WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

Bachelor Thesis

Scientometric Evaluation of Organizational Research Performance: Developing Organizational-Level Indicators for WU PURE

Philip Gruchow

Date of Birth: 16.05.2003
Student ID: 12020313

Subject Area: Information Systems

Studienkennzahl: 033 561

Supervisor: Prof. Dr. Axel Polleres
Co-Supervisor: Dr. Daniel Szöllösi

Date of Submission: 20.05.2026

Department of Information Systems & Operations Management, Vienna University of Economics and Business, Welthandelsplatz 1, 1020 Vienna, Austria

DEPARTMENT FÜR WIRTSCHAFTSINFORMATIK UND OPERATIONS MANAGEMENT DEPARTMENT OF INFORMATION SYSTEMS AND OPERATIONS MANAGEMENT

Contents

1 Introduction **5**
1.1 Motivation 5
1.2 Research Problem 6
1.3 Research Questions 8

2 Theoretical Background **9**
2.1 Foundations of Scientometrics 9
2.2 The h-index 11
2.3 The g-index 14
2.4 Organizational Scientometrics 16
2.5 CRIS Systems and PURE WU 18

3 Methodology **20**
3.1 Data Sources 20
3.2 Data Preparation 21
3.3 Attribution Model: Full Counting (Union Approach) 23
3.4 Organizational Levels 25
3.5 Computation of h-index and g-index 26
3.6 Limitations of the method 28

4 Empirical Results **29**
4.1 Dataset Overview 29
4.2 Child-Level Results 31
4.3 Institute-Level Results 33
4.4 Department-Level Results 34
4.5 Observations 35

5 Discussion **37**
5.1 Adaptation of h-index/g-index to Organizational Units 37
5.2 Implications of Full Counting 38
5.3 Impact of Meta Data Quality 40
5.4 Limitations of SCOPUS Data 41
5.5 Practical Implications for WU 42

6 Conclusion **43**
6.1 Summary of Research Objective and Approach 43
6.2 Summary of Key Findings 44
6.3 Theoretical and Practical Contributions 44
6.4 Limitations and Future Research 45

List of Figures

1 h-index and g-index at institute and department levels 35

List of Tables

1 Running example: h-index computation 11
2 Running example: g-index computation 14
3 Child-Level Scientometric Indicators 32
4 Institute-Level Scientometric Indicators 33
5 Department-Level Scientometric Indicators 34

Abstract

Universities increasingly rely on quantitative indicators to evaluate research performance, yet most scientometric indices were designed for individual researchers. This thesis develops a methodological framework for computing organizational-level h-index and g-index values using publication and citation data from the PURE research information system at the Vienna University of Economics and Business (WU). Publications are attributed to organizational units through author affiliation metadata using a full counting approach, which preserves the original computational logic of both indices by constructing complete, unpartitioned publication lists for each unit. Scientometric indicators are computed at three hierarchical levels: child units, institutes, and departments. The empirical results show that g-index values consistently exceed h-index values across all organizational levels, reflecting the sensitivity of the g-index to highly cited publications. Hierarchical aggregation systematically increases indicator values, as larger publication sets raise the probability that multiple publications exceed the respective citation thresholds. The analysis further demonstrates that computed indicators depend on the quality of affiliation metadata and the coverage of the SCOPUS citation database, which may underrepresent certain disciplines and publication types. The findings confirm that organizational h-index and g-index values can be computed from institutional research information systems, while highlighting the importance of transparent attribution rules and careful interpretation when applying individual-level indices to collective entities.

1 Introduction

1.1 Motivation

With the ever-growing importance of data, universities face an increased need to deliver competitive research performance. The quantity and quality of research output have an influence on the reputation of a university, its strategic development, and funding allocation, making research evaluation an important governance instrument. The increasing reliance on quantitative indicators of research performance allows for a structured and comparable assessment between individuals or organizational units. This enables the evaluation to be more objective and metric-supported, rather than purely qualitative. The field of scientometrics provides the basis for this analysis, defined as the study of science, technology, and innovation from a quantitative perspective [1]. Within the context of academic research performance evaluation, scientometrics provides a methodological basis for evaluation by focusing on publications and citations. In the field of academic scientometrics, many indicators were originally designed for individuals. However, universities increasingly require performance evaluation at the organizational level, creating a shift from individual researchers to collective entities.

The Vienna University of Economics and Business (which will be referred to as WU for the remainder of the thesis) serves as an academic institutional setting, therefore relying on structured research information systems. Research evaluation plays an important role in internal reporting and planning within the university. The Current Research Information System (CRIS) used by WU is "PURE", which stores publication metadata, author information, organizational affiliations, and citation data, which is integrated using SCOPUS data. SCOPUS is a commercial database for peer-reviewed literature [2]. PURE WU is currently used primarily as a data management and reporting tool by providing statistics. PURE WU provides publication lists, as well as information on individual researchers or organizational units, which includes publications, citations, and other descriptive statistics. The system currently does not offer built-in computation of any specific scientometric performance indicators, beyond publication and citation counts. The lack of performance indicators causes the need for a standardized method of aggregating performance indices at the department or institute level, which requires a methodological approach to organizational performance evaluation.

The h-index and the g-index are commonly used scientometric indicators that

aim to combine aspects of research productivity, as well as research impact in the form of citations. At their core, both indices were designed for individual researchers [3]. They assume a single identifiable entity and a clearly defined set of publications, which poses some challenges at the organizational level. At an organizational level the entity consists of multiple individuals, introducing the structural shift from individual to collective, furthermore the publication set becomes less clearly bounded. The h-index and g-index are both non-linear as they cannot be summed or averaged, therefore, organizational indices must be computed separately and require some methodological definitions. Publications often involve multiple authors, which sometimes have multiple affiliations that can also change over time, which raises attributional questions such as: Which unit receives credit? Should multiple units receive credit?. Different approaches to attribution rules can produce different results, which introduces a degree of complexity. Therefore, applying individual-level indices to organizational units requires methodological clarification, creating the central research problem.

The current CRIS at WU poses a solid foundation for the research problem, as PURE WU stores the necessary data, citation information, as well as a clear definition of organizational structures. However, due to the lack of a computed h-index, g-index, and a standardized attribution methodology, a direct organizational-level scientometric evaluation is currently not supported, which poses a central gap. Organizational units require structured evaluation, as decision making is based on comparable performance indicators, thus creating the need for a methodological framework. The challenge is not only in the computation of performance indices, but also in the conceptual adaptation of scientometric indices, which leads to the research problem addressed in this thesis.

1.2 Research Problem

The h-index was originally proposed by J.E. Hirsch in 2005 as "an index to quantify an individual's scientific research output" [4]. A year later, Egghe proposed the g-index as an improvement to the previously accepted h-index, putting more emphasis on the exponential nature of highly influential academic works [5]. Both indices evaluate a single author based on their publication-citation record, while assuming a clearly bounded publication set tied to one entity. Organizational units are not single entities, but rather made up of multiple researchers and have fewer clear boundaries of their publication sets, while also being dynamic and hierarchal. These characteristics of organizational units pose problems when attempting to apply

the original h-index or g-index, as they cannot simply be aggregated from individual values, due to their non-linear nature. The core conceptual research problem therefore lies not only in the computational integration of indices but also in the methodological adaptations of scientometric indicators.

Both the h-index and the g-index are non-linear measures, as their values depend on the distribution of citations, rather than simply considering the total number of citations an individual has received [4]. Furthermore, they are threshold-based indicators, which means that they measure whether certain citation conditions (thresholds) are met, depending on which index is used. This results in a simple aggregation being not a viable solution, as individual h-indices or g-indices cannot be summed or averaged meaningfully. For example, two researchers with an h-index of 10 respectively do not imply an institute or department h-index of 20. Furthermore, aggregating individual index scores is prone to citation overlap and distribution issues. Therefore, an organizational h-index or g-index must be calculated from the entire set of publications of the unit, which in turn requires a definition regarding which publications belong to a unit, respectively, and how individuals with multiple affiliations are handled.

Academic publications are often co-authored by multiple different individuals, in which case the authors may belong to the same organizational unit; however, they may also belong to different organizational units or different institutions entirely. This creates overlapping credit, as one publication now technically belongs to multiple individuals or organizational units. In addition, authors may hold multiple affiliations simultaneously, which can span different institutes, departments, or even universities. Therefore, it can become unclear which organizational unit should receive credit for the publication. Affiliations are also prone to change over time, introducing the complexity regarding historical attribution, as publications can be produced during one affiliation but be published and subsequently recorded under another affiliation. As organizational attribution is not self-evident, different attribution rules can be implemented, such as full counting or fractional counting, leading to different outcomes. Without clearly defined rules, organizational indices lack comparability, thus creating the need for a transparent attribution framework.

PURE WU stores publication metadata and affiliation information, as well as citation data sourced from SCOPUS. Currently, PURE WU is used for documentation and reporting, rather than methodological indicator computation. For organizational attribution to function reliably, accurate affilia-

tion metadata is crucial. Furthermore, the quality of SCOPUS citation data is fundamental for the calculation of organizational performance indicators later. The current PURE WU system poses multiple challenges regarding the computation of organizational level scientometric indicators, as there is currently no built-in standardized solution and data constraints must be considered. A structured and transparent methodological approach is required, as the research problem lies in adapting individual level indicators to organizational level indicators under real world data constraints. This forms the basis of the research questions addressed in the next section.

1.3 Research Questions

To address the challenges described in the previous section and develop a scientifically grounded method for assessing organizational units at WU, the following research questions guide this thesis:

Main research question:

- How can scientometric indicators potentially be adapted from simple author metrics such as the h-index and accurately computed for organizational units at WU using SCOPUS-based data stored in PURE or other external, publicly available sources?

Sub-questions:

1. How can individual-centric indicators such as the h-index or g-index be adapted for organizational-level evaluation?
2. How should publications be attributed to organizational units, particularly in cases of multi-authorship, multi-affiliation, and affiliation changes over time?
3. How does the quality of metadata in PURE (affiliations, duplicates, citation completeness) influence the results of unit-level scientometric indicators?
4. What limitations arise from relying on SCOPUS data for organizational scientometrics, and which other additional sources could be considered?

2 Theoretical Background

2.1 Foundations of Scientometrics

Scientometrics as a field of study was first introduced in 1971 in order to develop quantitative methods of research on the development of science as an informational process [6]. The focus is on the measurement of the research output and impact in a quantitative manner, grounded in bibliometrics, which is concerned with the estimation of the impact and/or influence of research articles on future work [7]. Publications are the primary units of scientific communication, which makes them an essential part of the scientific community. Citations of scientific publications can therefore be considered as indicators of the impact or influence of the publication. Scientometrics focuses on quantifiable patterns within the indicators used to measure scientific output. Scientometrics, bibliometrics and informetrics are often compared because they are related in nature [8]. While Scientometrics is concerned with applying quantitative methods to analyze and evaluate scientific activity, bibliometrics studies measurable patterns in scientific publications, whereas informetrics studies quantitative patterns in information, regardless of domain [8]. Scientometrics is used mainly in research evaluation and supports institutional decision-making in an academic context. Furthermore, scientometrics provides the quantitative tools required for appropriate performance analysis, through its logic involved in measurement.

Scientometrics fundamentally focuses on two dimensions of scientometric measurement. Scientific output can be measured through publication counts, which can also be referred to as the productivity of a researcher or entity. The second dimension views scientific output through citation counts, which can be viewed as the impact of a publication. It is important to note that these two dimensions are not interchangeable, as the quantity and influence of an entity are distinct dimensions [9]. Citations are used as a key metric within scientometrics as they reflect acknowledgment by peers to a certain extent and are therefore used as a proxy for scientific influence. Although citations are not entirely perfect as an indicator, they are widely accepted due to their measurability and comparability. When looking at citation counts as an indicator, the structural properties of citation data must be considered, as they tend to be skewed [10]. Few publications receive many citations, while the majority of publications receive few or none. Therefore, the distribution of citations plays an larger role in determining the impact of an entity than the total number of citations. As simple citation counts are often insufficient for a balanced scientometric evaluation, metrics that combine productivity

and impact are needed [4]. Composite indices attempt to integrate both dimensions, such as threshold-based indicators such as the h-index or the g-index.

When looking at the applications and institutional relevance of scientometric indicators, it can be said that they are widely applied in research evaluation, as there is an increasing demand for accountability and transparency in academic communities [11]. Furthermore, quantitative metrics complement peer review and add a more objective dimension, which can be used to create structured comparability between entities [11]. Scientometric indicators were originally designed for individuals, but can be applied to research groups, such as departments or even at institutional or national levels [9]. If a methodologically sound approach is used, the publication and citation data enable comparisons between different organizational units. These indicators can be used as a tool for strategic planning and internal monitoring, which can influence the allocation of funding in universities [9]. In addition, indicators can be incorporated into ranking systems at a national or international level or integrated into broader research policy frameworks [11]. Extending metrics originally designed for individuals to collective entities raises methodological questions, and aggregation requires clear conceptual and procedural definitions.

Scientometrics as a field of study is debated and critics have expressed some concerns [11]. Quantitative indicators often simplify complex research processes, and metrics are only able to capture certain dimensions of scientific performance. Although they offer a more objective approach to performance evaluation, there is a risk of over-reliance on numerical indicators, as evaluation context plays a role in the interpretation of scientific publications. One of the two main scientometric dimensions, citation counts, is also field-dependent, as citation behavior varies between different fields [10]. Furthermore, there is also the element of time-dependent accumulation of citations, which naturally favors older publications that have had more time to accumulate citations [3]. In addition, self-citations may further distort results, as this can result in artificially inflated citation counts [3]. Beyond the structural issues in citation-based evaluation, there are also broader methodological challenges that must be considered. Citation distributions are skewed and indicators may favor larger organizational units [10]. Furthermore, quantitative metrics may obscure qualitative contributions, which can cause problems with comparability between entities. When applied beyond individuals, these scientometric limitations become more complex, and questions regarding aggregation and attribution further complicate interpretation.

2.2 The h-index

The h-index was proposed by Jorge E. Hirsch in 2005 as "an index to quantify an individual's scientific research output" [4]. The index was developed to quantify both the productivity and the impact of the citation of individual researchers, with the intention of being a simple and robust performance indicator. Furthermore, the proposed h-index was designed as an alternative to the total number of citations that had previously been used. The h-index is best explained with an example: a researcher has an index h if h publications each received at least h citations. The remaining publications (if existing) have equal to or less than h citations. This makes the h-index threshold based by definition based on the publication of a certain citation threshold. In order for the index to be computable, a ranking of publications by citation count is required. Conceptually, the h-index captures a combination of productivity and impact, requiring sustained citation performance [3]. The index aims to balance both output and influence of publications, as the index is not determined by a single highly cited publication but rather a distributed citation performance across a number of publications.

To illustrate the computation of the h-index, consider a fictional researcher with 10 publications ranked by citation count, as shown in Table 1.

Table 1: Running example: publications of a fictional researcher ranked by citation count. The h-index equals the largest rank h for which the citation count is at least h. Here, $h = 6$, as the 6th-ranked publication has exactly 6 citations, while the 7th has only 4.

