RCE (Rationale-Cogency-Extent) Criterion Unravels Features affecting Citation Impact of Top-ranked Systematic Literature Reviews: Leaving the impression…*is all you need*.

Orošnjak, Marko[1][0000-0003-0929-1425], **Štrbac, Branko**[2*][0000-0003-3892-2767], **Vulanović, Srđan**[1][0000-0002-0527-505X], **Runje, Biserka**[3][0000-0002-7786-0864], **Horvatić Novak, Amalija**[3][0000-0003-4665-255X], **Razumić, Andrej**[3][0000-0002-1257-0233]

[1]*University of Novi Sad, Faculty of Technical Sciences, Department of Industrial Engineering and Management, 21000 Novi Sad, Serbia*

[2*]*University of Novi Sad, Faculty of Technical Sciences, Department of Production Engineering, 21000 Novi Sad, Serbia*

[3]*University of Zagreb, Faculty of Mechanical Engineering and Naval Architecture, Department of Quality, 10000 Zagreb, Croatia*

corresponding author: strbacb@uns.ac.rs

Abstract: Hijacked from medical and health sciences, Systematic Literature Reviews (SLRs) are widely (ab)used in many scientific domains. Considering the ability to provide transparency and replicability of research results, many scientists consider an SLR a safe avenue for attaining scientific impact, given that the theoretical probability of acceptance is relatively high. Relying on dual analysis of Partial Least Squares Discriminant Analysis (PLS-DA) and Network Analysis (NA), the study identifies key features associated with citation impact within top-tier SLRs. Next, the study introduces Rationale, Cogency, and Extent (RCE) criterion for evaluating potential markers that predict citation impact using two case studies of SLRs from engineering domain. The findings suggest that the informal logic for starting a review significantly correlates with citation impact. Additionally, journal- and author-level metrics, along with RCE composite scores, display significant difference between top- and bottom-ranked SLRs. Through NA, reporting the quality assessment of studies (QATR) emerges as the most influential node within the RCE network. Despite its lack of direct correlation with citation impact, we conclude that QATR is a moderating variable. Finally, the study concludes that a well-articulated research question, alignment with existing evidence, and rigorous data use collectively serve as a blueprint for producing a high-quality SLR.

keywords: rationale-cogency-extent criterion; scientometrics; partial least squares discriminant analysis; network analysis; systematic literature reviews; citation impact;

1 Introduction

1.1 Background

Systematic Literature Reviews (SLRs) are recognised as gold standards in evidence-based practice mainly because they provide transparency and replicability (Han et al. 2020). Besides, assuring extensiveness (Bramer et al. 2016), impartiality (Furley and Goldschmied 2021), reducing bias (e.g., type I and II errors (Kung et al. 2010)), providing generalisability, consistency or inconsistency (Mulrow 1994) of summarised evidence is among many attributes that an SLR offers. Ultimately, these characteristics contribute to a study's overall reliability, making an SLR compelling apparatus in evidence-based synthesis. As a result, an SLR has become a strategic move (Schniedermann 2021) for increasing research impact considering the theoretical probability of acceptance (Montori et al. 2003), in addition to the number of citations it receives (Blümel and Schniedermann 2020; Knottnerus and Knottnerus 2009; Kousha and Thelwall 2023; Mäntylä and Garousi 2019; Tahamtan et al. 2016), is relatively high.

Bearing in mind what an SLR offers, many engaged in the mass production of SLRs (Ioannidis 2016; Page and Moher 2016), believing that it would provide a safe avenue for making a scientific footprint. The practice is even acquired by editors under pressure as a feasible strategy for maintaining and increasing the journal's position in the "impact factor game" (Blümel and Schniedermann 2020; Knottnerus and Knottnerus 2009). However, although many believe that SLRs are infallible, the validity and reliability of an SLR are highly dependent on the extent of summarised information (Garcia-Doval et al. 2017). Recognising the existence of potential issues of SLRs, several protocols emerged as a way to reduce bias in retrieval and evidence synthesis. For instance, PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher et al. 2015) and Quality Assessment Tools (QAT) (Memon et al. 2020) have became common practice to ensure trustworthiness (Paré et al. 2016) of reporting. Despite the presence of these guidelines, the proliferation (McColl 2022) and criticism of SLRs conduct pervade even from their proponents arguing that is contributing to the research waste (Roberts and Ker 2015; Uttley et al. 2023).

