Bakar, A., Ismail, N. H., Syed Omar, S. N. S., & Nordine, N. (2023). Practices Malaysian research universities: patent profile in the lens database. *Journal of Information and Knowledge Management (JIKM)*, 13(1), 75-89.
- Morrow, E. M., Goreham, H., & Ross, F. (2017). Exploring research impact in the assessment of leadership, governance and management research. *Evaluation*, 23(4), 407-431.
- Morton, S. (2015) 'Progressing Research Impact Assessment: A 'Contributions' Approach', *Research Evaluation*, 24: 405–19
- Mounce, R. (2013). Open access and altmetrics: Distinct but complementary. *Bulletin of the American Society for Information Science and Technology*, 39(4), 14-17.
- Mugabushaka, A. M. (2020). Linking publications to funding at project level: A curated dataset of publications reported by FP7 projects. *arXiv preprint arXiv:2011.07880*.
- Mukherjee, B., Subotić, S., & Chaubey, A. K. (2018). And now for something completely different: the congruence of the Altmetric Attention Score's structure between different article groups. *Scientometrics*, 114(1), 253–275. <https://doi.org/10.1007/s11192-017-2559-8>
- Mulrow, C. D., Cook, D. J., & F. Davidoff. (1997). Systematic reviews: Critical links in the great chain of evidence. In *Annals of Internal Medicine* (Vol. 126, Issue 5, pp. 389–390). American College of Physicians. <https://doi.org/10.7326/0003-4819-126-5-199703010-00008>
- Myers, J. L., Well, A. D., & Lorch Jr, R. F. (2013). *Research design and statistical analysis*. Routledge.
- NAMAZI, M. R. & 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 & Leprology*, 76, 229-230.
- Neavill, G. B. (1975). Role of the Publisher in the Dissemination of Knowledge. *The Annals of the American Academy of Political and Social Science*, 421(1), 23-31.
- Newson, R., Rychetnik, L., King, L. *et al.* (2018). Does citation matter? Research citation in policy documents as an indicator of research impact – an Australian obesity policy case-study. *Health Res Policy Sys*, 16, 55. <https://doi.org/10.1186/s12961-018-0326-9>
- Neylon, C., & Wu, S. (2009). Article-Level Metrics and the evolution of scientific impact. *PLoS biology*, 7(11), e1000242.
- Oberst, U. (2017). Measuring the Societal Impact of Research with Altmetrics: An Experiment. *Journal for Library Culture*, 5(1), 16–21. <https://openaccess.leidenuniv.nl/handle/1887/57323>
- Offutt, J. (2016). STVR policy on extending conference papers to journal submissions. *Software Testing*,

Verification and Reliability, 26(4), 274-275.

- Orduña-Malea, E., & Delgado-López-Cózar, E. (2018). Dimensions: re-discovering the ecosystem of scientific information. *Profesional de La Informacion*, 27(2), 420–431. <https://doi.org/10.3145/epi.2018.mar.21>
- Ortega, J. L. (2018a). Disciplinary differences of the impact of altmetric. *FEMS Microbiology Letters*, 365(7), 1–6. <https://doi.org/10.1093/femsle/fny049>
- Ortega, J. L. (2018b). Reliability and accuracy of altmetric providers: a comparison among Altmetric.com, PlumX and Crossref Event Data. *Scientometrics*, 116(3), 2123–2138. <https://doi.org/10.1007/s11192-018-2838-z>
- Ortega, J. L. (2019). The coverage of blogs and news in the three major altmetric data providers. Conference proceedings, *17th International conference of the international society for scientometrics and informetrics*, Rome, Italy.
- Ortega, J. L. (2020). Blogs and news sources coverage in altmetrics data providers: a comparative analysis by country, language, and subject. *Scientometrics*, 122(1), 555–572. <https://doi.org/10.1007/s11192-019-03299-2>
- Overton (2024). *Explore the data*. Retrieved July 15, 2024, from https://app.overton.io/policy_sources.php
- Patthi, B., Prasad, M., Gupta, R., Singla, A., Kumar, J. K., Dhama, K., Ali, I., & Niraj, L. K. (2017). Altmetrics – A collated adjunct beyond citations for scholarly impact: A systematic review. *Journal of Clinical and Diagnostic Research*, 11(6), ZE16–ZE20. <https://doi.org/10.7860/JCDR/2017/26153.10078>
- Penfield, T., Baker, M. J., Scoble, R., & Wykes, M. C. (2014). Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation*, 23(1), 21–32. <https://doi.org/10.1093/reseval/rvt021>
- Penfold, R. (2020). Using the Lens database for staff publications. *Journal of the Medical Library Association: JMLA*, 108(2), 341.