Rank	Citations	Citations $\geq$ Rank?	
1	45	$45 \geq 1$	✓
2	20	$20 \geq 2$	✓
3	15	$15 \geq 3$	✓
4	12	$12 \geq 4$	✓
5	8	$8 \geq 5$	✓
6	6	$6 \geq 6$	✓
7	4	$4 \geq 7$	×
8	3	$3 \geq 8$	×
9	1	$1 \geq 9$	×
10	0	$0 \geq 10$	×

The researcher's h-index is 6, as 6 publications have each received at least 6 citations. Notably, the most-cited publication (45 citations) does not in-

crease the h-index beyond this value, as the index only reflects whether the citation threshold is met, not by how much it is exceeded. This insensitivity to highly cited publications is a key structural property of the h-index and motivates the development of the g-index, discussed in the following section.

The structural mechanisms of the h-index must also be considered. The h-index of a researcher increases only when a certain citation threshold is crossed, which results in the non-linear nature of the h-index [12]. Furthermore, additional citations do not always increase the index, which makes the index growth stepwise. The h-index is thus relatively insensitive to small citation increases below the required threshold. The characteristics of the h-index make it robust; for example, a highly cited article does not drastically increase the h-index, as sustained performance in multiple publications is required [12]. This naturally protects the h-index from distortion by single high citation outliers and creates stability. However, the h-index also shows a level of structural dependence, as it favors longer publication histories, which are generally correlated with total research output [4]. Therefore, researchers who possess a larger set of publications allow for higher possible h values, as the h-index does not normalize for the career length. The h-index mainly reflects the cumulative performance of the citation, with the aim of balancing productivity and impact over time. It does not measure the average quality of publications and is not directly equivalent to the total influence of a researcher.

The h-index has become widely used and popular within the academic community [12]. The conceptual attractiveness of the h-index is due to a variety of reasons. The index is simple to compute and easy to interpret, creating low barriers to entry [4]. Furthermore, the index combines productivity and impact into a single interpretable number, which was previously not possible. The transparent definition of the index that does not include complex weighting has also contributed to its acceptance and popularity. The h-index is generally perceived as balanced as it avoids the pure publication count or the pure citation count. Additionally, the requirement of sustained citation performance, as well as the reduced distortion by single outliers, further contribute to its balance. The combination of these factors has led to the h-index being viewed as a compromise indicator. The h-index has also been institutionally adopted, such as the implementation of the index in major citation databases [12]. Furthermore, the index can be utilized in hiring or promotion contexts and is also easily comparable between individuals who operate in the same or similar fields. Despite known limitations of the index, it remains widely used due to its simplicity.

Although the h-index provides a variety of benefits, it also has some limitations and critiques. A limitation of the h-index is related to its field and its disciplinary dependence [12]. Citation practices differ between different disciplines, as certain disciplines have a larger total number of publications, while others have a higher citation density, resulting in higher h-values. Therefore, h-index values are not directly comparable across fields, as there is no field normalization implemented in the h-index. Furthermore, the h-index shows temporal bias as the index accumulates over time, in turn favoring senior researchers, while disadvantaging researchers in the early stages of their career [4]. The index does not account for the age of a publication and is therefore sensitive to citation windows. Although some consider the insensitivity of the h-index to highly cited publications to be a benefit, it can also pose certain limitations. Highly cited publications do not proportionally increase the h value and therefore do not reflect citation intensity beyond the threshold. This may lead to an underestimation of breakthrough contributions and has led to the development of alternative indices, such as the g-index [5]. Furthermore, the index faces limitations through challenges with aggregation and non-additivity [12]. The h-index cannot be averaged or summed meaningfully, as the non-linear structure of the index prevents additive aggregation. Thus, collective entities cannot be evaluated by purely summing individual h-values, therefore requiring recomputation from unified publication sets. These limitations become more pronounced at collective levels as application beyond individual researchers requires methodological clarification.

The h-index has a few implications when applied to the organizational level. Transfer from individual researchers to organizational units must be approached critically. The h-index was originally designed for individual researchers, but is increasingly applied to collective entities, such as research groups, departments, or other institutions [12]. Transfer is not straightforward; however, the structural properties remain unchanged at the organizational level. The core structural issue of the index's non-additivity means that organizational h-index values must be computed separately from a combined publication set attributed to that organizational unit, with author overlap complicating the aggregation [12]. The non-linear threshold structure remains at a group level. An organizational level h-index depends on how publication sets are defined as multi-affiliation and co-authorship create ambiguity. Thus, attribution rules influence the results, as there is no inherent solution to the h-index itself. Methodological clarification is required for collective application, creating the necessity of defined attribution models.

2.3 The g-index

The g-index was proposed by Leo Egghe in 2006 and was developed as a refinement of the h-index proposed the previous year [5]. The g-index is similar to the h-index; however, it is intended to address the limitations of the h-index, specifically designed to better account for highly cited publications. In order to compute the g-index of an individual, a list of publications ranked in descending order by citation count is necessary. Describing the index value is again easy with a real word example: A researcher has a g-index of g if the top g publications together received at least g^2 citations. The g-index, therefore, utilizes a cumulative citation threshold and places a stronger weight on highly cited publications. Conceptually, this is where the g-index separates itself from the h-index, as cumulative citations are used, which allows highly cited publications to increase the index further. The less restrictive threshold mechanism of the g-index makes it more sensitive to citation intensity.

Returning to the fictional researcher introduced in Section 2.2, Table 2 illustrates the computation of the g-index using the same set of publications.

Table 2: Running example: g-index computation for the same fictional researcher. The g-index equals the largest rank g for which the cumulative citations are at least g^2. Here, $g = 10$, as the cumulative citation total of 114 exceeds $10^2 = 100$.

Rank	Citations	Cumulative	g^2	Cumulative $\geq g^2$?
1	45	45	1	✓
2	20	65	4	✓
3	15	80	9	✓
4	12	92	16	✓
5	8	100	25	✓
6	6	106	36	✓
7	4	110	49	✓
8	3	113	64	✓
9	1	114	81	✓
10	0	114	100	✓

The g-index for this researcher is 10, compared to an h-index of 6. The difference is driven primarily by the highly cited first publication (45 citations), which contributes substantially to the cumulative citation total but does not influence the h-index beyond the threshold. This example demonstrates how the g-index captures citation intensity that the h-index does not reflect, particularly when a small number of publications account for a disproportionate

share of total citations.

When comparing the g-index and the h-index, it is important not to take into account the structural difference and measurement logic. The threshold mechanisms of the two indices differ significantly. Although the h-index uses individual publication thresholds ($>$ h citations each), the g-index utilizes cumulative citation thresholds ($> g^2$ total citations) [5]. The h-index counts qualifying publications individually, whereas the g-index aggregates citation strength across multiple top publications. The impact of the cumulative threshold structure of the g-index lies in the sensitivity to citation concentration. Highly cited publications contribute more strongly to the index value; therefore, the concentration of citations can significantly increase the value of the g-index. This allows the g-index to differentiate more at higher citation levels and has a less restrictive threshold than the h-index. Practically, this means that researchers with few highly cited publications have a higher g-index value relative to their h-index value. Furthermore, there is a greater spread between the two indices at high citation levels, since the g-index may be better suited to reflect citation intensity than the h-index, which allows for structural differentiation among high-impact researchers. Both indices measure citation performance differently; however, the structural distinction influences their interpretation.

Egghe proposed the g-index as a refinement of the h-index [5]. The h-index has been criticized for insensitivity to highly cited publications, therefore, the g-index was introduced to address this limitation. The g-index is intended to better reflect the intensity of the citation and was developed as a refinement of the h-index rather than a replacement of it. The core conceptual strength of the g-index is that strongly cited publications influence the index more, which allows better differentiation among high-performing researchers, allowing for the recognition of breakthrough contributions. In practice, this reduces the ties compared to the h-index and provides finer granularity at higher impact levels. This becomes particularly useful in highly cited disciplines. Although the g-index offers an alternative perspective on citation performance, structural complexity remains.

Although the g-index offers certain improvements over the h-index, some limitations and critiques remain. As with the h-index, the g-index also faces many of the same structural issues. The g-index is still field-dependent and time-dependent [12]. Furthermore, the index is also cumulative and sensitive to size with respect to publication sets, favoring larger organizational units. There is also no inherent normalization; thus the issue of a lack of cross-field

compatibility remains. In addition to the structural issues faced by both indices, the g-index also introduces new vulnerabilities. Strong sensitivity to highly cited publications can lead to an overestimation of citation concentration when measuring research performance. This can amplify the influence of significant outliers, which naturally results in greater volatility than the h-index. In an organizational context, issue with non-additivity and aggregation remains, as the g-index is non-additive and cannot be meaningfully averaged or summed. Therefore, separate computation from combined publication sets is required, with aggregation challenges remaining unchanged from the h-index. Generally, the g-index improves certain aspects of the h-index, while also introducing new complexities, and does not eliminate fundamental methodological concerns.

When applying the g-index to organizational level research performance, the shift from individual metrics to organizational use must be established. The g-index was originally designed for individuals but has become increasingly considered for collective entities. Like the h-index, the g-index is non additive and must be computed separately, which is the core structural constraint of both indices [12]. Overlapping authorship complicates aggregation and must be carefully considered. At an organizational level, highly cited publications can disproportionally affect the organizational g-index. With larger publication pools, this can increase the sensitivity of the g-index, as organizational size interacts with the citation concentration. Therefore, the risk of volatility increases at aggregated organizational levels. The application of the g-index to collective entities requires defined attribution rules, as aggregation and attribution become the central methodological issues.

2.4 Organizational Scientometrics

Research evaluation increasingly operates at organizational levels, as universities, departments, or research groups are collectively evaluated [9]. Institutional performance assessment plays an increasing role in governance, as aggregated metrics are used for benchmarking and comparison. This collective research performance evaluation can be referred to as organizational scientometrics. The shift towards organizational scientometrics is due to a variety of reasons, such as increasing accountability in higher education, the use of scientometrics in funding allocation mechanisms or their role in competitive ranking systems [11]. Furthermore, it can affect resource distribution at universities, which naturally creates the need for comparative performance metrics. Conceptually, organizational scientometrics differs from individual evaluation, as organizational entities differ conceptually from individuals.

Collective performance is not simply the sum of individuals, as aggregation introduces methodological challenges, and publication sets must be defined at the group level. The application of citation based indicators at an organizational level therefore requires methodological clarity and raises questions of aggregation and attribution.

Individual level scientometric indices cannot be directly transferred to organizational entities due to structural aggregation challenges. The discussed citation indicators have a core mathematical constraint, as both the h-index and the g-index are non-additive [12]. When shifting to organizational level metrics, the non-linear threshold structure of both indices persists. Furthermore, complexities regarding co-authorship are introduced, as publications often involve multiple authors that may belong to the same or different organizational units. The summation of individual indicators can lead to an inflation of the indices, and overlapping publication sets can distort aggregation. Organizational scientometrics involves methodological ambiguity, as organizational performance depends on the definition of the set of publications of the organizational unit. Multi-affiliation complicates attribution, while temporal affiliation changes influence inclusion. Neither of the discussed indices suggest an inherent definition; therefore, individual attribution rules directly influence the results of the computed indices. Organizational scientometrics requires explicit aggregation rules, and the structural properties of indices demand methodological clarity.

Organizational performance in the context of scientometrics depends primarily on how publication are assigned to organizational units [12]. Multi-authorship and multi-affiliation complication complicates attribution, hence defined attribution models are required, as without clear rules, indicators can become inconsistent. Publications may be assigned using different attribution approaches. The two most used approaches are full counting and fractional counting. These two approaches differ in how credit is distributed among authors or affiliations. Full counting assigns full credit to each contributing entity in a publication, whereas fractional counting distributes credit proportionally to contributing entities. The choice of attribution influences publication counts and citation indicators and, therefore, may produce substantially different results. Attribution models play a critical role in organizational scientometrics and therefore directly influence computed performance indicators.

Organizational scientometrics is not only a conceptual challenge but also a practical one. Implementing indicators at the organizational level requires

operational decisions. Real-world research information systems must translate theoretical indices into computable metrics [9]. Organizational publication rely on existing affiliation metadata; therefore, multi-affiliation and inconsistent affiliation reporting may complicate attribution. Institutional structures change over time, due to reorganizations or, for example, department changes. These issues influence how publications are attributed to organizational units. Organizational indices require recomputation from aggregated publication sets, with large datasets increasing computational complexity [12]. Index values may change significantly depending on the attribution rules. Furthermore, the sensitivity of the indicator increases on an organizational scale. These empirical and structural challenges show that organizational scientometrics lacks a commonly accepted straightforward implementation, which leads to the absence of standardized methodological approaches.

Scientometric indices were originally designed for individuals, with organizational applications remaining methodologically ambiguous [4]. Currently, there is no universally accepted framework for computing organizational h- or g-indices. Different studies adopt different attribution rules and aggregation procedures [12]. Methodological differences produce inconsistent results, as organizational level indicators depend heavily on attribution models and publication set definitions. The lack of standardization limits the comparability between different institutions. Due to this methodological gap, studies must clearly define their computational approach. This thesis adopts a defined methodological framework for computing organizational h- and g-indices.

2.5 CRIS Systems and PURE WU

CRIS stands for Current Research Information System. CRIS are information systems used by research institutions and are designed to collect, manage, and organize research related information [13]. CRIS systems manage publication metadata, author profiles, institutional affiliations, as well as research projects and funding. These datasets create a structure representation of an institution's research activity. Universities use CRIS to support research management. They enable reporting and monitoring of research output and provide transparency for institutional evaluation. CRIS systems provide the structured datasets required for scientometric analysis and enable the extraction of publication and citation data for evaluation purposes.

CRIS act as the data infrastructure required for research evaluation, as universities rely on data-driven evaluation of research performance [9]. CRIS

systems provide centralized datasets for institutional analysis as they consolidate publication and citation data required for the computation of performance indicators. CRIS often integrate data from external sources, such as "SCOPUS" or "Web of Science" [13]. These integrations allow institutions to track citations and research impact. CRIS allow for aggregation of research output at institutional or department levels, which enables the comparison of different organizational units and provides the basis for scientometric performance indicators. Many universities implement CRIS platforms for these purposes. One widely used CRIS platform is "PURE", which is used in WU.

PURE is a widely used CRIS, used internationally by universities and other research institutions, as it is designed to manage institutional research data [14]. PURE integrates external citation data sources, and the citation information being typically retrieved from SCOPUS. This allows citation-based evaluation of publications. WU uses PURE as its institutional CRIS, with the system storing publication records and organizational structures to an extent. The data set extracted from PURE WU provides the data basis for this thesis. Due to this structured research information infrastructure, the computation of scientometric indicators requires a clearly defined methodological approach.