Raised concerns over the misuse of SLRs started in medical science (Yuan and Hunt 2009) but extended to business and management sciences (Coombes 2023) and library and information sciences (Soheili et al. 2022). The main question is raised over the rigorous modus operandi and quality of reporting of SLRs (Coombes 2023), followed by rationale of Research Questions (RQs) (Karunananthan et al. 2021), motivation for the research topic (Webster and Watson 2002), the relevance of content (Blümel and Schniedermann 2020), the argumentation scheme (Bornmann and Daniel 2008), study selection and retrieval (Yuan and Hunt 2009), and many others. Although such issues are inherently of great concern in health sciences (Kousha and Thelwall 2023), we have experienced the decline in quality in the engineering-based domain. As a response, we hypothesise that the main concerns in engineering-based SLRs' are attributed to (i) logic and rationale when starting a review, (ii) study design and used search strategy, and (iii) quality and transparency of reporting.

1.2 Research problem

Nowadays, the logical sequence for justifying the need for an SLR is built from tokenistic citations and ampliative rationale (Ghidalia et al. 2023). The prior is observed where authors only capture articles' meta-data (Oelen et al. 2020) for suggesting the importance of an SLR, while the latter exemplifies cases where authors explain the need for a review by simply implying that the new information should be genuinely more interesting because it is merely new – an argument "…*to the best of our knowledge, no SLR study is conducted*…" usually follows. Such statements can be analogous to the idea that the lack of proof is the proof itself, which is absurd. We argue that exposing and allocating gaps must be assured not by previous outcomes or lack of arguments but by looking at evidence at face value. In such instances, more weight is added because the sole aim of an SLR is not to synthesise recent findings but to appraise evidence behind outcomes within retrieved studies critically. Rhetorically, Blümel and Schniedermann (2020) argue that in order to establish a specific argumentation scheme for starting a review, the author(s) should have a compelling argumentation narrative (e.g., assumptive, contrastive) for rationalising the need for a problem to be investigated (Bornmann and Daniel 2008). The recent transition from *BigData* to *SmartData* (Triguero et al. 2019) pretty much sums up the importance of capturing relevant evidence instead of descriptive and ampliative reasoning for justifying the need for an SLR study.

Drawing preliminary inferences, we question recent SLRs' methodological rigour in engineering disciplines since many SLRs yield excessive information or, in some instances, lack robustness. This can be attributed to today's practice, where authors frequently introduce new terminology to reform existing concepts (Chawla 2020). Likewise, there is an apparent shift in the number of original scientific contributions to generating more SLRs (van der Braak et al. 2022) questioning the rigour of peer-review process. To contextualise such observations, consider the following statistics: (**i**) the expected number of SLRs in 2023 is 10k (Appendix 1), while by the end of 2025, the existing number of SLRs should double ($y_{pred} = 7^{-227}e^{0.262*\mathrm{Year}}$; $R^2 = 99.7\%$); (**ii**) at the moment there are at least 50 authors that produced >20 SLR studies (some topping over 50); (**iii**) Some Q1-Q2 journals published more than >200 SLRs in 2022 alone(!). Consequently, the mismatch between the published SLRs and their peer-review standard forced us to investigate existing SLR practices and delineate features that constitute high-quality SLR research. By synthesising insights from other high-quality SLR research, the present study seeks to preemptively identify and underscore potential pitfalls by mitigating the risk of publishing substandard systematic reviews.