- Peters, I., Jobmann, A., Eppelin, A., Hoffmann, C. P., Künne, S., & Wollnik-korn, G. (2014). Altmetrics for large, multidisciplinary research groups: A case study of the Leibniz Association. *Libraries in the Digital Age (LIDA) Proceedings*, 13.
- Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2015). Research data explored: Citations versus altmetrics. *Proceedings of ISSI 2015 Istanbul: 15th International Society of Scientometrics and Informetrics Conference*, 172–183.
- Peters, I., Kraker, P., Lex, E., Gumpenberger, C., & Gorraiz, J. (2016). Research data explored: an extended analysis of citations and altmetrics. *Scientometrics*, 107(2), 723–744. <https://doi.org/10.1007/s11192-016-1887-4>
- Pinkowitz, L. (2012). Research dissemination and impact: Evidence from web site downloads. *The Journal of Finance*, 57(1), 485-499.
- Piwowar, H. (2013a). Introduction Altmetrics: What, Why and Where? *Bulletin of the American Society for Information Science and Technology*, 39(4), 8–9.
- Piwowar, H. (2013b). Value all research products. *Nature*, 493, 159.

- Plum Analytics (2021, October 13). *Citation Metrics*. Retrieved June 20, 2024, from <https://plumanalytics.com/learn/about-metrics/citation-metrics/>
- Polka, J. K., & Vale, R. D. (2015). Preprints: An Underutilized Mechanism to Accelerate Scholarly Communication and Science. *PLOS Biology*, 13(2), e1002093.
- Poulin-Dubois, D., & Rakison, D. H. (1999). A developmental theory of implicit and explicit knowledge? *Behavioral and Brain Sciences*, 22(5), 782–782. <https://doi.org/10.1017/S0140525X99522188>
- Priem, J., Groth, P., & Taraborelli, D. (2012). The Altmetrics Collection. *PLoS ONE*, 7(11), e48753. <https://doi.org/10.1371/journal.pone.0048753>
- Priem, J., Piwowar, H. A., & Hemminger, B. M. (2012). *Altmetrics in the wild: Using social media to explore scholarly impact*. <http://arxiv.org/abs/1203.4745>
- Priem, J., Piwowar, H., & Orr, R. (2022). OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). Altmetrics: A manifesto. Retrieved July 20, 2020, from <http://altmetrics.org/manifesto>
- Pulido, C. M., Redondo-Sama, G., Sordé-Martí, T., & Flecha, R. (2018). Social impact in social media: A new method to evaluate the social impact of research. *PLOS ONE*, 13(8), e0203117. <https://doi.org/10.1371/journal.pone.0203117>
- Qi, S., & Zheng, J. (2017). Analysis of Distribution of Research Strengths and Hotspots of Altmetrics Based on Web of Science. *Journal of Modern Information*, 37(10), 128-136.
- Qiu, J., et al. (2010). Evaluation Science – Theory, Method and Practice (in Chinese). *Science Publisher*, Beijing.
- RADICCHIA, F., FORTUNATO, S. & CASTELLANO, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, 105.
- Rawls, A. W. (1996). Durkheim's epistemology: the neglected argument. *American Journal of sociology*, 102(2), 430-482.
- Reale, E., Avramov, D., Canhial, K., Donovan, C., Flecha, R., Holm, P., Larkin, C., Lepori, B., Mosoni-Fried, J., Oliver, E., Primeri, E., Puigvert, L., Scharnhorst, A., Schubert, A., Soler, M., Soès, S., Sordé, T., Travis, C., & Van Horik, R. (2018). A review of literature on evaluating the scientific, social and political impact of social sciences and humanities research. *Research Evaluation*, 27(4), 298–308. <https://doi.org/10.1093/reseval/rvx025>
- Reed, M. S., Ferré, M., Martin-Ortega, J., Blanche, R., Lawford-Rolfe, R., Dallimer, M., & Holden, J. (2021). Evaluating impact from research: A methodological framework. *Research Policy*, 50(4), 104147.