Scientometric computation depends on the underlying system data structures [9]. CRIS datasets are based on structured metadata; therefore, the completeness and accuracy of the data influence the analysis results. Publication records and affiliations form the basis for scientometric indicators. Organizational structures are defined within the CRIS system, with publications being attributed based on recorded affiliations. Multi-affiliation and metadata inconsistencies can in turn influence attribution. Organizational scientometric indicators depend on the definition of publication sets; therefore, system constraints influence indicator computation, as results may vary accordingly. Due to these structural considerations, the calculation of organizational scientometric indicators requires a clearly defined methodological framework. The following section describes the methodological approach used in this study.

3 Methodology

3.1 Data Sources

The empirical analysis of this thesis is based on publication and citation data extracted from PURE WU, as PURE serves as the institutional CRIS used to manage research output. The data set contains publication records and associated citation information. The data set contains various components, such as publication metadata, citation counts, author identifiers, and organizational affiliations. These elements enable scientometric analysis at the organizational level. The data is used to construct publication sets for organizational units, as well as to compute scientometric indicators, which aim to enable the comparison of research performance across different organizational units at WU. The following paragraphs describe the publication, citation, and affiliation data used in the analysis.

The publication data set used for the analysis consists of peer-reviewed publication records stored in PURE WU. Each record contains structured publication metadata from various fields, such as title, publication year, document type, and authors involved. These records represent the scientific output used in the analysis. Citation data are included through citation counts that are linked to specific publication records. The citation data originate from the SCOPUS integration within PURE WU. Furthermore, they serve as the basis for impact measurement later, as they are required to compute scientometric indices, such as the h-index or the g-index. The data enables scientometric computation, as publications form the core unit of analysis used. Citation counts determine the ranking of publications, which is necessary as a ranking by citation counts enables the calculator of the h-index, as well as the g-index. In addition to publication and citation data, the data set includes affiliation information, which links publications and authors to organizational units within WU.

Organizational scientometric analysis requires the linking of publications to institutions or specific organizational units. This linkage is established through author affiliation data, which is stored within PURE WU, connecting authors and their respective publications to organizational units within the university. PURE WU stores the institutional organizational structure to an extent, meaning that authors are associated with specific organizational units. Therefore, publications inherit organizational attribution through their respective author affiliations, and in turn organizational affiliations. This allows publications to be assigned to institutes, departments, or other

organizational units in the analysis. Publications may be linked to multiple authors, which in turn may belong to different organizational units within WU. This relationship allows for the construction of organizational level publication sets. The following paragraph clarifies the scope and boundaries of the data set used for the analysis in this thesis.

The WU PURE system stores a wide range of research information. For this analysis, only publication records relevant to the scientometric analysis at hand are used, as this analysis focuses specifically on peer-reviewed research output recorded in the system. Publications serve as the main unit of analysis, and their citation counts are used to determine the relative impact of publications. These elements form the basis for computing the organizational level h- and g-indices. The analysis is limited to organizational units within WU, meaning that publication records are therefore considered only so far as they can be attributed to WU authors and organizational units. Before these indicators can be computed, the dataset must first be prepared and structured for the following analysis.

3.2 Data Preparation

Raw data extracted from CRIS systems often contains inconsistencies or incomplete records, which means that the raw data must be prepared and cannot be used directly. Raw datasets typically require cleaning and standardization prior to analysis. Data preparation ensures reliable computation of scientometric indicators for analysis. The cleaning of raw data involves the removal of incomplete or irrelevant records within the dataset, as well as the verification of publication metadata. Furthermore, consistency checks are implemented for key variables, such as publication year or authorship, to ensure systematic preparation. For this study specifically, scientometric indicators rely on accurate publication and citation data, meaning that errors in metadata or affiliations could significantly distort indicator computation. Therefore, preparing the data set is necessary before attribution and aggregation can be applied. The following paragraphs describe the specific data preparation steps applied to the dataset.

The imported publication data set from PURE WU contained column names originating from the original PURE export. These columns were standardized and renamed to a consistent naming scheme to further simplify the processing. Furthermore, key identifiers within the dataset, such as publication ID, person ID, publication year, and metric fields, were harmonized. The data set contained different metric types stored within the metric fields;

therefore, an inspection of the available metric names was performed to identify the relevant citation-based metric. The data set was then restricted to records that contained citation counts. Since citation values could appear in different formats within the data set, a helper function was implemented to extract numeric citation values from the metric field. This ensures that the citation counts were stored in a consistent numerical format, allowing them to be used for subsequent indicator computation. After standardized and extracted citation values, the data set was validated to ensure that the citation counts were successfully parsed and usable for further analysis.

Within the extracted PURE data, a separate data set containing affiliation data was extracted. The affiliation data set containing organizational PURE IDs and author affiliations was loaded for integration. This data set provides the link between individual researchers and their organizational units. Affiliation records include person identifiers and associated organizational identifiers, which allow the organization of the assignment of publications. First, identifier columns across data sets were standardized to a consistent string format, ensuring that identifiers such as publication IDs, person IDs, and organization IDs could be reliably matched across different data tables. This standardization of identifiers prevents mismatches during data set merging. Citation records at the publication-person level were merged with the affiliation data set. The merge operation used the person PURE ID as the linking variable, thus attaching organizational identifiers and affiliation information to each publication record. The resulting data set links publications, authors, and organizational identifiers within a single table. This enriched structure enables the subsequent assignment of publications to organizational units. The following paragraph describes how organizational identifiers were validated and completed to ensure consistent organizational assignment of publication records.

The affiliation data set contains organizational PURE IDs that link authors to organizational units. However, these identifiers are not consistently populated across all records, therefore missing organization identifiers would prevent reliable assignment of publications to organizational units. In turn, an effective organization identifier was constructed to increase coverage. When the primary Org Pure ID was missing, the alternative PURE ID field was used as a fallback. This procedure ensured that most of the affiliation records could still be associated with an organizational identifier. After constructing the effective organization identifier, coverage was evaluated. This was done by calculating the proportion of affiliation records that contain a valid identifier. This step ensured that organizational identifiers were available for nearly

all affiliation records. The effective organizational identifier was then merged into the publication-person data set. This allowed publications to be associated with organizational identifiers for subsequent organizational mapping. After establishing a reliable organizational identifier, publications could be assigned to specific organizational units within the university's structure.

After establishing reliable organization identifiers, the publications were assigned to child-level organizational units within WU's institutional structure. The child level represents the lowest organizational level available in the PURE organizational hierarchy, with the level serving as the primary unit for initial publication assignment. Only publication records with a valid child-level organizational assignment were retained for analysis. This filtering ensures that each publication included in the data set can be reliably attributed to an organizational unit. Afterwards, coverage statistics were calculated to determine the proportion of publications that have been successfully mapped. A deduplicated child-level publication table was constructed, where each publication was counted once per child-level organizational unit, even if multiple authors from the same organizational unit were involved. Duplicate combinations of publication and organization were removed to ensure consistent attribution. The resulting data set contains unique publication-organization pairs with associated citation counts, thus forming the basis for the subsequent computation of scientometric indicators. Based on this structured publication data set, scientometric indicators can be computed for different organizational levels.

3.3 Attribution Model: Full Counting (Union Approach)

Publications often involve multiple authors and organizational affiliations; therefore, they must be assigned to specific organizational units when evaluating research performance, thus requiring an explicit attribution model. The way publications are attributed influences publication counts, citation tools, and derived indicators. Without clearly defined attribution procedures, organizational indicators can become inconsistent or difficult to interpret. Consequently, this study adopts a clearly defined attribution approach for assigning publications to organizational units. The following paragraph introduces the attribution approach used in this study.

This study adopts a full counting attribution model. In full counting, each publication is fully attributed to every organizational unit associated with the publication. If multiple authors from different organizational units contribute to a publication, each unit receives full credit for the publication.

Publications may therefore contribute to multiple organizational units simultaneously. The same publication may appear in the publications sets of several organizational entities. Full counting therefore reflects collaborative research contributions without dividing publication credit across different units. Furthermore, full counting preserves the complete research output associated with each organizational unit. This allows organizational units to receive full recognition for their participation in collaborative research. As a result, publication sets represent the complete set of publications associated with a given organizational unit. The following paragraph discusses methodological considerations associated with the use of full counting in organizational scientometrics.

The choice of full counting is further motivated by the original design of the scientometric indicators used in this study. Both the h-index and the g-index were developed to operate on a complete, unpartitioned publication list belonging to a single entity [4, 5]. The h-index identifies a citation threshold across a ranked list of all publications, while the g-index evaluates cumulative citation counts across the same complete list. These computational procedures assume that each publication in the list contributes fully to the indicator value. Fractional counting would assign partial credit to publications, effectively reducing citation counts or publication weights before indicator computation. This would alter the input structure on which the indices were originally designed to operate. Full counting preserves the original computational logic of both indices by constructing complete publication lists for each organizational unit, treating the unit as the evaluated entity in the same way that Hirsch and Egghe treated an individual researcher. Therefore, full counting represents the most methodologically consistent approach when adapting these individual-level indicators to organizational entities.

Because publications can appear in multiple organizational publication sets, publication counts and citation totals may overlap across units. As a result, aggregated institutional totals cannot be interpreted as simple sums between organizational units. Indicators therefore reflect organizational participation in research output rather than exclusive ownership of publications. However, full counting may inflate publication counts at aggregated levels, particularly when collaborative publications involve multiple units with the same university. Furthermore, the approach does not distinguish between primary and secondary contributions or the distribution of work within a publication. However, despite these limitations, full counting remains an appropriate approach when calculating scientometric indicators at an organizational level. After defining the attribution model used in this study, the following sec-

tion describes how organizational levels are constructed for the subsequent analysis.

3.4 Organizational Levels

Universities are typically structured through multiple organizational levels, such as institutes and departments in the case of WU. These hierarchical structures organize research activities and administrative responsibilities within the university. Therefore, scientometric evaluation at the organizational level must consider how publication data is attributed within these hierarchical structures. The PURE organizational data set contains hierarchical identifiers linking organizational units to higher level entities. In the extracted data, the child level represents the lowest organizational entity associated with publications, thus forming the initial level of analysis. Research performance may vary between different organizational levels; therefore, evaluating only a single level would limit the ability to analyze how research output is distributed between the different units within the university. Therefore, scientometric indicators are computed across multiple hierarchical levels within the organizational structure. The following paragraph describes how publications are first assigned to organizational units of the child-level before aggregation to higher levels of the university's hierarchy.

Child-level organizational units represent the lowest level of the university's hierarchy available in the PURE organizational structure. These units correspond to specific research groups led by specific professors, chairs, or, if unclear, other organizational entities within WU. Because publication affiliations are recorded at this level, it serves as the initial level of publication attribution. Publications are linked to child units through organizational identifiers derived during the affiliation mapping process. Each publication may be associated with multiple child units when authors belong to different organizational entities. As a result, publications are assigned to the child-level organizational units associated with their contributing authors. A deduplicated publication table was constructed containing unique combinations of publication identifiers and child-level organizational units. Therefore, each publication is counted once per child unit, even if multiple authors from the same unit contributed to the publication. This data set forms the foundation for all subsequent organizational aggregation and subsequent indicator computation. After constructing the child-level publication data set, publications can be aggregated to higher organizational levels within WU.

Within WU's hierarchy, institutes represent intermediate organizational units,

as they typically group several child-level units under a broader research or administrative structure. Evaluating research performance at the institute level allows for analysis on a more aggregated scale. Within this study, child-level units were linked to their corresponding institutes using the organizational structure data set. A mapping table was constructed to connect child-level identifiers with institute-level identifiers, which in turn defines the hierarchical relationship between the two organizational levels. The resulting data set contains unique combinations of publications and institute-level units. Duplicate publication-institute pairs were removed to ensure that each publication was counted once per institute. The following paragraph describes the aggregation of publication data to the department level, representing a higher level of the university's hierarchy.

Departments represent the highest level of organizational entities within the university structure, as they group multiple institutes and child units under broader disciplinary or administrative categories. Evaluating research performance at the department level enables university-wide comparison across major units. Child-level organizational units were linked to departments through the organizational structure data set. Another mapping table was constructed connecting child-level identifiers with department-level identifiers, thus establishing the hierarchical relationship between lower-level units and departments. Publication assignments from the child-level data set were aggregated to the department level using the mapping table, resulting in a data set that contains unique publication-department combinations. Duplicate entries were removed once again to ensure that each publication was counted once per department. Based on these hierarchical publication data sets, scientometric indicators can be computed for each organizational level.

3.5 Computation of h-index and g-index

After constructing publication data sets for each organizational level, scientometric indicators can be computed. These data sets contain unique publication-organization combinations with associated citation counts; therefore, each organizational unit has a defined publication set that serves as the basis for indicator computation. Each publication in the data set is associated with a citation count extracted during data preparation, thus forming a citation distribution for each organizational unit. Citation-based indicators are computed from these computations. For each organizational unit, the set of associated publications and citation counts is processed since the indicator values are calculated separately for each organizational entity. The same computation procedure is applied across all organizational levels. The

following paragraphs describe the computation procedures for the specific scientometric indicators used in this study.

To calculate the scientometric indicators, the set of associated publications for each organizational unit and their corresponding citation counts were extracted from the publication data set. These citation values represent the citation distribution of the publications attributed to each unit, thus forming the basis for the computation of the h-index. The citation counts of the publications belonging to each organizational unit were sorted in descending order, creating a ranking. This ranking created an ordered list of citations in which the publications with the highest citation counts appear first. The ranked citation distribution enables identification of the citation threshold used to determine the value of the h-index. The h-index is determined by identifying the largest value h for which at least h publications have received h or more citations. This condition is evaluated using the ranked citation list, and the resulting value represents the organizational h-index, capturing both the publication output and the impact of the citation of the unit. The same computational procedure was applied across all organizational levels considered in the analysis. Therefore, the h-index is calculated for child units, institutes, and departments. This consistent approach ensured the comparability of indicator values across the different levels of WU's hierarchy. In addition to the h-index, the g-index is calculated to capture the influence of highly cited publications within the WU data set.

The computation of the g-index uses the same organizational publication data sets described in the previous sections, as for each organizational unit, the associated publications and their citation counts form the input for the calculation. Once again, these citation counts represent the citation distribution of the publication set belonging to the unit. As in the h-index computation, the publications of each organizational unit are sorted in descending order according to their citation counts, creating a ranking that produces an ordered list of citation values. The ordered list is required to evaluate the cumulative citation thresholds used in the g-index calculation. The g-index is determined by identifying the largest value, such that the top g publications collectively received at least g^2 citations. This condition is evaluated by computing the cumulative citation counts of the ranked publication list, with the index value corresponding to the largest rank position for which the cumulative citation threshold is satisfied. The same computational procedure is applied across all organizational levels used in the analysis; therefore, the g-index is computed for child units, institutes, and departments. Beyond the procedural computation of the indices, certain methodological considerations

arise from the computational approach used in this study.

Both the h-index and the g-index are computed directly from the organizational data sets constructed earlier in the methodology. Each organizational unit is evaluated on the basis of its associated publication set and citation distribution, thus ensuring that indicator values reflect the citation performance of the publications attributed to the unit. Both the h-index and the g-index are non-additive indicators, meaning they cannot be derived by summing or averaging indices of smaller units but must be computed directly from the full publication set associated with a unit instead. The same computational framework is applied across all organizational levels used in the analysis, which ensures that the indicator values remain comparable throughout the hierarchy. Despite the systematic computational framework used for the computation of indicators, certain methodological and data-related limitations must be considered.