It is difficult to answer what qualifies as "high-quality" research. Some posit that such property only qualified experts can judge (Kousha and Thelwall 2023). Others state that it must impact practice and grand challenges (Fitzgerald et al. 2019), which does not account where some, however, rely on the use of citations as an impact measure (Fitzgerald et al. 2019). Despite the plethora of research metrics (e.g., citations, altmetrics), unfortunately, the majority still lean on citation count as a quality measure of the research impact (Y.-S. Ho and Shekofteh 2021). This comes as no surprise since authors' (and/or institutions') impact is mainly measured through citations, consequently regarded as "scientific monetary value" (Rousseau et al. 2021). Driven by this metrics-focused ideology, some content that SLRs have become an ideal endeavour for seeking funding (Kousha and Thelwall 2023), rewards (Judge et al. 2007), or anyone wanting to improve their CV or impact score (Wormald and Evans 2018). However, as much as citations indicate scientific merit, even Nobel Laureates have uncited work (Glänzel et al. 2006). For that matter, while this study uses citation output as a measurement for assessing and delineating features associated with high-quality SLRs, the elaboration and interpretation of results are done with caution.

1.3 Literature review

The review by Tahamtan et al. (2016) shows that at least 28 factors contribute to the increase of citations of papers, dividing them into study factors, journal factors, and author factors. These factors consider standard variables that are researched within the existing scientometric domain. For instance, most studies suggest that JIF (Journal Impact Factor) (Bornmann and Leydesdorff 2017; Yu et al. 2014), length of the paper (Xie, Gong, Cheng, et al. 2019), number of authors (Cheng et al. 2017), and international and inter-institutional collaboration (Chen et al. 2023; So et al. 2015) are primarily associated with the rise of citations. For instance, Uthman et al. (2013) show the existence of a correlation between citations and authors ($r = 0.32$) and JIF ($r = 0.24$). Liskiewicz et al. (2021) suggest that mean citation count positively correlates with stated references ($r \cong 0.15$), paper length ($r \cong 0.15$) (Xie, Gong, Cheng, et al. 2019), and the number of collaborating authors (So et al. 2015). The evidence also suggest favouring open-access publishing over closed-access as an advantage regarding citation impact. Similarly, a study by Zong et al. (2020) shows that an open peer review policy increases citation impact by increasing reviewers' and authors' accountability. Wang et al. (2019) show that with the help of machine learning algorithms, bibliometric indices and alternative metrics are valuable predictors of an article's success, leading to an increase in citation impact.

The studies that allocate factors affecting the citation impact of literature reviews are limited ((Royle et al. 2013; Wagner et al. 2016, 2021). Wagner et al. (2021) used paper-level attributes (transparency, research agenda, topic popularity), author-level attributes (h-index), and journal-level (impact factor) as independent variables and scientific impact (number of citations) as dependent variables. The results of GLM (Generalized Linear Model) on four different review papers show a significant relationship between JIF, *h*-index and topic popularity. At the paper level, methodological transparency and a developed research agenda are most impactful in distinguishing between high-impact Information Sciences (IS) reviews. The study by Royle et al. (2013) shows the presence of a correlation between citations on one side and JIF ($r = 0.453$), 5-year JIF ($r = 0.444$), SCImago journal rank ($r = 0.438$), and the number of authors ($r = 0.215$) of an SLR, on the other. Also, they provide a comparison of the top 50 cited and bottom 50 cited SLRs showing significant differences ($p < 0.05$) with all factors (e.g., JIF, number of authors) except the number of pages. Next, the study conducted by Blümel and Schniedermann (2020) first emphasised the lack of knowledge about the citation patterns of SLRs. They provide exciting remarks on citation impact of review articles (de Almeida and Guimarães 2013; M. H.-C. Ho et al. 2017; Jokic and Ball 2006; Knottnerus and Knottnerus 2009; J. S. Liu and Kuan 2016) showing that the length of reference list correlates with citations, which is also found in other studies emphasising the "quid pro quo" approach (Grover et al. 2014) as the citable favour. Grover et al. (2014) study suggests the impact of universalistic and particularistic variables on citation impact, explicitly excluding methodological variables. This was also noticed in the review conducted by Tahamtan et al. (2016), where they highlighted the research gaps regarding the influence of methodological variables. Rhetorically, others discussed the lack of evidence of methodological-level factors on the citation impact (Fitzgerald et al. 2019). We assume that the underlying reason why little research is done on methodological-level factors is due to conflicting results. Some studies show the effect, while others show the lack of the effect (Tahamtan et al. 2016).