- Reedijk, J. and Moed, H. (2008) Is the impact of journal impact factors decreasing? *Journal of Documentation*, 64(2), pp.183-192.
- REF (2021). Guidance on submissions (2019/01). Research Excellence Framework. Retrieved July 20, 2020, from <https://2021.ref.ac.uk/publications-and-reports/guidance-on-submissions-201901/index.html>
- Repiso, R., Castillo-Esparcia, A., & Torres-Salinas, D. (2019a). Altmetrics, alternative indicators for Web of

- Science Communication studies journals. *Scientometrics*, 119(2), 941–958.
<https://doi.org/10.1007/s11192-019-03070-7>
- Repiso, R., Castillo-Esparcia, A., & Torres-Salinas, D. (2019b). Altmetrics, alternative indicators for Web of Science Communication studies journals. *Scientometrics*, 119(2), 941–958.
<https://doi.org/10.1007/s11192-019-03070-7>
- Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PLOS ONE*, 12(8), e0183551.
<https://doi.org/10.1371/journal.pone.0183551>
- Robinson-Garcia, N., van Leeuwen, T. N., & Ràfols, I. (2018). Using altmetrics for contextualised mapping of societal impact: From hits to networks. *Science and Public Policy*, 45(6), 815–826.
<https://doi.org/10.1093/SCIPOL/SCY024>
- Rodgers, E. P., & Barrow, S. (2013). A Look at Altmetrics and Its Growing Significance to Research Libraries. *The University of Michigan University Library*, 0–14.
- Roemer, R. C., & Borchardt, R. (2015). *Altmetrics: library technology reports*. 42.
<https://journals.ala.org/index.php/ltr/article/view/5745/7185>
- Rosenkrantz, A. B., Ayoola, A., Singh, K., & Duszak, R. (2017). Alternative Metrics (“Altmetrics”) for Assessing Article Impact in Popular General Radiology Journals. *Academic Radiology*, 24(7), 891–897.
<https://doi.org/10.1016/j.acra.2016.11.019>
- Rousseau, R., & Ye, F. Y. (2013). A multi-metric approach for research evaluation. *Chinese Science Bulletin*, 58(26), 3288–3290. <https://doi.org/10.1007/s11434-013-5939-3>
- RUANE, F. & TOL, R. S. J. 2008. Rational (successive) h-indices: An application to economics in the Republic of Ireland. *Scientometrics*, 75, 395-405.
- Saheb, T., & Saheb, T. (2020). Understanding the development trends of big data technologies: an analysis of patents and the cited scholarly works. *Journal of Big Data*, 7(1), 12.
- Salinas DT, Clavijo AC, Contreras EJ.(2013). Altmetrics: New Indicators for Scientific Communication in Web 2.0. Comunicar. *Media Education Research Journal*, 1, 21(1).
- Salvi, R. (2020). The process of knowledge dissemination. *Lingue e Linguaggi*, 34, 15-30.
- Samuel, G. N., & Derrick, G. E. (2015). Societal impact evaluation: Exploring evaluator perceptions of the characterization of impact under the REF2014. *Research Evaluation*, 24(3), 229–241.
<https://doi.org/10.1093/reseval/rvv007>
- Sanghani, N., & MacDonald, M. (2024). The Future of Public Money: From Fiscal Fantasy to Reality. *Public Finance and Management*, 15239721241300572.
- SEGLÉN, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *BMJ*.
- Schlögl, C., Gorraiz, J., Gumpenberger, C., Jack, K., & Kraker, P. (2014). Comparison of downloads, citations and readership data for two information systems journals. *Scientometrics*, 101, 1113-1128.
- Schnell, J. D. (2017). Web of Science: The first citation index for data analytics and scientometrics. In F. J.

- Cantu-Ortiz (Ed.), Research analytics: Boosting university productivity and competitiveness through scientometrics (pp. 15–29). Abingdon: Taylor & Francis. DOI: <https://doi.org/10.1201/9781315155890-2>
- Schnitzler, K., Davies, N., Ross, F., & Harris, R. (2016). Using Twitter™ to drive research impact: A discussion of strategies, opportunities and challenges. *International Journal of Nursing Studies*, 59, 15–26. <https://doi.org/10.1016/j.ijnurstu.2016.02.004>
- Sebastian-Coleman, L. (2012). *Measuring data quality for ongoing improvement: a data quality assessment framework*. Newnes.