3.6 Limitations of the method

Organizational scientometric analysis requires assigning publications to organizational units. As publications frequently involve multiple authors and affiliations, attribution decisions inevitably introduce methodological simplifications, thus influencing the resulting publication sets. In full counting, publications are fully attributed to all participating units, which can lead to overlapping publication counts between units, particularly when collaborative publications involve multiple units within the same institution. As a result, aggregated publication counts may exceed the total number of unique publications. Computed indicator values should therefore be interpreted as representing organizational participation in research output, rather than exclusive ownership, since the attribution model captures collaborative research contributions, but does not distinguish between relative contributions by different units. In addition to attribution-related considerations, the scientometric indicators used in this study also exhibit structural limitations when applied to organizational entities.

As the h-index and the g-index were originally developed to evaluate individual researchers, their interpretation changes when applied to organizational units, as organizational units represent aggregations of multiple researchers and publication sets, which alters the underlying measurement context. The non-linear and non-additive properties of both indices complicate comparisons across hierarchical organizational levels. As larger units generally possess larger publication sets, which increases the potential index values, the

values may partly reflect organizational size, rather than purely research performance. Beyond indicator-specific considerations, empirical results are also influenced by characteristics of the underlying data and the research information system.

The analysis in this study relies on the metadata for the publication and affiliation stored in PURE WU. Since organizational attribution of publication depends on author affiliation data recorded in the system, incomplete or inconsistent affiliation records may negatively influence the according of publication assignment. Citation counts used in the analysis originate from the SCOPUS integration within PURE WU; therefore, citation coverage depends on the indexing and update frequency of SCOPUS. Publications not indexed in SCOPUS may therefore exhibit lower or missing citation counts, which influences the computed indicators. Furthermore, organizational assignment relies on the organizational identifiers stored in the PURE data set; therefore, missing or incomplete organization identifiers required the fallback identifier implemented during data preparation. Although this approach improves coverage, it still depends on the accuracy of the underlying system metadata. In addition to methodological and data-related constraints, the scope of the present analysis is intentionally limited to specific methodological objectives.

This thesis focuses specifically on the methodological computation of organizational h-index and g-index values and aims to demonstrate how scientometric indicators can be computed using publication and citation data stored in PURE WU. Therefore, the study focuses on the methodological implementation of the indicator calculation within the university's data set. The thesis does not attempt to develop new scientometric indicators or perform cross-institutional benchmarking or comparative evaluation between universities, since the analysis is limited to the WU data set. Broader questions of research evaluation policy therefore fall outside the scope of this thesis.

4 Empirical Results

4.1 Dataset Overview

The empirical analysis of this study is based on the data set constructed in section 3, containing publication records attributed to organizational units at WU. Each record includes publication identifiers, organizational assignment, and citation counts used for indicator computation. After filtering the

PURE export data into citation-based metrics, the working data set contains 7,351 publication-metric records. These records represent publications with citation information that can be used for scientometric computation, as each record includes the publication identifier, associated author(s), and organizational information, as well as the extracted citation count. Through the affiliation linkage and organizational mapping described in section 3, the data set is transformed into publication-organization tables used for indicator computation. These tables represent the empirical basis for computing the h-index and g-index across organizational units at WU, with the resulting data sets enabling scientometric analysis across multiple levels. Beyond the overall size of the data set, it is also important to examine how publications are distributed between organizational units within the university.

Publications are assigned to organizational units through the affiliation mapping process described in Section 3. The assignment is initially performed at the child-unit level, which represents the lowest organizational entities within the extracted PURE data. At the child-unit level, the deduplicated data set contains 1,753 publication-unit assignments representing 1,584 unique publications. When aggregated to higher organizational levels, the data set contains 1,626 publication-institute assignments (1,514 unique publications) and 1,636 publication-department assignments (1,577 unique publications). Because publications can be associated with multiple organizational units, the number of assignment of publication-units exceeds the number of unique publications, reflecting the collaborative nature of research activity between units within WU. The hierarchical aggregation therefore produces different publication sets at each organizational level. In addition to the organizational distribution of publications, it is also useful to examine the citation characteristics of the data set used.

Citation counts represent the impact dimension of the publication included in the data set, which forms the basis for computing the h-index and g-index indicators. Therefore, understanding the citation distribution is crucial for providing context for interpreting the resulting scientometric indicators. The empirical citation distribution exhibits a mean citation count of 11.87, with a median citation count of 5. The upper quartile of the data set has 11 citations, with citations in the entire data set ranging between 0 and 409 citations. The citation distribution is highly skewed, with only a small number of publications receiving significantly high citation counts and most publications receiving comparatively few citations. Such a skewed citation distribution is not unusual for scientometric data sets, as they also influence the behavior of the computed citation-based indicators. In addition to the

distribution of citations, it is also important to examine the completeness of the data and the coverage of the data set used in this analysis.

During data preparation, an effective organizational identifier was constructed to improve affiliation coverage, with the fallback strategy ensuring that organizational identifiers were available for approximately 99.98 percent of affiliation records. As a result, nearly all author-affiliation entries could be linked to an organizational unit. After linking publications to organizational units, 1,584 publications could be assigned to at least one child-level organizational unit, which corresponds to a coverage rate of approximately 84.0 percent of the publications in the mapped publication-person data set. The remaining publications lacked sufficient affiliation information for reliable organizational assignment. The high coverage of organizational identifiers ensured that most publications can be included in the subsequent analysis. Although some records remain unassigned, the available data set provides a reasonable empirical basis for the computation of scientometric indicators. Based on this data set, the following sections present the computed scientometric indicators at different levels.

4.2 Child-Level Results

Child units represent the lowest organizational level used in the empirical analysis, as these units correspond to the most specific research entities within the PURE data. Evaluating indicators at this level provides a detailed view of research output and citation performance. The child-level analysis is based on the deduplicated publication-child data set constructed in Section 3, containing 1,753 publication-unit assignments, representing 1,584 unique publications. Each child unit therefore possesses an associated publication set from which scientometric indicators can be computed. For each child unit, the h-index and g-index are computed on the basis of the citation distribution of the assigned publications, capturing different aspects of citation performance. The following paragraphs describe the distribution and structural characteristics of these indicator values in different child units.

The h-index values are derived from the child-level publication data set constructed earlier. In the code, publications are assigned to each child unit, grouped by organizational identifier, forming unit-specific publication sets, with the citation counts then being used to compute the h-index for every child unit. The empirical spread of data can be seen in table 1 below:
A small number of units show substantially higher h-index values, such as the organizational unit "Child Unit 1". These correspond to units with larger

Table 3: Child-Level Scientometric Indicators

ID	Unit	Pubs	Citations	CPP	h	g
C001	Child Unit 1	73	1473	20.18	19	36
C002	Child Unit 2	60	1027	17.12	12	31
C003	Child Unit 3	48	559	11.65	12	22
C004	Child Unit 4	21	510	24.29	7	21
C005	Child Unit 5	18	500	27.78	6	18
C006	Child Unit 6	116	499	4.30	10	17
C007	Child Unit 7	63	423	6.71	13	18
C008	Child Unit 8	16	404	25.25	9	16
C009	Child Unit 9	24	387	16.13	9	19
C010	Child Unit 10	42	381	9.07	12	16
C011	Child Unit 11	19	349	18.37	9	18
C012	Child Unit 12	17	346	20.35	9	17
C013	Child Unit 13	46	340	7.39	11	16
C014	Child Unit 14	34	334	9.82	11	17
C015	Child Unit 15	44	332	7.55	11	16

publication sets or more highly cited papers. Although the h-index reflects threshold-based citation performance, the g-index places greater emphasis on highly cited publications, which will be examined in the following paragraph.

The table presented in the previous paragraph reports the publication counts, total citations, citations per publication (CPP), h-index and g-index values for child-level units. Both indicators are computed using the same publication sets derived from the child-level data set constructed in the data preparation pipeline, therefore providing a direct comparison of threshold-based and cumulative citation indicators across the same entities. Across all child units in the table, the g-index values are consistently higher than the corresponding h-index values. This pattern reflects the cumulative citation characteristic of the g-index, which considers highly cited publications, therefore, unit with relatively high citation counts exhibit larger differences between the two indices. Organizational units with a larger set of publications and higher citation totals tend to exhibit higher indicator values; however, the magnitude of the g-index substantially exceeds the h-index when citation counts are concentrated among highly cited publications. Although the child-level analysis provides a detailed view of scientometric performance at the lowest organizational level, the following section examines how these indicators behave when publications are aggregated at the institute level.

4.3 Institute-Level Results

The previous section examined scientometric indicators at the child-unit level, representing the lowest level of organizational entities within WU. Although this level provides a detailed view of research performance, a higher organizational level analysis is required to gain broader insights. Institute-level publication sets are constructed by combining the publications associated with child units belonging to the same institute, with this aggregation being implemented through the mapping of child-level organizational identifiers to institute-level identifiers. Publications linked to multiple child units within the same institute are duplicated prior to the calculation of the indicator. Based on the aggregated publication sets, the h-index and g-index are recomputed for each institute, as they cannot be derived from child-level values. The following paragraph presents the distribution of indicator values in different institutes.

In the code, publications associated with child units are mapped to their corresponding institutes and group by institute organizational identifier, with the citation counts within each institute being used to compute the scientometric indicators. Table 2 presents the resulting institute-level scientometric indicators, including publication counts, total citations, citations per publication (CPP), and the calculated values of the h-index and g-index for each institute.

Table 4: Institute-Level Scientometric Indicators

ID	Institute	Pubs	Citations	CPP	h	g
I001	Institute 1	73	1473	20.18	19	36
I002	Institute 2	60	1027	17.12	12	31
I003	Institute 3	67	695	10.37	13	23
I004	Institute 4	38	628	16.53	9	24
I005	Institute 5	29	519	17.90	12	22
I006	Institute 6	21	510	24.29	7	21
I007	Institute 7	116	499	4.30	10	17
I008	Institute 8	38	431	11.34	11	19
I009	Institute 9	63	423	6.71	13	18
I010	Institute 10	46	410	8.91	11	18
I011	Institute 11	57	409	7.18	11	17
I012	Institute 12	24	387	16.13	9	19
I013	Institute 13	43	363	8.44	13	16
I014	Institute 14	38	358	9.42	6	18
I015	Institute 15	19	349	18.37	9	18

The h-index values vary between institutes, reflecting differences in publication volume and citation performance, with institutes with larger aggregated publication sets generally exhibiting higher h-index and g-index values. As in child-level analysis, the g-index values for each institute tend to be higher than their respective h-index values, reflecting the impact of highly cited publications. The data structure exported from PURE shows some limitations here, as in some cases the child-level unit associated with a publication is the respective department directly. Therefore, these entries cannot be directly aggregated to the institute level, as they have been entered directly to the department level. This can be seen by "Institute 2", for example. The following section discusses the effects of aggregation to the highest organizational level at WU, that being, the department level.

4.4 Department-Level Results

The previous section examined scientometric indicators at the institute level. While institute-level analysis captures aggregated research activity within individual institutes, an analysis at the department level is necessary to gain an overview at a university wide level. Department-level publication sets are constructed by mapping institutes to their corresponding departments, with the same procedure being applied that was used to aggregate publications from the child level to the institute level. Based on these aggregated publication sets, the h-index and g-index are again recomputed for each department. Table 3 presents the results of the analysis at the department level.

Table 5: Department-Level Scientometric Indicators

ID	Department	Pubs	Citations	CPP	h	g
D001	Department A	300	3219	10.73	27	45
D002	Department B	208	1949	9.37	18	37
D003	Department C	178	1652	9.28	17	32
D004	Department D	212	1545	7.29	17	27
D005	Department E	153	1422	9.29	20	29
D006	Department F	105	1208	11.50	17	30
D007	Department G	201	1197	5.96	15	27
D008	Department H	134	1036	7.73	14	25
D009	Department I	80	263	3.29	9	12
D010	Department J	21	186	8.86	7	13
D011	Department K	17	51	3.00	5	6
D012	Department L	27	24	0.89	2	3

The table presents publication counts, total citations, citations per publica-

tion (CPP), as well as the h-index and g-index values for each department, making the table representative of the highest level of aggregation in the analysis. The h-index values at the department level vary considerably between different departments, as departments with large publications sets and citation totals show higher h-index values. As in the previous analysis, the g-index values remain consistently higher than the corresponding h-index values, again reflecting the impact of highly cited publications. Furthermore, aggregation at this level naturally increases the publication set size, thus resulting in generally higher indicator values compared to the child-level and institute-level analysis. These patterns show the structural effects of aggregation, which are summarized across organizational levels in the following section.

4.5 Observations

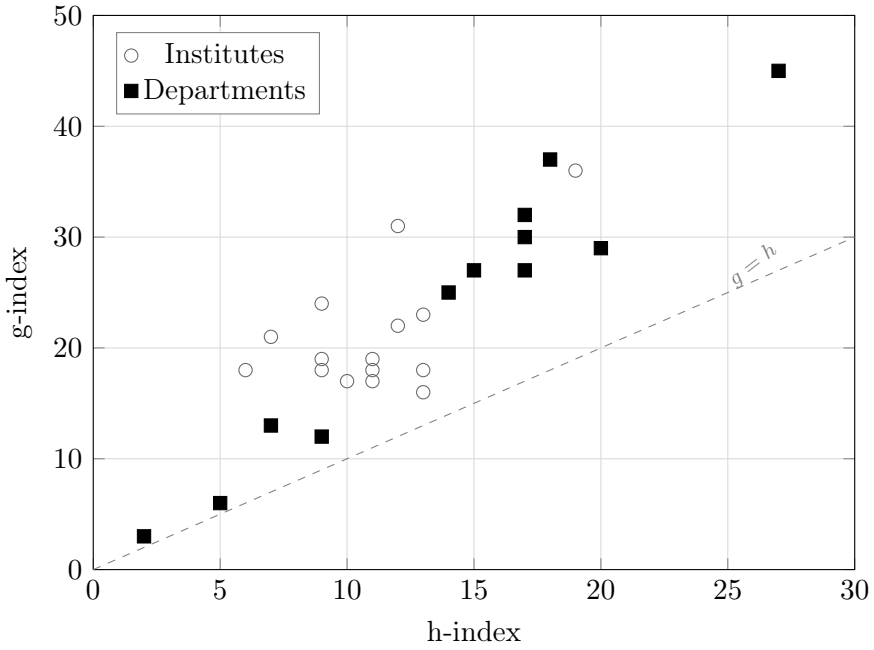

Figure 1: Scatter plot of h-index and g-index values at the institute and department levels. The dashed line represents $g = h$; all points lie above this line, indicating that g-index values consistently exceed h-index values across all organizational units.

Figure 1 provides a visual overview of the computed indicator values at the institute and department levels. The upward shift of department-level

values reflects the systematic effect of hierarchical aggregation on both indicators.