To justify the need to understand methodological-level factors influencing citation impact, we highlight an earlier remark given by Judge et al. (2007) stating that "*Regardless of the quality of an idea, the ability to draw inferences about a phenomenon is constrained by the quality of the methods used to gather data about it.*", suggesting that methodology carries most of the weight considering the quality of an outcome. Rhetorically, Patnode et al. (2015) synthesised existing issues of SLRs, showing that without exception, the quality assessment of individual studies is the central issue, followed by the comprehensiveness of literature search, selection criteria, and transparency.

Similarly, Yuan and Hunt (2009) identified seven signs of bad SLR practices: study design, study selection, quality assessment, heterogeneity of studies, excluding non-statistically significant studies, poor data handling, and sample size. From a more critical perspective, Wagner et al. (2021) emphasised that the scientific impact of reviews should not be confined to a positivist, commensurable perspective. However, it should capture all contributions, including disagreement with and refutation of previous ideas. Their study of 220 Information Science (IS) reviews concluded that methodological transparency and developed research agenda are the main contributing factors to the citation impact, aside from usual universal metrics such as JIF and h-index.

In summary, while there is an extensive amount of scientometric research dedicated to exploring journal and author-level factors associated with citation impact, there is a noticable gap in examining the role of methodological-level factors within review studies. This is especially important in SLRs since methodological rigor directly influences the perceived quality, i.e., validity and reliability of SLRs. Understanding these factors is crucial for delineating characteristics of high-quality SLRs. Therefore, we expose the gap and conduct a comprehensive analysis of methodological-level factors influencing the citation impact. The effort aimed at bridging this gap, particularly in understanding how these methodological factors shape and impact the quality of SLRs, is of our primary concern. Ultimately, our objective is to reinforce the importance of methodological rigor in producing high-quality SLR.

1.4 Aims and Objectives

In defense of the rationale, many argue that the RQ is pivotal (Booth 2006; Nishikawa-Pacher 2022) since it logically follows that conclusions can only be generated from a corpus of empirical studies. Although existing SLR protocols (e.g., PRISMA – Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and RQ frameworks (e.g., PICO – Patient Intervention Comparison Outcome) are used to scope and guide the review (Booth et al. 2016), the lack of appropriate search strategy is also the reason why many authors usually fail to provide a transparent and replicable results. Nonetheless, it becomes apparent that the RQ being asked, strategy used for extracting evidence and data used (Garcia-Doval et al. 2017) to answer the RQ (Munn et al. 2018) play essential roles in producing high-quality SLRs. For such reasons we believe that rationale for starting an SLR, the methodological quality and transparency of reporting are indeed three pivotal parts of a high-quality SLR.

Since most of previous studies rely on regression models (Judge et al. 2007; Soheili et al. 2022; Wagner et al. 2016, 2021; Xie, Gong, Li, et al. 2019) for measuring the effect of individual variables on citation impact, our work is inspired by Wang et al. (2011, 2012) who used machine learning tools for classification of low-, medium- and highly-cited papers. However, not to differ from the usual scientific practice, we used PLS-DA (Partial-Least Square Discriminant Analysis) as a way to combine both traditional and advanced statistics for delineating features affecting high-quality SLRs. We believe that the use of an advanced machine learning approach would not only highlight the novelty of the approach but would also position this study at the intersection of scientometrics and evidence-based practice. Finally, we extracted features to develop composite scores of RCE (Rationale-Cogency-Extent) criterion proposed in this study. However, our work did not stop here. After allocating most important features from high-quality SLRs using PLS-DA, we employed Network Analysis (NA) to deliver a regularised model allocating most impactful features within RCE network.