- Serrano-López, A. E., Ingwersen, P., & Sanz-Casado, E. (2017). Wind power research in Wikipedia: Does Wikipedia demonstrate direct influence of research publications and can it be used as adequate source in research evaluation? *Scientometrics*, 112(3), 1471–1488. <https://doi.org/10.1007/s11192-017-2447-2>
- Shema, H., Bar-Ilan, J., & Thelwall, M. (2012). Research blogs and the discussion of scholarly information. *PloS one*, 7(5), e35869.
- Shema, H., Bar-Ilan, J., & Thelwall, M. (2014). Do blog citations correlate with a higher number of future citations? Research blogs as a potential source for alternative metrics. *Journal of the Association for Information Science and Technology*, 65(5), 1018–1027. <https://doi.org/10.1002/asi.23037>
- Shema, H., Bar-Ilan, J., & Thelwall, M. (2015). How is research blogged? A content analysis approach. *Journal of the Association for Information Science and Technology*, 66(6), 1136–1149. <https://doi.org/10.1002/asi.23239>
- Singh, A., & Singh, H. (2016). Knowledge Publishing: An Integral Component of Knowledge Management. *International Journal of Library and Information Studies*, 6(4), 1-9.
- Singh, V. K., Singh, P., Karmakar, M., Leta, J., & Mayr, P. (2021). The journal coverage of Web of Science, Scopus and Dimensions: A comparative analysis. *Scientometrics*, 126(6), 5113–5142. <https://doi.org/10.1007/s11192-021-03948-5>
- Solutions. Available online: <https://help.altmetric.com/support/solutions/> (accessed on 1 Aug 2023)
- Stacy, K., Natalia, M., & Amy, R. (2016). *Altmetrics for librarians: 100+ tips, tricks, and examples*. <https://doi.org/10.6084/m9.figshare.3749838.v2>
- Stoica, P., & Selen, Y. (2004). Model-order selection: a review of information criterion rules. *IEEE Signal Processing Magazine*, 21(4), 36-47.
- Subbiah, V. (2023). The next generation of evidence-based medicine. *Nature medicine*, 29(1), 49-58.
- Sud, P., & Thelwall, M. (2014). Evaluating altmetrics. *Scientometrics*, 98(2), 1131–1143. <https://doi.org/10.1007/s11192-013-1117-2>
- Suzuki, Y. (2017). Validity of new measures of implicit knowledge: Distinguishing implicit knowledge from automatized explicit knowledge. *Applied Psycholinguistics*, 38(5), 1229-1261.
- Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly use of social media and altmetrics: A review of the literature. *Journal of the Association for Information Science and Technology*, 68(9), 2037–2062. <https://doi.org/10.1002/asi.23833>

- Swift, L., & Piff, S. (2014). *Quantitative methods: for business, management and finance*. Bloomsbury Publishing.5
- Szomszor, M., & Adie, E. (2022). Overton: A bibliometric database of policy document citations. *Quantitative Science Studies*, 3(3), 624–650. https://doi.org/10.1162/qss_a_00204
- Tananbaum, G. (2013). Article-level metrics: a SPARC primer. Retrieved July 20, 2020, from <http://www.sparc.arl.org/sites/default/files/sparc-alm-primer.pdf>
- Tahamtan, I., & Bornmann, L. (2020). Altmetrics and societal impact measurements: Match or mismatch? a literature review. *Profesional de La Informacion*, 29(1), 1–30. <https://doi.org/10.3145/epi.2020.ene.02>
- Thelwall, Michael. (2016). Web Indicators for Research Evaluation: A Practical Guide. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 8(4), i–155. <https://doi.org/10.2200/s00733ed1v01y201609icr052>
- Thelwall, Mike. (2017). Three practical field normalised alternative indicator formulae for research evaluation. *Journal of Informetrics*, 11(1), 128–151. <https://doi.org/10.1016/j.joi.2016.12.002>
- Thelwall, Mike. (2018). Altmetric Prevalence in the Social Sciences, Arts and Humanities: Where are the Online Discussions? *Journal of Altmetrics*, 1(1), 1–12. <https://doi.org/10.29024/joa.6>
- Thelwall, M. (2018). Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis. *Journal of Informetrics*, 12(1), 1–9. <https://doi.org/10.1016/j.joi.2017.11.001>.