The study examined scientometric indicators across three different organizational levels derived from the existing PURE structure, with the child level representing the lowest level and the department level representing the highest level of aggregation. Child units aggregate into institutes, which, in turn, aggregate into departments. Child level h-index values reached approximately 6 to 19 on average, with the respective institute level indices remaining in a similar range. However, due to the aggregation effects, the h-index values at the department level increased to 27 in the highest case. These increases in indicator values can be traced back to the aggregation of publication sets, with most child level units falling below 100 publications, whereas departments often ranged between 100 and 300 publications. Larger publication sets increase the probability that multiple publications exceed the citation threshold, thus increasing indicator values. The results therefore indicate that hierarchical aggregation within the WU structure systematically increases the magnitude of scientometric indicators.

Both the h-index and the g-index were calculated for each of the three organizational levels, allowing for a direct comparison between the two indicators in identical publication sets. The differences between the two indicators therefore arise from their structural definitions rather than from different data inputs. Across all levels, the values of the g-index were consistently higher than the corresponding values of the h-index, such as "Department A", which has a value of the h-index of 27, while its value of the g-index is 45. This shows that the difference between the two indicators can be substantial, due to differences in citation concentration, as highly cited publications increase the g-index values more than their corresponding h-index values. Differences between the two indices tend to become more visible when publication sets grow larger, which naturally occurs during the aggregation process. The comparison therefore shows that while both indicators measure the citation-based research performance, the g-index reflects the intensity of the citation more strongly, whereas the h-index emphasizes the sustained citation performance across different publications.

The results in the previous sections reveal several structural patterns in the data set, such as the relationships between the volume of the publication, the total number of citations, and the indicator values. Larger publication sets produce higher potential h-index and g-index values, with publication sets such as "Department A" exhibiting the largest publication volume, as

well as the highest h-index and g-index values across all departments. Furthermore, units with large citation totals relative to their publication counts also tend to produce higher indicators, although this effect is stronger for the g-index values. The observations presented in this chapter show how publication volume, citation distribution, and hierarchical aggregation influence scientometric indicators within the analyzed data set. The following chapter discusses the methodological and practical implications of these findings for the evaluation of organizational research.

5 Discussion

5.1 Adaptation of h-index/g-index to Organizational Units

Both the h-index and the g-index measure the relationship between publications and citations within a single author's publication set; therefore, the evaluated entity is a single researcher with a clearly defined set of publications. In this thesis, the indices are applied to organizational units within WU rather than individuals; therefore, the publication sets are constructed from multiple researchers linked through affiliation metadata in PURE. These sets of publications represent the combined research output of organizational entities, which becomes substantially larger than an individual set of publications. In the data set analyzed in this thesis, child units typically contain tens of publications, whereas some departments contain more than 200 publications. Because both indices operate on citation-ranked publication lists, their behavior strongly depends on the size and structure of the publication set. When publication sets are aggregated across multiple researchers, the indicators reflect collective citation performance rather than individual productivity. Therefore, applying individual-level indices to organizational entities requires explicit methodological handling of publication aggregation and attribution, which forms the basis of the empirical analysis conducted in this thesis.

The h-index values computed in the analysis reflect the set of citation-ranked publications constructed from PURE data, representing the aggregated research output of WU organizational units. The indicator was consistently calculated for child units, institutes, and departments. At the child-unit level, the values of the h-index in the data set range approximately between 6 and 19, with similar magnitudes appearing at the institute level, where the indicator remains within a comparable range. At the department level, the indicator increases further, reaching 27 at most. Furthermore, organizational

units with larger sets of publications tend to exhibit higher h-index values, such as "Department A" reaching 300 publications, as well as comparatively higher indicator values than smaller organizational units. Therefore, the h-index at the organizational level reflects how many publications within the aggregated publication set exceed a certain citation threshold. The empirical results therefore show that organizational h-index values are strongly influenced by the size and composition of the aggregated publication sets derived from affiliation data.

The g-index was calculated in the same manner as the h-index was calculated; however, the results varied between the two indicators. At the child-unit level, the values of the g-index reached as high as 36, which is considerably higher than their h-index counterparts for the same units. Organizational units with high citation totals relative to their publication counts show particularly larger g-index values, as the g-index uses cumulative thresholds, where highly cited publications contribute more strongly. Therefore, the empirical results show that the g-index is particularly sensitive to citation concentration within the publication sets, which explains the larger indicator values observed in the units analyzed in this thesis.

Both indicators were calculated from the same sets of organizational publications, allowing for a direct comparison between the two. Since both indices were calculated for identical publication data sets derived from PURE affiliation data, the observed differences between the indicators can be attributed to their structural differences rather than data differences. The consistent pattern in the empirical data shows higher relative g-index values at all three organizational levels. This occurs because the h-index only increases when additional publications exceed a citation threshold, whereas the g-index can increase if a single publication gains more recognition in the form of citations. Therefore, the two indices capture different aspects of collective research performance, with the h-index favoring consistent citation performance, while the g-index favors a high citation concentration. The empirical comparison therefore demonstrates that the choice of indicator influences how organizational research performance is represented, particularly in aggregated publication sets.

5.2 Implications of Full Counting

Computing scientometric indicators at the organizational level requires assigning publications to organizational units. The computed indicators, therefore, depend on the chosen attribution rules. In this thesis, publication attri-

bution is based on author affiliation metadata stored in the PURE system, which means that publications inherit organizational membership through affiliated researchers. The analysis applies a full counting approach, which means that the publication is fully attributed to each organizational unit represented by its authors. If authors of a publication belong to multiple units within WU, the publication is included in the publication sets of all relevant units. This allows publication sets to reflect the complete research activity associated with each unit, making it consistent with the aggregation of publication sets used to compute the h-index and g-index. Therefore, full counting provides transparent and operational attribution rules for constructing organizational publication sets, which form the basis for indicator computation in this thesis.

The full counting approach used in this thesis directly shapes the set of organizational publications used for the computation of the indicators. Because publications are attributed to all organizational units represented by their authors, the same publication can appear in multiple organizational publication sets, which occurs when co-authors belong to different institutes or departments within WU. Both the h-index and the g-index must be computed from the publication sets of each unit separately, which increase in size through the aggregation process used. Full counting therefore interacts with the non-additive structure of the indicators by expanding the organizational publication sets from which the indices are computed, thus influencing the size of the resulting index values.

Fractional counting is the main methodological alternative to the full counting approach used in this thesis, as it distributes publication credit proportionally among contributing authors or organizational units. Instead of assigning a full publication to each affiliated unit, the contribution is divided between the participating entities. In fractional counting, publications with authors from multiple units would contribute only partially to each unit's publication set. As a result, the effective publication and citation counts of units would be reduced compared to the full counting approach applied in this thesis. Because both computed indices depend on the size and citation distribution of publication sets, fractional attribution would likely produce lower indicator values for units, with the extent of this difference depending on the degree of cross-unit collaboration. Although fractional counting represents an alternative attribution model, this applied full counting to preserve complete publication membership within units.

Full counting provides a clear and operational rule for assigning publications

to organizational units, which is directly compatible with the affiliation-based publication records stored in the PURE system, allowing organizational publication sets to reflect the complete research activity associated with each unit. Because publications may be attributed to multiple units, full counting can lead to overlapping publication sets between units. As a result, aggregated publication counts may overrepresent collaborative publications that involve multiple units. The resulting h-index and g-index values should therefore be interpreted as indicators of organizational participation in research output, rather than strictly independent publication contributions. Despite these trade-offs, full counting provides a transparent and consistent methodological basis for constructing publication sets and computing scientometric indicators within the data set analyzed in this thesis.

5.3 Impact of Meta Data Quality

Organizational publication sets in this thesis were constructed using author affiliation information stored in PURE, in which publications are linked to organizational units through research identifiers and institutional affiliation metadata. Publications inherit organizational attribution through the affiliations of their authors recorded in PURE, which determine whether a publication becomes part of the publication set of a child unit, institute, or department. Because the h-index and the g-index are computed from these organizational publication sets, the resulting indicator values depend directly on the accuracy and completeness of the underlying affiliation metadata. Therefore, the reliability of organizational scientometric indicators is strongly dependent on the integrity of the affiliation data used to assign publications to different units within the PURE system.

Within the PURE data set used in this thesis, individual researchers may hold affiliations with more than one organizational unit, therefore, publications authored by such researchers may be linked to multiple institutes or departments at WU. Under the full counting approach used in this thesis, publications with authors affiliated with multiple units are included in the publication sets of all relevant organizational units, which creates overlapping publication sets across different units. Furthermore, researchers may also change their organizational affiliation over time, which may affect how publications are attributed within data sets, depending on how affiliation records are stored in the system. Therefore, multi-affiliation and temporal changes in organizational membership introduce additional complexity into the attribution of publications to units, which can influence the resulting scientometric indicators.

The h-index and the g-index values computed in this thesis are derived from the available PURE affiliation metadata, therefore any change in publication attribution alters the citation-ranked publication lists used for indicator computation. Adding or removing even a small number of publications from an organizational publication set may therefore change the threshold at which the index value is determined. As a result, the computed organizational indicators should be interpreted with awareness that metadata inconsistencies or attribution adjustments may influence the resulting values. Maintaining accurate and consistent affiliation metadata within the system is therefore essential for ensuring the reliability of scientometric indicators computed in this thesis.

5.4 Limitations of SCOPUS Data

The citation counts used to compute the h-index and the g-index values in this thesis originate from the SCOPUS database, which are then integrated into the PURE system at WU. SCOPUS primarily indexes peer-reviewed journal publications, while other types of publications may be less comprehensively represented. As a result, the citation counts used in this thesis reflect the coverage and indexing practices of the SCOPUS database. Because the computed indicators are based directly on SCOPUS citation counts, they represent the impact of the citation captured within this specific database. Therefore, the indicators computed in this thesis depend on the coverage and scope of the SCOPUS database integrated into the PURE WU system.

The citation counts used in this thesis represent the citation state recorded in the SCOPUS database at the time of data extraction from PURE. As citation accumulation is a time-dependent process, publications typically receive citations gradually after publication, with older publications having had more time to accumulate citations. As a result, the computed indicators may be influenced by the age distribution of publication within the data set. Therefore, organizational units with larger numbers of older publications may show higher indicator values, independent of research performance. The scientometric indicators computed in this thesis therefore reflect the impact of the citation accumulated within a specific time window defined by the SCOPUS data available at the time of analysis.

As the citation data used in this thesis is derived exclusively from the SCOPUS database, computed metrics can vary depending on the indexing policies, journal coverage, and citation tracking methods used in different bibli-

ographic databases. As a result, the scientometric indicators computed for WU represent the impact of the citation as measured within the SCOPUS indexing framework. Differences in database coverage and citation indexing practices represent an inherent limitation when interpreting indicators derived from a single citation database.

The reliance on SCOPUS as the sole citation source introduces several specific limitations that may affect the computed indicators. SCOPUS primarily indexes peer-reviewed journal articles, which means that other forms of research output, such as books, book chapters, conference proceedings, or working papers, are less comprehensively represented [12]. This is particularly relevant for disciplines in which non-journal publications constitute a significant share of research output, such as law or the humanities. Furthermore, SCOPUS has stronger coverage of English-language publications, which may disadvantage organizational units that frequently publish in German or other languages [11]. As a result, the computed indicators may systematically underrepresent the research output of units whose publication practices do not align with the coverage profile of SCOPUS. This could partly explain the low indicator values observed for certain units, such as Department L, where publication types and languages common in legal scholarship may fall outside the scope of the SCOPUS database. Since no alternative citation source was used in this analysis, it is not possible to assess how indicator values would differ if computed from a different database, such as Web of Science or Google Scholar.

5.5 Practical Implications for WU

The organizational h-index and g-index values computed in this thesis provide quantitative indicators of the research output and the impact of citations for organizational units within WU. The computed indicators may support internal research monitoring and performance reporting. Because the indicators were calculated for the child, institute, and department levels, they provide a structured overview of research activity across different levels within WU. Therefore, the computed indicators demonstrate how research information systems such as PURE WU can support the analysis of research performance.

Although the computed h-index and g-index values provide quantitative insights into organizational research activity, they represent simplified representations of complex research performance, as they capture publication output and citation impact, but they do not fully represent all dimension of research

activity. The indicator values in this thesis depend on the attribution model used, specifically the full counting approach applied in the analysis. Therefore, different attribution models could produce different publication sets and therefore different indicator values. The indicators also depend on the quality of the affiliation metadata in PURE and the citation coverage provided by the SCOPUS database. The indicators should therefore be interpreted within the methodological and data context in which they were computed.

The methodology implemented in this thesis could support internal analytical reporting on research activity in different units at WU. The use of these indicators should ensure transparent methodological definitions, including clearly defined attribution rules and aggregation procedures, therefore, maintaining accurate affiliation metadata within PURE is also crucial to ensure reliable attribution of publications. The values of the organizational h-index and the g-index should be used as complementary indicators within broader research evaluation frameworks, rather than as standalone measures of research performance.

6 Conclusion

6.1 Summary of Research Objective and Approach

Universities increasingly rely on quantitative indicators to evaluate research performance, with research systems such as PURE WU storing structured publication and citation data, which enables scientometric analysis at the organizational level. Scientometric indicators such as the h-index or the g-index were originally designed for individual researchers; therefore, applying them to organizational entities introduces structural challenges, particularly with regard to the definition of the aggregation, attribution and publication set. The objective of this thesis was to develop and implement a methodological framework for computing organizational h-index and g-index values using publication and citation data from PURE WU.

The analysis was based on publication and citation data extracted from the PURE WU system, with citation counts originating from the SCOPUS integration. Publication records were prepared and linked to organizational affiliation data to construct organizational publication sets. The data set was cleaned, standardized, and merged with author affiliation records, enabling publications to be assigned to organizational units. A full counting attribution model was applied to construct publication sets for different or-

ganizational levels. Based on these set of publications, the values of the organizational h-index and g-index were calculated for the child units, institutes, and departments within WU.

6.2 Summary of Key Findings

The results demonstrate that both the h-index and the g-index can be computed for organizational publication sets derived from institutional research information systems. Across all organizational levels analyzed, the g-index values were consistently higher than the corresponding h-index values, reflecting the cumulative citation structure of the g-index, with units with highly cited publications exhibiting greater differences between the two indicators. Furthermore, aggregation across organizational levels increased publication counts and citation totals, which resulted in higher indicator values at the institute and department levels compared to child units.

The results highlight that organizational scientometric indicators strongly depend on the attribution model used to construct the publications. Furthermore, the analysis also demonstrated the importance of reliable affiliation metadata in the PURE system, as the assignment of the publication to organizational units depends on the accurate author affiliation record. Therefore, research information systems play a central role as data infrastructure for organizational scientometric analysis.

6.3 Theoretical and Practical Contributions

This thesis contributes to the literature on organizational scientometrics by examining how individual-level indicators behave when applied to collective entities. It demonstrates a practical methodological approach for constructing organizational publication sets and computing the h-index and g-index values using data from the research information system. The analysis highlights the structural challenges of applying non-additive indicators to aggregated publication sets.