Taking altogether, the rationale for the study is identified from two aspects. From scientometric standpoint, the rationale is identified through (1) the lack of scientometric studies on the methodological-level features affecting citation impact in general; (2) the lack of scientometric studies investigating features affecting citation impact of SLRs; (3) the lack of understanding of critical features associating with high-quality SLRs. Facing the framework of existing problems in SLRs (see (Uttley et al. 2023)) this seem to be a compelling task considering that nowadays most SLRs are expected to follow imposed protocols (e.g., PRISMA). Although there are many studies assessing the relationship between bibliometric indices and citation count, there is a lack of empirical evidence considering methodological-level factors, which is where we believe this study resides.

The rest of the study is structured as follows. The second chapter explains the development of the manifest variables (items), their application on the two case studies of engineering-based SLRs and identification of factors that are used to assess the quality and citation impact using PLS-DA machine learning classification. The third chapter provides research results from the PLS-DA and NA analysis, including model's architecture, bootstrapping results and multiple comparison tests. The fourth chapter discusses obtained results, limitations, implications, and concluding remarks of the study.

2 Methodology

2.1 Coding of RCE criterion

Although extensive literature is available regarding Quality Assessment Tools (QATs) for reporting quality of primary studies, little efforts are made to develop quality assessment tools for SLR studies. Namely, in the medical sphere, the Joanna Briggs Institute (JBI) proposed a JBI-QAT 11-item checklist for assessing the research evidence in SLRs (Joanna Briggs Institute and JBI 2022). The CASP (Critical Appraisal Skills Programme) checklist (CASP 2018) for Systematic Reviews provides ten questions to evaluate the validity of results in SLRs. The ROBIS (Risk of Bias in Systematic Reviews) (Whiting et al. 2016) contains 24 signalling questions used for assessing the SLRs' bias and are split into two phases to evaluate study eligibility criteria, identification and selection of studies, data collection and study appraisal and synthesis of findings as the first phase, and risk of bias in the review (3 questions) as the second. Aside from those mentioned, two checklists stand out the most for guiding and assessing SLRs (D. Liu et al. 2015), PRISMA and AMSTAR-2 (Assessing the Methodological Quality of Systematic Reviews) (Shea et al. 2017), respectively. The AMSTAR-2 tool consists of 16 items with *yes*, *partial yes* and *no* responses to the items. However, even after the immensive amount of protocols have been published and become a normative in medical sciences a recent study showed that only 38% of published SLR of interventions reported a protocol in SLR (van der Braak et al. 2022). Faggion et al. (2017) used a sample of reviews (n = 275; 97 SLRs) from five highest ranked medical journals and showed that up to 37/97 of SLRs did not report full methodology (search+selection+extraction+quality assessment).

On the other hand, Kitchenham et al. (2007) report is most familiar among engineering-based SLR studies on their work on EBSE (Evidence Based in Software Engineering) tool. Their report provides instructions and guidelines for conducting an SLR and QATs for qualitative and quantitative studies. The research by Dybå and Dingsøyr (2008) and Zhou et al. (2015) extend the EBSE with additional insights concluding 4Rs (Relevance, Retrieval, Rigour, and Reporting) are the main predictors for successful SLR. In addition, a study by Templier and Paré (2018) of 142 review articles extracted six essential steps (problem formulation; literature search; screeing for inclusion; quality assessment; data extraction; data analysis and interpretation) from existing protocols. Their analysis emphasises that data extraction plan in reviews are rarely made explicit, while also showing the lack of quality assessment tools. Without exception all agree that the proposed and well-formulated RQ is arbitrary in starting a review – guiding the search strategy, extracting evidence and producing an outcome. However, although it is argued that RQ is a crucial, little attention is given to the informal logic, rationale, strength of the RQ, evidence and data behind the proposed RQ. Many authors report and discuss generated results without asserting and providing data behind the analysis, which additionally fuels the replication crisis that science encounters today. As a consequence, we identified three major factors *Rationale* – describing the arugments and logic for staring an SLR, *Cogency* – the methodological quality, search strategy and rigour for extracting evidence and *Extent* – the level of transparency in providing evidence and data of the study findings. These RCE factors are used to generate manifest variables. Additionally, since most existing items for guiding the review and QAT are assessed by "yes", "partial yes", or "no" outcome, the extension is performed with an ordinal scale for determining the level of items included.