- Thelwall, Mike. (2020). The Pros and Cons of the Use of Altmetrics in Research Assessment. *Scholarly Assessment Reports*, 2(1). <https://doi.org/10.29024/sar.10>
- Thelwall, Mike. (2021). *MEASURING SOCIETAL IMPACTS OF RESEARCH WITH ALTMETRICS ? COMMON PROBLEMS AND MISTAKES*. 35(5), 1302–1314. <https://doi.org/10.1111/joes.12381>
- Thelwall, Mike, Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, 8(5), e64841. <https://doi.org/10.1371/journal.pone.0064841>
- Thelwall, Mike, & Nevill, T. (2018). Could scientists use Altmetric.com scores to predict longer term citation counts? *Journal of Informetrics*, 12(1), 237–248. <https://doi.org/10.1016/j.joi.2018.01.008>
- Thelwall, Mike, & Wilson, P. (2016). Mendeley readership altmetrics for medical articles: An analysis of 45 fields. *Journal of the Association for Information Science and Technology*, 67(8), 1962–1972. <https://doi.org/10.1002/asi.23501>
- TODESCHINI, R. 2011. The j-index: a new bibliometric index and multivariate comparisons between other common indices. *Scientometrics*, 87.
- Torres-Salinas, D., Cabezas-Clavijo, Á., & Jiménez-Contreras, E. (2013). Altmetrics: New indicators for scientific communication in Web 2.0. *Comunicar*, 21(41), 53–60. <https://doi.org/10.3916/C41-2013-05>
- Torres-Salinas, D., Robinson-Garcia, N., & Gorraiz, J. (2017). Filling the citation gap: measuring the multidimensional impact of the academic book at institutional level with PlumX. *Scientometrics*, 113(3), 1371–1384. <https://doi.org/10.1007/s11192-017-2539-z>

- Van Eck, N. J., Waltman, L., van Raan, A. F., Klautz, R. J., & Peul, W. C. (2013). Citation analysis may severely underestimate the impact of clinical research as compared to basic research. *PLoS one*, 8(4), e62395.
- VAN Raan, A. F. J. (2005). Fatal attraction: Conceptual and methodological problems in the ranking of universities by bibliometric methods. *Scientometrics*, 62, 133-143.
- van Raan, A. F., & Winnink, J. J. (2019). The occurrence of ‘Sleeping Beauty’ publications in medical research: Their scientific impact and technological relevance. *PLoS One*, 14(10), e0223373.
- Victora, C. G., Black, R. E., Boerma, J. T., & Bryce, J. (2017). Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. *The lancet*, 377(9759), 85-95.
- VINKLER, P. (2009). The π -index: a new indicator for assessing scientific impact. *Journal of Information Science*, 35, 602-612.
- Visser, M., Eck, N. J. Van, & Waltman, L. (2021). Large-scale comparison of bibliographic data sources : Scopus , Web of Science , Dimensions , Crossref , and Microsoft Academic. <https://doi.org/10.1162/qss>
- Wagner, C. & Leydesdorff, L. (2012). An Integrated Impact Indicator (I3): A new definition of “Impact” with Policy Relevance. *Research Evaluation*, 21(3):183-188
- Waltman, L., Costas, R., & Van Eck, N. (2012). Some limitations of the h index: A commentary on Ruscio and colleagues' analysis of bibliometric indices. *Measurement: Interdisciplinary Research and Perspectives*, 10(3), pp.172-175.
- Wang, Jiabo, Ma, Z., Niu, M., Zhu, Y., Liang, Q., Zhao, Y., Song, J., Bai, Z., Zhang, Y., Zhang, P., Li, N., Meng, Y., Li, Q., Qin, L., Teng, G., Cao, J., Li, B., Chen, S., Li, Y., ... Xiao, X. (2015). Evidence chain-based causality identification in herb-induced liver injury: exemplification of a well-known liver-restorative herb *Polygonum multiflorum*. *Frontiers of Medicine*, 9(4), 457–467. <https://doi.org/10.1007/s11684-015-0417-8>
- Wang, Justin, Alotaibi, N. M., Ibrahim, G. M., Kulkarni, A. V., & Lozano, A. M. (2017). The Spectrum of Altmetrics in Neurosurgery: The Top 100 “Trending” Articles in Neurosurgical Journals. *World Neurosurgery*, 103, 883-895.e1. <https://doi.org/10.1016/j.wneu.2017.04.157>
- Wang, W., Liu, W., & Mingers, J. (2015). A systemic method for organisational stakeholder identification and analysis using Soft Systems Methodology (SSM). *European Journal of Operational Research*, 246(2), 562-574.