The study demonstrates how PURE data can be used to compute scientometric indicators at multiple organizational levels within WU, which provides a structured analytical framework that could support internal research monitoring and reporting. At the same time, the study emphasizes the importance of methodological transparency and careful interpretation when using such indicators.

6.4 Limitations and Future Research

The analysis implemented in this thesis focuses on organizational scientometric indicators derived from publication and citation data stored in PURE WU. The methodological framework applied in the analysis uses a single attribution model, namely, the full counting (union) approach, to construct organizational publication sets. Therefore, the computed indicators depend on the affiliation metadata, which determines how publications are assigned to organizational units. Inaccuracies or incomplete affiliation records may influence the composition of publication sets and therefore the computed indicator values. Citation counts used for indicator computation originate from the SCOPUS database integrated within PURE WU; therefore, scientometric indicators reflect the impact of citations captured in this specific database. These limitations should be considered when interpreting the organizational h-index and g-index values computed in this thesis.

A particular limitation concerns the exclusive reliance on SCOPUS as the citation data source. As discussed in Section 5.4, SCOPUS coverage varies by discipline, publication type, and language, which may systematically affect the computed indicators for certain organizational units. Future research could address this limitation by comparing indicator values computed from multiple citation databases, such as Web of Science or Google Scholar, to assess the sensitivity of organizational scientometric indicators to the choice of data source. Such a comparison would help determine whether the observed differences between organizational units reflect actual variation in research performance or are partly attributable to source-specific coverage gaps.

Future research could explore alternative attribution models for constructing organizational publication sets, such as fractional counting approaches. Comparing full counting and fractional counting could provide further insight into how attribution rules influence scientometric organizational indicators. Further studies could also analyze additional scientometric indicators or field normalized metrics to provide a broader perspective on organizational research performance, as such extensions could help evaluate how different indicators behave at aggregated organizational levels. Future research could also apply the methodological framework developed in this thesis to other institutions or research information systems, allowing comparisons between different environments. These extensions would contribute to a deeper understanding of how scientometric indicators can be applied and interpreted in organizational research evaluation contexts.

References

[1] Loet Leydesdorff and Staša Milojević. Scientometrics, 2013.

[2] ScienceDirect. Scopus, n.d.

[3] Lutz Bornmann and Hans-Dieter Daniel. What do we know about the h index? *Journal of the American Society for Information Science and technology*, 58(9):1381–1385, 2007.

[4] Jorge E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National academy of Sciences*, 102(46):16569–16572, 2005.

[5] Leo Egghe. An improvement of the h-index: The g-index, 2006.

[6] John Mingers and Loet Leydesdorff. A review of theory and practice in scientometrics. *European journal of operational research*, 246(1):1–19, 2015.

[7] I Diane Cooper. Bibliometrics basics. *Journal of the Medical Library Association: JMLA*, 103(4):217, 2015.

[8] Siluo Yang, Qingli Yuan, Jiahui Dong, et al. Are scientometrics, informetrics, and bibliometrics different?, 2017.

[9] Moed. *Synopsis*, pages 35–68. Springer Netherlands, Dordrecht, 2005.

[10] Per O Seglen. The skewness of science. *Journal of the American society for information science*, 43(9):628–638, 1992.

[11] Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. Bibliometrics: the leiden manifesto for research metrics. *Nature*, 520(7548):429–431, 2015.

[12] Ludo Waltman. A review of the literature on citation impact indicators. *Journal of informetrics*, 10(2):365–391, 2016.

[13] Keith Jeffery and Anne Asserson. Institutional repositories and current research information systems. *New Review of Information Networking*, 14(2):71–83, 2009.

[14] Elsevier. Pure: Research information management system, n.d. Accessed: 2026.

Appendix A
Python Code for Data Preparation and Indicator Computation

The following appendix contains the Python code used for data preparation, organizational attribution, and the computation of h-index and g-index values described in Section 3 of this thesis.

01_cleaning_and_union

February 20, 2026

1 1 - Data Preparation

1.1 1.1 - Importing and Project Setup

The project root is automatically detected by locating the data_raw folder, ensuring consistent and location-independent access to the input data.

[11]:
```
from pathlib import Path
import pandas as pd

candidates = [Path.cwd()] + list(Path.cwd().parents)

PROJECT_ROOT = None
for p in candidates:
    if (p / "data_raw").exists():
        PROJECT_ROOT = p
        break

if PROJECT_ROOT is None:
    raise FileNotFoundError(
        f"Could not find 'data_raw' folder from cwd={Path.cwd()} or its parents.
↪"
    )

DATA_RAW = PROJECT_ROOT / "data_raw"

import warnings
warnings.filterwarnings("ignore")
```

1.2 1.2 - Environment Setup and Library Import

Initialization of the analysis environment through library imports and definition of the data directory structure.

[2]:
```
import pandas as pd
import numpy as np
import re
from pathlib import Path
```

```
pd.set_option("display.max_columns", 200)
pd.set_option("display.max_rows", 50)

BASE_DIR = Path(".")
DATA_RAW = BASE_DIR / "data_raw"
```

1.3 1.3 - Data Import

The core datasets are loaded from the data_raw directory, including the organizational structure, affiliation data, and publication citation datasets.

[3]:
```
df_org_structure = pd.read_excel(DATA_RAW / "Org_Structure.xlsx")
df_affiliations  = pd.read_excel(DATA_RAW / "Personen_Affiliation_Start_Schluss.
 ↪xlsx")
df_popc          = pd.read_excel(DATA_RAW / "Publikation_Org_Person_Zitation.
 ↪xlsx")
df_poc           = pd.read_excel(DATA_RAW / "Publikation_Org_Zitation.xlsx")

print("Org Structure:", df_org_structure.shape)
print("Affiliations:", df_affiliations.shape)
print("Pub-Org-Person-Citation:", df_popc.shape)
print("Pub-Org-Citation:", df_poc.shape)
```

```
Org Structure: (186, 15)
Affiliations: (47009, 5)
Pub-Org-Person-Citation: (12639, 9)
Pub-Org-Citation: (4051, 7)
```

1.4 1.4 - Initial Data Inspection

Preliminary inspection of column names to verify the structure of the imported dataset.

[4]:
```
df_popc.columns
```

[4]:
```
Index(['Jahr', 'Organisationen von Mitwirkenden', 'Personen', 'Typ',
       'Pub-Pure-ID', 'Name', 'Person-Pure-ID', 'Name der Metrik',
       'Numerischer Wert'],
      dtype='object')
```

1.5 1.5 - Column Standardization

The column names of the publication dataset are standardized and translated into a consistent format to simplify subsequent data processing and analysis steps.

[5]:
```
rename_map = {
    "Pub-Pure-ID": "pub_id",
    "Organisationen von Mitwirkenden": "org_unit",
```

```
    "Person-Pure-ID": "person_id",
    "Jahr": "pub_year",
    "Typ": "pub_type",
    "Name der Metrik": "metric_name",
    "Numerischer Wert": "metric_value",
}

df = df_popc.rename(columns=rename_map).copy()
df.columns
```

[5]:
```
Index(['pub_year', 'org_unit', 'Personen', 'pub_type', 'pub_id', 'Name',
       'person_id', 'metric_name', 'metric_value'],
      dtype='object')
```

1.6 1.6 - Metric Inspection

Inspection of available metric types to identify the citation-based indicator used for the analysis.

[6]:
```
df["metric_name"].value_counts()
```

[6]:
```
metric_name
Anzahl der Zitationen    7351
Name: count, dtype: int64
```

1.7 1.7 - Citation Value Extraction

A helper function is defined to extract numeric citation values from potentially formatted string entries. This ensures that citation counts are consistently stored as numeric values for subsequent analysis.

[7]:
```
def extract_citations(value):

    if pd.isna(value):
        return np.nan

    if isinstance(value, (int, float, np.integer, np.floating)):
        return float(value)

    m = re.search(r"\d+", str(value))
    return float(m.group()) if m else np.nan
```

1.8 1.8 - Citation Filtering and Numeric Transformation

The dataset is restricted to citation-based metrics, and citation values are converted into a numeric format to enable quantitative analysis.

[8]:
```
df_cit = df[df["metric_name"] == "Anzahl der Zitationen"].copy()

df_cit["citations"] = df_cit["metric_value"].apply(extract_citations)

df_cit["citations"].describe()
```

[8]:
```
count    7351.000000
mean       11.867501
std        30.191234
min         0.000000
25%         1.000000
50%         5.000000
75%        11.000000
max       409.000000
Name: citations, dtype: float64
```

1.9 1.9 - Data Validation Check

Verification that all citation values were successfully parsed into numeric format.

[9]:
```
unparseable = df_cit["citations"].isna().sum()
print("Unparseable citation values:", unparseable)

df_cit[df_cit["citations"].isna()][["metric_value"]].head(20)
```

```
Unparseable citation values: 0
```

[9]:
```
Empty DataFrame
Columns: [metric_value]
Index: []
```

2 2 - ID-Based Affiliation Mapping

2.1 2.1 - Reload Affiliation Data with Organization IDs

The affiliation dataset containing organization-level Pure IDs is reloaded. This dataset forms the basis for ID-based organizational assignment of publications.

[12]:
```
expected_file = "Personen_Affiliation_Start_Schluss_Org_ID.xlsx"

df_affiliations = pd.read_excel(DATA_RAW / expected_file)

print("Reloaded df_affiliations from:", expected_file)
print("Columns:", list(df_affiliations.columns))
display(df_affiliations.head(3))
```

```
Reloaded df_affiliations from: Personen_Affiliation_Start_Schluss_Org_ID.xlsx
Columns: ['Name', 'Person-Pure-ID', 'Organisationseinheit', 'Startdatum',
```

```
'Schlussdatum', 'Frühere Organisationseinheit', 'Pure-ID', 'Primär
Organisationseinheit', 'Org-Pure-ID']
```

```
                        Name  Person-Pure-ID  \
0  Bijlsma-Frankema, Katinka        80543515
1            Colak, Muhammed        80146317
2              Riener, Jakob        79987790

                                Organisationseinheit          Startdatum  \
0  Interdisziplinäres Institut für verhaltenswiss… 2026-01-09 12:00:00
1                         Statistics and Mathematics 2026-01-05 12:00:00
2                    Entrepreneurship und Innovation 2026-01-07 12:00:00

  Schlussdatum Frühere Organisationseinheit Pure-ID  \
0          NaT                          NaN     NaN
1          NaT                          NaN     NaN
2          NaT                          NaN     NaN

                         Primär Organisationseinheit  Org-Pure-ID
0  Interdisziplinäres Institut für verhaltenswiss…   16470441.0
1                         Statistics and Mathematics   16474420.0
2                    Entrepreneurship und Innovation   16474183.0
```

2.2 2.2 - ID Normalization and Structural Validation

Working copies of all relevant datasets are created. Identifier columns are standardized to string format to ensure robust merging across datasets.

[13]:
```python
cit_ext = df_cit.copy()
aff_ext = df_affiliations.copy()
org_ext = df_org_structure.copy()

cit_ext.columns = cit_ext.columns.str.strip()
aff_ext.columns = aff_ext.columns.str.strip()
org_ext.columns = org_ext.columns.str.strip()

def to_id_str(x):
    if pd.isna(x):
        return pd.NA
    try:
        fx = float(x)
        if fx.is_integer():
            return str(int(fx))
    except Exception:
        pass
    s = str(x).strip()
    if s.endswith(".0"):
        s = s[:-2]
```

```
    return s

PERSON_COL = "Person-Pure-ID"
ORG_COL = "Org-Pure-ID"
CHILD_COL = "Child-Pure-ID"

if PERSON_COL not in aff_ext.columns:
    raise KeyError(f"{PERSON_COL} not found in df_affiliations")

if ORG_COL not in aff_ext.columns:
    raise KeyError(f"{ORG_COL} not found in df_affiliations")

if CHILD_COL not in org_ext.columns:
    raise KeyError(f"{CHILD_COL} not found in df_org_structure")

if "pub_id" not in cit_ext.columns or "person_id" not in cit_ext.columns:
    raise KeyError("pub_id or person_id missing in df_cit")

cit_ext["pub_id"] = cit_ext["pub_id"].apply(to_id_str)
cit_ext["person_id"] = cit_ext["person_id"].apply(to_id_str)

aff_ext[PERSON_COL] = aff_ext[PERSON_COL].apply(to_id_str)
aff_ext[ORG_COL] = aff_ext[ORG_COL].apply(to_id_str)

org_ext[CHILD_COL] = org_ext[CHILD_COL].apply(to_id_str)

print("ID normalization complete.")
print("cit_ext shape:", cit_ext.shape)
print("aff_ext shape:", aff_ext.shape)
print("org_ext shape:", org_ext.shape)
```

```
ID normalization complete.
cit_ext shape: (7351, 10)
aff_ext shape: (47115, 9)
org_ext shape: (186, 15)
```

2.3 2.3 - Link Publication–Person Records to Affiliations

Citation records at the publication–person level are merged with the affiliation table via the person Pure ID. This step attaches organization identifiers and time information needed for ID-based organizational assignment.

[14]:
```
PERSON_COL = "Person-Pure-ID"
ORG_COL = "Org-Pure-ID"

aff_min = aff_ext[[PERSON_COL, ORG_COL, "Startdatum", "Schlussdatum", "Primär␣
↪Organisationseinheit", "Organisationseinheit"]].copy()
```

```
pub_person_org = cit_ext.merge(
    aff_min,
    left_on="person_id",
    right_on=PERSON_COL,
    how="left"
)

print("pub_person_org shape:", pub_person_org.shape)
print("Share rows with Org-Pure-ID:", pub_person_org[ORG_COL].notna().mean())

display(pub_person_org[["pub_id", "person_id", ORG_COL, "Organisationseinheit",␣
↪"Startdatum", "Schlussdatum"]].head(10))
```

```
pub_person_org shape: (38312, 16)
Share rows with Org-Pure-ID: 0.6778816036750888
```

```
     pub_id person_id Org-Pure-ID  \
0  79626455  16960296    16474458
1  79626455  16960296    16474458
2  79626455  16960296    16474458
3  79626455  16960296    16474458
4  79626455  16960296    16474458
5  79626455  16960296    16474458
6  79626455  16960296    16474458
7  79626455      <NA>         NaN
8  79626455      <NA>         NaN
9  79626455      <NA>         NaN

                              Organisationseinheit          Startdatum  \
0    Wirtschaftsinformatik und Operations Management 2025-10-01 12:00:00
1    Wirtschaftsinformatik und Operations Management 2022-10-01 12:00:00
2                                  Complex Networks 2011-03-01 12:00:00
3                                    Kryptoökonomie 2011-08-01 12:00:00
4     Data, Process and Knowledge Management (Sabou) 2011-08-01 12:00:00
5  Kompetenzzentrum für Sustainability Transforma… 2011-08-01 12:00:00
6             Data, Process and Knowledge Management 2021-06-01 12:00:00
7                                                NaN                 NaT
8                                                NaN                 NaT
9                                                NaN                 NaT

         Schlussdatum
0                 NaT
1 2025-09-30 12:00:00
2 2011-07-31 12:00:00
3 2021-03-31 12:00:00
4 2021-03-31 12:00:00
5 2021-03-31 12:00:00
6 2022-09-30 12:00:00
```

```
7                 NaT
8                 NaT
9                 NaT
```

2.4 2.4 - Baseline Mapping from Organization ID to Child Unit

Organization Pure IDs are mapped to child units using the organizational structure table. This baseline mapping provides an initial coverage estimate before applying the effective ID fallback strategy.

[15]:
```
CHILD_COL = "Child-Pure-ID"

org_min = org_ext[["Child", CHILD_COL]].copy()

pub_person_child = pub_person_org.merge(
    org_min,
    left_on=ORG_COL,
    right_on=CHILD_COL,
    how="left"
)

print("pub_person_child shape:", pub_person_child.shape)
print("Share rows mapped to Child-Pure-ID:",
      pub_person_child[CHILD_COL].notna().mean())

display(
    pub_person_child[[
        "pub_id",
        "person_id",
        ORG_COL,
        CHILD_COL,
        "Child"
    ]].head(10)
)
```

```
pub_person_child shape: (38312, 18)
Share rows mapped to Child-Pure-ID: 0.6727135101273752
```

```
     pub_id person_id Org-Pure-ID Child-Pure-ID  \
0  79626455  16960296    16474458      16474458
1  79626455  16960296    16474458      16474458
2  79626455  16960296    16474458      16474458
3  79626455  16960296    16474458      16474458
4  79626455  16960296    16474458      16474458
5  79626455  16960296    16474458      16474458
6  79626455  16960296    16474458      16474458
7  79626455      <NA>         NaN           NaN
8  79626455      <NA>         NaN           NaN
9  79626455      <NA>         NaN           NaN
```

```
                                                Child
0  Wirtschaftsinformatik und Operations Management
1  Wirtschaftsinformatik und Operations Management
2  Wirtschaftsinformatik und Operations Management
3  Wirtschaftsinformatik und Operations Management
4  Wirtschaftsinformatik und Operations Management
5  Wirtschaftsinformatik und Operations Management
6  Wirtschaftsinformatik und Operations Management
7                                                NaN
8                                                NaN
9                                                NaN
```

2.5 2.5 - Derive Effective Organization ID via Fallback Strategy

Missing organization Pure IDs in the affiliation table are addressed by falling back to the alternative Pure ID field. This yields an effective organization identifier with near-complete coverage for downstream mapping.

[16]:
```
aff_test = aff_ext.copy()

if "Pure-ID" in aff_test.columns:
    aff_test["Pure-ID"] = aff_test["Pure-ID"].apply(to_id_str)

aff_test["effective_org_pure_id"] = aff_test["Org-Pure-ID"]

if "Pure-ID" in aff_test.columns:
    aff_test.loc[
        aff_test["effective_org_pure_id"].isna(),
        "effective_org_pure_id"
    ] = aff_test["Pure-ID"]

print("Original Org-Pure-ID coverage:",
      aff_ext["Org-Pure-ID"].notna().mean())

print("Coverage after fallback to Pure-ID:",
      aff_test["effective_org_pure_id"].notna().mean())
```

```
Original Org-Pure-ID coverage: 0.4286108458028229
Coverage after fallback to Pure-ID: 0.9997665287063567
```

2.6 2.6 - Validate Effective Organization ID Coverage

The effective organization identifier is validated by reporting coverage after applying the fallback strategy. A small preview is shown to document the constructed ID fields.

[17]:
```
aff_ext = aff_ext.copy()

if "Pure-ID" in aff_ext.columns:
```

```
    aff_ext["Pure-ID"] = aff_ext["Pure-ID"].apply(to_id_str)

aff_ext["effective_org_pure_id"] = aff_ext["Org-Pure-ID"]

aff_ext.loc[
    aff_ext["effective_org_pure_id"].isna(),
    "effective_org_pure_id"
] = aff_ext["Pure-ID"]

print("Effective ID coverage:",
      aff_ext["effective_org_pure_id"].notna().mean())

display(
    aff_ext[[
        "Person-Pure-ID",
        "Org-Pure-ID",
        "Pure-ID",
        "effective_org_pure_id"
    ]].head(10)
)
```

```
Effective ID coverage: 0.9997665287063567
```

	Person-Pure-ID	Org-Pure-ID	Pure-ID	effective_org_pure_id
0	80543515	16470441	<NA>	16470441
1	80146317	16474420	<NA>	16474420
2	79987790	16474183	<NA>	16474183
3	79987774	16467282	<NA>	16467282
4	79987766	<NA>	<NA>	<NA>
5	79987758	16467033	<NA>	16467033
6	79872851	44599087	<NA>	44599087
7	79799914	16467259	<NA>	16467259
8	79799893	69919703	<NA>	69919703
9	79799887	16468660	<NA>	16468660

2.7 2.7 - Merge Effective Organization IDs and Assess Child-Level Match Rate

Publication–person records are merged with the affiliation table to attach the effective organization identifier. Coverage is reported and the share of effective IDs that directly match child-level Pure IDs is quantified.

[18]:
```
child_ids = set(org_ext["Child-Pure-ID"].dropna())

pub_person_org = cit_ext.merge(
    aff_ext[["Person-Pure-ID", "effective_org_pure_id"]],
    left_on="person_id",
    right_on="Person-Pure-ID",
    how="left"
```

```
)

share_effective_present = pub_person_org["effective_org_pure_id"].notna().mean()

share_matching_child = pub_person_org.loc[
    pub_person_org["effective_org_pure_id"].notna(),
    "effective_org_pure_id"
].isin(child_ids).mean()

print("Share rows with effective ID:", share_effective_present)
print("Share of effective IDs matching Child-Pure-ID:", share_matching_child)
```

```
Share rows with effective ID: 0.8863541449154312
Share of effective IDs matching Child-Pure-ID: 0.8283467813180988
```

2.8 2.8 - Inspect Resulting Column Structure

The column structure of the publication–person–organization table is inspected to ensure that all required identifiers and metadata fields are available for subsequent aggregation steps.

[19]:
```
print(pub_person_child.columns.tolist())
```

```
['pub_year', 'org_unit', 'Personen', 'pub_type', 'pub_id', 'Name', 'person_id',
'metric_name', 'metric_value', 'citations', 'Person-Pure-ID', 'Org-Pure-ID',
'Startdatum', 'Schlussdatum', 'Primär Organisationseinheit',
'Organisationseinheit', 'Child', 'Child-Pure-ID']
```

2.9 2.9 - Publication Coverage at Child Level

The share of publications with at least one valid child-level organization assignment is calculated. This indicates the proportion of publications usable for subsequent organizational analysis.

[20]:
```
child_col = "Child-Pure-ID"

mapped = pub_person_child.dropna(subset=[child_col])

pubs_total = pub_person_child["pub_id"].nunique()
pubs_mapped = mapped["pub_id"].nunique()

print("Total publications:", pubs_total)
print("Publications with at least one child mapping:", pubs_mapped)
print("Share of publications usable:", pubs_mapped / pubs_total)
```

```
Total publications: 1885
Publications with at least one child mapping: 1584
Share of publications usable: 0.8403183023872679
```

3 3 - Construction of Final Child-Level Analysis Table

3.1 3.1 - Build Deduplicated Child-Level Publication Table

A final child-level publication table is constructed by retaining unique combinations of publication and child unit. Each publication is counted once per child unit, forming the basis for metric computation.

[21]:
```
child_col = "Child-Pure-ID"

mapped = pub_person_child.dropna(subset=[child_col]).copy()

df_union_child = (
    mapped[["pub_id", child_col, "Child", "citations"]]
    .drop_duplicates(subset=[child_col, "pub_id"])
    .rename(columns={
        child_col: "child_pure_id",
        "Child": "child_name"
    })
)

print("df_union_child shape:", df_union_child.shape)
display(df_union_child.head(10))
```

```
df_union_child shape: (1753, 4)
```

	pub_id	child_pure_id	child_name	citations
0	79626455	16474458	Wirtschaftsinformatik und Operations Management	2.0
10	79462810	16474420	Statistics and Mathematics	1.0
26	79462274	16470002	Transportwirtschaft und Logistik (Kummer)	0.0
32	79323578	16472824	Project Management Group	0.0
53	79300411	16472824	Project Management Group	1.0
103	79299890	16472824	Project Management Group	0.0
135	79257647	16468660	Sprache und Diskurs in der Wirtschaft	0.0
145	79178077	60155265	Data, Energy, and Sustainability	0.0
173	79107299	16472366	Public Management und Governance	0.0
177	79067571	16467036	Ethics in Management	1.0

4 4 - Organizational Level Computation

4.1 4.1 - Publication Count at Child Level

The number of unique publications is computed for each child-level organizational unit. Each publication is counted once per child unit.

[22]:
```
pub_count_child = (
    df_union_child
    .groupby(["child_pure_id", "child_name"])["pub_id"]
    .nunique()
    .reset_index(name="publication_count")
)
```

4.2 4.2 - Total Citations at Child Level

The total number of citations is aggregated for each child-level organizational unit by summing citation counts across all assigned publications.

[23]:
```
citations_child = (
    df_union_child
    .groupby(["child_pure_id", "child_name"])["citations"]
    .sum()
    .reset_index(name="total_citations")
)
```

4.3 4.3 - Citations per Publication (CPP)

Publication count and total citation metrics are combined to compute citations per publication (CPP) for each child-level organizational unit.

[24]:
```
metrics_child = pub_count_child.merge(
    citations_child,
    on=["child_pure_id", "child_name"],
    how="left"
)

metrics_child["cpp"] = (
    metrics_child["total_citations"] /
    metrics_child["publication_count"]
)
```

4.4 4.4 - h-Index at Child Level

The h-index is computed for each child-level organizational unit based on the distribution of citation counts across its publications.

[25]:
```python
def h_index(series):
    s = sorted(series, reverse=True)
    return max((min(i + 1, c) for i, c in enumerate(s)), default=0)

h_child = (
    df_union_child
    .groupby(["child_pure_id", "child_name"])["citations"]
    .apply(h_index)
    .reset_index(name="h_index")
)
```

4.5 4.5 - g-Index at Child Level

The g-index is computed for each child-level organizational unit, placing greater emphasis on highly cited publications than the h-index.

[26]:
```python
def g_index(series):
    s = sorted(series, reverse=True)
    cumulative = 0
    g = 0
    for i, c in enumerate(s):
        cumulative += c
        if cumulative >= (i + 1) ** 2:
            g = i + 1
    return g

g_child = (
    df_union_child
    .groupby(["child_pure_id", "child_name"])["citations"]
    .apply(g_index)
    .reset_index(name="g_index")
)
```

4.6 4.6 - Consolidate Child-Level Metrics Table

All child-level indicators (publication count, citations, CPP, h-index, and g-index) are merged into a single table. The table is sorted for inspection and reporting.

[27]:
```python
metrics_child = (
    metrics_child
    .merge(h_child, on=["child_pure_id", "child_name"])
    .merge(g_child, on=["child_pure_id", "child_name"])
)

metrics_child = metrics_child.sort_values(
    "total_citations",
    ascending=False
```

```
).reset_index(drop=True)

display(metrics_child.head(15))
```

```
   child_pure_id                                        child_name  \
0       16468227                                 Ecological Economics
1       16467176           Department für Volkswirtschaft (Prettner)
2       16468980                                              Novy
3       16474037                                   Complex Networks
4       16468157  Data, Process and Knowledge Management (Polleres)
5       16474420                          Statistics and Mathematics
6       16472366                 Public Management und Governance
7       16474165  Responsibility and Sustainability in Global Bu…
8       16469483                               Marketing-Management
9       16474458    Wirtschaftsinformatik und Operations Management
10      16473196                                 Personalmanagement
11      16468088            Gesellschaftswandel und Nachhaltigkeit
12      16474072                    Finance, Banking and Insurance
13      16468141                    Produktionsmanagement (Reiner)
14      16472431                                 Organization Studies

    publication_count  total_citations        cpp  h_index  g_index
0                  73           1473.0  20.178082     19.0       36
1                  60           1027.0  17.116667     12.0       31
2                  48            559.0  11.645833     12.0       22
3                  21            510.0  24.285714      7.0       21
4                  18            500.0  27.777778      6.0       18
5                 116            499.0   4.301724     10.0       17
6                  63            423.0   6.714286     13.0       18
7                  16            404.0  25.250000      9.0       16
8                  24            387.0  16.125000      9.0       19
9                  42            381.0   9.071429     12.0       16
10                 19            349.0  18.368421      9.0       18
11                 17            346.0  20.352941      9.0       17
12                 46            340.0   7.391304     11.0       16
13                 34            334.0   9.823529     11.0       17
14                 44            332.0   7.545455     11.0       16
```

4.7 4.7 - Inspect Institute-Level Coverage in the Organization Structure

Institute identifiers are inspected to understand hierarchy coverage and potential missing mappings. This provides context for institute-level aggregation and explains why some child-level publications may not roll up to institutes.

[28]:
```
print("Institute-Pure-ID non-null count:",
      org_ext["Institute-Pure-ID"].notna().sum())

print("Unique Institute-Pure-ID:",
```

```
      org_ext["Institute-Pure-ID"].dropna().nunique())

display(
    org_ext.loc[
        org_ext["Institute-Pure-ID"].notna(),
        ["Child", "Institute", "Institute-Pure-ID", "Department"]
    ].head(10)
)
```

```
Institute-Pure-ID non-null count: 175
Unique Institute-Pure-ID: 91
```

	Child	Institute	Institute-Pure-ID	Department
11	Financial Research	Financial Research	16474692.0	Finance, Accounting and Statistics
12	Statistics and Mathematics	Statistics and Mathematics	16474420.0	Finance, Accounting and Statistics
13	Finance, Banking and Insurance	Finance, Banking and Insurance	16474072.0	Finance, Accounting and Statistics
14	Department-Office FASd4	Department-Office FASd4	16472554.0	Finance, Accounting and Statistics
15	Accounting and Auditing	Accounting and Auditing	16470672.0	Finance, Accounting and Statistics
16	Familienunternehmen	Familienunternehmen	46081346.0	Management
17	Institut für Gender und Diversität in Organisa…	Institut für Gender und Diversität in Organisa…	16473399.0	Management
18	Personalmanagement	Personalmanagement	16473196.0	Management
19	Wirtschaftspädagogik	Wirtschaftspädagogik	16473170.0	Management
20	Change Management und Management Development	Change Management und Management Development	16473108.0	Management

4.8 4.8 - Create Child-to-Institute Mapping Table

A deduplicated mapping table is constructed to roll up child-level units to their corresponding institutes. This mapping is used to aggregate publications from child level to institute level.