- Wang, X., Fang, Z., & Yang, Y. (2014). *Continuous, Dynamic and Comprehensive Article-Level Evaluation of Scientific Literature*. arxiv:1411.7004v1
- Wang, X., Fang, Z., & Wang, H. (2015). Study on the Continuous, Dynamic and Comprehensive Article-Level Evaluation System. *Science of Science and Management of S. & T.* 2015(8): 37-48.
- Wang, X., Mao, W., Xu, S., & Zhang, C. (2014). Usage history of scientific literature: Nature metrics and metrics of Nature publications. *Scientometrics*, 98(3), 1923–1933. <https://doi.org/10.1007/s11192-013-1167-5>
- Wanzer, D. L. (2021). What is evaluation?: Perspectives of how evaluation differs (or not) from research. *American Journal of Evaluation*, 42(1), 28-46.

- Weinstock, M. (1971). *Encyclopedia of Library and Information Science*, volume 5, Chapter: Citation indexes. Dekker, New York, NY.
- Weller, K. (2015). Social media and altmetrics: An overview of current alternative approaches to measuring scholarly impact. In *Incentives and Performance: Governance of Research Organizations* (pp. 261–276). Springer International Publishing. https://doi.org/10.1007/978-3-319-09785-5_16
- WEST, R. & MCILWAINE, A. 2002. What do citation counts count for in the field of addiction? An empirical evaluation of citation counts and their link with peer ratings of quality *Addiction*, 97, 501-504.
- Westland, J. C. (2015). Structural equation models. *Stud. Syst. Decis. Control*, 22(5), 152.
- Willis, C. D., Riley, B., Stockton, L., Viehbeck, S., Wutzke, S., & Frank, J. (2017). Evaluating the impact of applied prevention research centres: Results from a modified Delphi approach. *Research Evaluation*, 26(2), 78–90. <https://doi.org/10.1093/reseval/rvx010>
- Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S. & Johnson, B. (2015). The metric tide: report of the independent review of the role of metrics in research assessment and management bristol. UK: Higher Education Funding Council for England (HEFCE)
- WIPO. (2014). *National Collections - Data Coverage*. Retrieved August 1, 2024, from https://patentscope.wipo.int/search/en/help/data_coverage.jsf
- Woelert, P. (2013). The ‘Economy of memory’: publications, citations, and the paradox of effective research governance. *Minerva*, 51(3), 341–362. <http://dx.doi.org/10.1007/s11024-013-9232-z>
- Wong, E. Y., & Vital, S. M. (2017). PlumX: a tool to showcase academic profile and distinction. *Digital Library Perspectives*, 33(4), 305–313. <https://doi.org/10.1108/DLP-12-2016-0047>
- World Health Organization (2024). *Clinical trials - overview*. Health Topics. Retrieved June 20, 2024, from https://www.who.int/health-topics/clinical-trials#tab=tab_1
- Wouters, P., & Costas, R. (2012). Users , narcissism and control – tracking the impact of scholarly publications in the 21 st century. *Image Rochester NY*, 50 pages.
- 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 & Technology*, 61(3), pp.609–614.
- Wu, S., & Zhao, R. (2016). Research on Status and Development Trend of Altmetrics Tools. *Documentation, Information & Knowledge*, 2016(01):84-93.
- Xia, F., Su, X., Wang, W., Zhang, C., Ning, Z., & Lee, I. (2016). Bibliographic analysis of Nature based on Twitter and Facebook altmetrics data. *PLoS ONE*, 11(12), 1–14. <https://doi.org/10.1371/journal.pone.0165997>
- Xu, F., Li, X. X., Meng, W., Liu, W. B., & Mingers, J. (2013). Ranking academic impact of world national research institutes-by the chinese academy of sciences. *Research Evaluation*, 22(5), 337–350. <https://doi.org/10.1093/reseval/rvt007>
- Xu, S. (2018). Issues in the Interpretation of “Altmetrics” Digital Traces: A Review. *Frontiers in Research Metrics and Analytics*, 3(October), 1–8. <https://doi.org/10.3389/frma.2018.00029>
- Yang, S. Q., & Dawson, P. H. (2018). Altmetrics and Their Potential as an Assessment Tool for Digital Libraries.