[29]:
```
child_to_inst = (
    org_ext[
        ["Child-Pure-ID", "Institute", "Institute-Pure-ID"]
    ]
    .dropna(subset=["Institute-Pure-ID"])
    .drop_duplicates()
)

print("Child→Institute mapping rows:", child_to_inst.shape)
display(child_to_inst.head(10))
```

```
Child→Institute mapping rows: (172, 3)
```

	Child-Pure-ID	Institute	Institute-Pure-ID
11	16474692	Financial Research	16474692.0
12	16474420	Statistics and Mathematics	16474420.0
13	16474072	Finance, Banking and Insurance	16474072.0
14	16472554	Department-Office FASd4	16472554.0
15	16470672	Accounting and Auditing	16470672.0
16	46081346	Familienunternehmen	46081346.0
17	16473399	Institut für Gender und Diversität in Organisa…	16473399.0
18	16473196	Personalmanagement	16473196.0
19	16473170	Wirtschaftspädagogik	16473170.0
20	16473108	Change Management und Management Development	16473108.0

4.9 4.9 - Construct Deduplicated Institute-Level Publication Table

Child-level publication assignments are rolled up to institutes via the child-to-institute mapping. The resulting institute-level table is deduplicated so that each publication is counted once per institute.

[30]:
```
df_union_institute = df_union_child.merge(
    child_to_inst,
    left_on="child_pure_id",
    right_on="Child-Pure-ID",
    how="inner"
)

df_union_institute = (
    df_union_institute[
        ["pub_id", "Institute-Pure-ID", "Institute", "citations"]
    ]
    .drop_duplicates(subset=["Institute-Pure-ID", "pub_id"])
)

print("Child-level rows:", df_union_child.shape[0])
print("Institute-level rows:", df_union_institute.shape[0])
print("Unique pubs (child level):", df_union_child["pub_id"].nunique())
print("Unique pubs (institute level):", df_union_institute["pub_id"].nunique())
```

```
Child-level rows: 1753
Institute-level rows: 1626
Unique pubs (child level): 1584
Unique pubs (institute level): 1514
```

4.10 4.10 - Publication Count at Institute Level

The number of unique publications is computed for each institute by counting distinct publications assigned through the child-to-institute aggregation.

[31]:
```
pub_count_inst = (
    df_union_institute
    .groupby(["Institute-Pure-ID", "Institute"])["pub_id"]
    .nunique()
    .reset_index(name="publication_count")
)
```

4.11 4.11 - Total Citations at Institute Level

Total citation counts are aggregated for each institute by summing citation values across all assigned publications.

[32]:
```
citations_inst = (
    df_union_institute
    .groupby(["Institute-Pure-ID", "Institute"])["citations"]
    .sum()
    .reset_index(name="total_citations")
)
```

4.12 4.12 - Citations per Publication (CPP) at Institute Level

Publication counts and total citation metrics are merged to compute citations per publication (CPP) for each institute.

[33]:
```
metrics_inst = pub_count_inst.merge(
    citations_inst,
    on=["Institute-Pure-ID", "Institute"]
)

metrics_inst["cpp"] = (
    metrics_inst["total_citations"] /
    metrics_inst["publication_count"]
)
```

4.13 4.13 - h-Index at Institute Level

The h-index is computed for each institute based on the citation distribution of its associated publications.

[34]:
```
h_inst = (
    df_union_institute
    .groupby(["Institute-Pure-ID", "Institute"])["citations"]
    .apply(h_index)
    .reset_index(name="h_index")
)
```

4.14 4.14 - g-Index at Institute Level

The g-index is computed for each institute, emphasizing highly cited publications within the aggregated institute-level citation distribution.

[35]:
```
g_inst = (
    df_union_institute
    .groupby(["Institute-Pure-ID", "Institute"])["citations"]
    .apply(g_index)
    .reset_index(name="g_index")
)
```

4.15 4.15 - Consolidate Institute-Level Metrics Table

All institute-level indicators (publication count, total citations, CPP, h-index, and g-index) are merged into a single institute-level metrics table. The table is sorted by total citation volume to facilitate ranking and comparison across institutes.

[36]:
```
metrics_inst = (
    metrics_inst
    .merge(h_inst, on=["Institute-Pure-ID", "Institute"])
```

```
    .merge(g_inst, on=["Institute-Pure-ID", "Institute"])
    .sort_values("total_citations", ascending=False)
    .reset_index(drop=True)
)

display(metrics_inst.head(15))
```

```
    Institute-Pure-ID                                          Institute  \
0           16468227.0                                 Ecological Economics
1           16467176.0                Department für Volkswirtschaft (Prettner)
2           62739340.0  Spatial and Social-Ecological Transformations …
3           16469751.0                   Data, Process and Knowledge Management
4           38679068.0  Responsibility and Sustainability in Global Bu…
5           16474037.0                                     Complex Networks
6           16474420.0                           Statistics and Mathematics
7           16469946.0                             Strategisches Management
8           16472366.0                    Public Management und Governance
9           16474056.0                               Produktionsmanagement
10          16473631.0                              International Business
11          16469483.0                                Marketing-Management
12          16469197.0                    Transportwirtschaft und Logistik
13          16470672.0                             Accounting and Auditing
14          16473196.0                                  Personalmanagement

    publication_count  total_citations        cpp  h_index  g_index
0                  73           1473.0  20.178082       19       36
1                  60           1027.0  17.116667       12       31
2                  67            695.0  10.373134       13       23
3                  38            628.0  16.526316        9       24
4                  29            519.0  17.896552       12       22
5                  21            510.0  24.285714        7       21
6                 116            499.0   4.301724       10       17
7                  38            431.0  11.342105       11       19
8                  63            423.0   6.714286       13       18
9                  46            410.0   8.913043       11       18
10                 57            409.0   7.175439       11       17
11                 24            387.0  16.125000        9       19
12                 43            363.0   8.441860       13       16
13                 38            358.0   9.421053        6       18
14                 19            349.0  18.368421        9       18
```

4.16 4.16 - Number of Child Units per Institute

The number of distinct child-level units assigned to each institute is computed. This provides structural context on the hierarchical composition of institutes and helps interpret aggregated institute-level metrics.

[37]:
```
inst_child_count = (
    child_to_inst
    .groupby("Institute-Pure-ID")["Child-Pure-ID"]
    .nunique()
    .reset_index(name="child_count")
    .sort_values("child_count", ascending=False)
)

display(inst_child_count.head(10))
```

```
    Institute-Pure-ID  child_count
80         16474368.0           13
31         16469061.0            7
81         16474401.0            7
44         16470672.0            7
29         16468874.0            6
62         16473631.0            6
32         16469083.0            6
20         16467710.0            5
40         16470063.0            4
36         16469751.0            4
```

4.17 4.17 - Create Child-to-Department Mapping Table

A deduplicated mapping table is constructed to roll up child-level units to their corresponding departments. This mapping enables aggregation of publication data from child level to department level.

[38]:
```
org_ext["Department-Pure-ID"] = org_ext["Department-Pure-ID"].apply(to_id_str)

child_to_dept = (
    org_ext[
        ["Child-Pure-ID", "Department", "Department-Pure-ID"]
    ]
    .dropna(subset=["Department-Pure-ID"])
    .drop_duplicates()
)

print("Child→Department mapping rows:", child_to_dept.shape)
display(child_to_dept.head(10))
```

```
Child→Department mapping rows: (182, 3)
```

```
  Child-Pure-ID                                 Department  \
0      16470647          Finance, Accounting and Statistics
1      16470779                                  Management
2      16470568                                   Marketing
3      16470793          Öffentliches Recht und Steuerrecht
4      16470484                                  Privatrecht
```

```
5      16474467                                    Sozioökonomie
6      16470846                          Strategy and Innovation
7      16474544                                  Volkswirtschaft
8      16474533                                      Welthandel
9      16474458  Wirtschaftsinformatik und Operations Management

  Department-Pure-ID
0            16470647
1            16470779
2            16470568
3            16470793
4            16470484
5            16474467
6            16470846
7            16474544
8            16474533
9            16474458
```

4.18 4.18 - Construct Deduplicated Department-Level Publication Table

Child-level publication assignments are rolled up to departments via the child-to-department mapping. The resulting department-level table is deduplicated so that each publication is counted once per department.

[39]:
```
df_union_department = df_union_child.merge(
    child_to_dept,
    left_on="child_pure_id",
    right_on="Child-Pure-ID",
    how="inner"
)

df_union_department = (
    df_union_department[
        ["pub_id", "Department-Pure-ID", "Department", "citations"]
    ]
    .drop_duplicates(subset=["Department-Pure-ID", "pub_id"])
    .rename(columns={
        "Department-Pure-ID": "department_pure_id",
        "Department": "department_name"
    })
)

print("Department-level rows:", df_union_department.shape[0])
print("Unique pubs (department level):", df_union_department["pub_id"].
 ↪nunique())

display(df_union_department.head())
```

```
Department-level rows: 1636
Unique pubs (department level): 1577
```

```
     pub_id department_pure_id  \
0  79626455           16474458
1  79462810           16470647
2  79462274           16474533
3  79323578           16470846
4  79300411           16470846

                                  department_name  citations
0  Wirtschaftsinformatik und Operations Management        2.0
1               Finance, Accounting and Statistics        1.0
2                                       Welthandel        0.0
3                          Strategy and Innovation        0.0
4                          Strategy and Innovation        1.0
```

4.19 4.19 - Publication Count at Department Level

The number of unique publications is calculated for each department by counting distinct publications assigned through the child-to-department aggregation.

[40]:
```
pub_count_dept = (
    df_union_department
    .groupby(["department_pure_id", "department_name"])["pub_id"]
    .nunique()
    .reset_index(name="publication_count")
)
```

4.20 4.20 - Total Citations at Department Level

Total citation counts are aggregated for each department by summing citation values across all associated publications.

[41]:
```
citations_dept = (
    df_union_department
    .groupby(["department_pure_id", "department_name"])["citations"]
    .sum()
    .reset_index(name="total_citations")
)
```

4.21 4.21 - Citations per Publication (CPP) at Department Level

Publication counts and total citation metrics are merged to compute citations per publication (CPP) for each department.

[42]:
```
metrics_dept = pub_count_dept.merge(
    citations_dept,
    on=["department_pure_id", "department_name"]
```

```
)

metrics_dept["cpp"] = (
    metrics_dept["total_citations"] /
    metrics_dept["publication_count"]
)
```

4.22 4.22 - h-Index at Department Level

The h-index is computed for each department based on the citation distribution of its associated publications.

[43]:
```
h_dept = (
    df_union_department
    .groupby(["department_pure_id", "department_name"])["citations"]
    .apply(h_index)
    .reset_index(name="h_index")
)
```

4.23 4.23 - g-Index at Department Level

The g-index is computed for each department, placing greater emphasis on highly cited publications within the department-level citation distribution.

[44]:
```
g_dept = (
    df_union_department
    .groupby(["department_pure_id", "department_name"])["citations"]
    .apply(g_index)
    .reset_index(name="g_index")
)
```

4.24 4.24 - Consolidate Department-Level Metrics Table

All department-level indicators (publication count, total citations, CPP, h-index, and g-index) are merged into a single department-level metrics table. The table is sorted by total citation volume to enable ranking and comparative analysis across departments.

[45]:
```
metrics_dept = (
    metrics_dept
    .merge(h_dept, on=["department_pure_id", "department_name"])
    .merge(g_dept, on=["department_pure_id", "department_name"])
    .sort_values("total_citations", ascending=False)
    .reset_index(drop=True)
)

display(metrics_dept.head(15))
```

	department_pure_id	department_name
0	16474467	Sozioökonomie
1	16474544	Volkswirtschaft
2	16474458	Wirtschaftsinformatik und Operations Management
3	16470779	Management
4	16474533	Welthandel
5	16470568	Marketing
6	16470647	Finance, Accounting and Statistics
7	16470846	Strategy and Innovation
8	16474501	Wirtschaftskommunikation
9	16474467	Management
10	16470484	Privatrecht
11	16470793	Öffentliches Recht und Steuerrecht

	publication_count	total_citations	cpp	h_index	g_index
0	300	3219.0	10.730000	27	45
1	208	1949.0	9.370192	18	37
2	178	1652.0	9.280899	17	32
3	212	1545.0	7.287736	17	27
4	153	1422.0	9.294118	20	29
5	105	1208.0	11.504762	17	30
6	201	1197.0	5.955224	15	27
7	134	1036.0	7.731343	14	25
8	80	263.0	3.287500	9	12
9	21	186.0	8.857143	7	13
10	17	51.0	3.000000	5	6
11	27	24.0	0.888889	2	3

5 5 - Experiments

5.1 5.1 - Top 10 Departments by h-Index

The ten departments with the highest h-index values.

[46]:
```
top10_dept_h = (
    metrics_dept
    .sort_values("h_index", ascending=False)
    .head(10)
    .reset_index(drop=True)
)

display(
    top10_dept_h[
        [
            "department_pure_id",
            "department_name",
            "publication_count",
            "total_citations",
```

```
            "cpp",
            "h_index",
            "g_index"
        ]
    ]
)
```

```
  department_pure_id                                department_name  \
0           16474467                                  Sozioökonomie
1           16474533                                     Welthandel
2           16474544                                Volkswirtschaft
3           16474458  Wirtschaftsinformatik und Operations Management
4           16470779                                     Management
5           16470568                                      Marketing
6           16470647           Finance, Accounting and Statistics
7           16470846                     Strategy and Innovation
8           16474501                   Wirtschaftskommunikation
9           16474467                                     Management

   publication_count  total_citations        cpp  h_index  g_index
0                300           3219.0  10.730000       27       45
1                153           1422.0   9.294118       20       29
2                208           1949.0   9.370192       18       37
3                178           1652.0   9.280899       17       32
4                212           1545.0   7.287736       17       27
5                105           1208.0  11.504762       17       30
6                201           1197.0   5.955224       15       27
7                134           1036.0   7.731343       14       25
8                 80            263.0   3.287500        9       12
9                 21            186.0   8.857143        7       13
```

5.2 5.2 - Top 10 Departments by g-index

The ten departments with the highest g-index values.

[47]:
```
top10_dept_g = (
    metrics_dept
    .sort_values("g_index", ascending=False)
    .head(10)
    .reset_index(drop=True)
)

display(
    top10_dept_g[
        [
            "department_pure_id",
            "department_name",
            "publication_count",
```

```
            "total_citations",
            "cpp",
            "h_index",
            "g_index"
        ]
    ]
)
```

	department_pure_id	department_name
0	16474467	Sozioökonomie
1	16474544	Volkswirtschaft
2	16474458	Wirtschaftsinformatik und Operations Management
3	16470568	Marketing
4	16474533	Welthandel
5	16470779	Management
6	16470647	Finance, Accounting and Statistics
7	16470846	Strategy and Innovation
8	16474467	Management
9	16474501	Wirtschaftskommunikation

	publication_count	total_citations	cpp	h_index	g_index
0	300	3219.0	10.730000	27	45
1	208	1949.0	9.370192	18	37
2	178	1652.0	9.280899	17	32
3	105	1208.0	11.504762	17	30
4	153	1422.0	9.294118	20	29
5	212	1545.0	7.287736	17	27
6	201	1197.0	5.955224	15	27
7	134	1036.0	7.731343	14	25
8	21	186.0	8.857143	7	13
9	80	263.0	3.287500	9	12