IEEE 5th International Symposium on Emerging Trends and Technologies in Libraries and Information Services, ETTLIS 2018, 351–354. <https://doi.org/10.1109/ETTLIS.2018.8485273>

- Yang, L., & Chen, M. (2015). Comparison study of the common used altmetrics tools. *Information Studies: Theory & Application*, 38(9), 114-119.
- Ye, F. & Leydesdorff, L. (2014). The “academic trace” of the performance matrix: A mathematical synthesis of the h-index and the integrated impact indicator (I3). *Journal of the Association for Information Science and Technology*, 65(4), pp.742-750.
- Ye, F., & Rousseau, R. (2010). Probing the h-core: an investigation of the tail–core ratio for rank distributions. *Scientometrics*, 84(2), 431-439.
- Yom, S. S. (2023). The Top 20 Cited and Downloaded Red Journal Articles of 2022. *International Journal of Radiation Oncology, Biology, Physics*, 117(3), 529-531
- Yong, A. G., & Pearce, S. (2013). A beginner’s guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in quantitative methods for psychology*, 9(2), 79-94.
- Yu, H. (2017). Context of altmetrics data matters: an investigation of count type and user category. *Scientometrics*, 111(1), 267–283. <https://doi.org/10.1007/s11192-017-2251-z>
- Yu, H., & Biekezhati, M. (2020). Interpreting the Latest Development of Altmetrics Research from ISSI2019 Conference. *Information Studies: Theory & Application*, 2020-01-09
- Yu, H., Cao, X., Xiao, T., & Yang, Z. (2020). How accurate are policy document mentions? A first look at the role of altmetrics database. *Scientometrics*, 125(2), 1517–1540. <https://doi.org/10.1007/s11192-020-03558-7>
- Yu, H., Murat, B., Li, J. *et al.* (2023). How can policy document mentions to scholarly papers be interpreted? An analysis of the underlying mentioning process. *Scientometrics*, **128**, 6247–6266. <https://doi.org/10.1007/s11192-023-04826-y>
- Yu, H., Xu, S., Xiao, T., Hemminger, B. M., & Yang, S. (2017). Global science discussed in local altmetrics: Weibo and its comparison with Twitter. *Journal of Informetrics*, 11(2), 466–482. <https://doi.org/10.1016/j.joi.2017.02.011>
- Zahedi, Z., & Costas, R. (2018). General discussion of data quality challenges in social media metrics: Extensive comparison of four major altmetric data aggregators. *PLoS ONE*, 13(5). <https://doi.org/10.1371/journal.pone.0197326>
- Zahedi, Z., Costas, R., & Wouters, P. (2014). How well developed are altmetrics? A cross-disciplinary analysis of the presence of ‘alternative metrics’ in scientific publications. *Scientometrics*, 101(2), 1491–1513. <https://doi.org/10.1007/s11192-014-1264-0>
- Zahedi, Z., Fenner, M., & Costas, R. (2015). *Consistency among altmetrics data provider/aggregators: what are the challenges?* 677(October), 5–7.
- Zahedi Z, Fenner M, Costas R. How consistent are altmetrics providers? Study of 1000 PLOS ONE publications using the PLOS ALM, Mendeley and Altmeteric. com APIs[C]//*altmetrics 14. Workshop at the Web Science Conference*, Bloomington, USA. 2014.

- Zhang, C. (2009). The e-Index, complementing the h-Index for Excess Citations. *PLoS ONE*, 4(5), pp.1-4.
- Zhang, X., Wang, X., Zhao, H., Ordóñez de Pablos, P., Sun, Y., & Xiong, H. (2019). An effectiveness analysis of altmetrics indices for different levels of artificial intelligence publications. *Scientometrics*, 119(3), 1311–1344. <https://doi.org/10.1007/s11192-019-03088-x>
- Zhao, R., & Wang, X. (2019). Evaluation and comparison of influence in international Open Access journals between China and USA. *Scientometrics*, 120(3), 1091–1110. <https://doi.org/10.1007/s11192-019-03159-z>

Appendix 1: The comprehensive schematic diagram of the generation of research impact for publications (large print)