The development of RCE items started with online brainstorming sessions since the development of items is conducted by the international team of experts. Firstly, the idea is to generate at least 5 items for each factor with 5-point ordinal scale. Secondly, each item should take into consideration four important aspects extracted from existing SLR issues that direct the development of items: (1) from general to specific; (2) from mild to rigorous; (3) from descriptive to critical; (4) from low to high impact. Each team member was asked to propose five items for each RCE criterion with 5-point ordinal scale considering proposed aspects. After two iterations the list comprised of 46 unique items. The list consisted mostly of items that are generically used within SLR protocols of experts involved in the development of the RCE. Through iterative discussions and consensus-building, the list was reduced to 22 items that are mostly associated with existing protocols (e.g., PRISMA), but with added additional ordinal scale, instead of tranditional binary (dummy) response to the item.

Again, after another two iterations and removing overlapping items, the final list included 17 items with 5-point ordinal scale that are used in this study. Two authors underwent training before data collection and data processing. For measuring intercoder reliability we used Cohen's Kappa coefficients. After the training, two authors separately evaluated the sample of n = 10 SLR articles and coded each paper according to RCE criterion. The measured agreement resulted in Cohen's K = 0.92. Considering individual RCE coefficients strongest variation was amongst Rationale items where coefficient ranged (K = 0.68-0.85) still indicating substantial agreement among coders (Xie, Gong, Li, et al. 2019). During the analysis of samples if there was a case of disagreement between two coders the third author stepped in to reach a consensus and make a final decision. If a disagreement could not be resolved the article is removed from the analysis. In order to validate the internal consistency of RCE scores later in the analysis, we used Cronbach's alpha for the reliability estimation of items. Finally, the complete list of coded RCE factors are given in following tables - Table 1 shows coded items for Rationale criterion; Table 2 shows coded items for Cogency criterion; Table 3 shows coded items for Extent criterion. In addition, based on literature review we provide a list of variables used in the analysis in the Table 4.

Table 1. Full list of coded Rationale items

	Informal Logic behind the research Question (ILQ)	Scale
Rationale	The study lacks arguments for justifying the need for starting a review study.	1
	Does not conclude cutting-edge research (e.g., outdated). Undermines the credibility of the RQ.	2
	Lacks criticality in challenging previous findings. Superficial understanding.	3
	Uses primary and secondary studies but fails to scope the setbacks of other systematic reviews.	4
	Comprehensive synthesis of the existing body of knowledge. Rigorous and insightful. Strong rationale.	5
	Motivation behind Research Question (MRQ)	**Scale**

The rationale is not convincing in explaining the RQ(s) significance.	1
Addresses review but not each specific RQ; needs more clarity and depth.	2
Relies on subjective reasoning; insufficient analysis and evidence from referenced studies.	3
Highlights gaps; comprehensive analysis; sound arguments; a clear understanding of research context.	4
Critically challenges previous evidence, understanding of research context, and compelling justification.	5
Research Question Formulation Logic (QFL)	**Scale**
Does not formulate explicit RQ.	1
Explain the study's intents and aims and implicitly state RQ (or the RQ is formulated later).	2
Explicit RQ, but lacks question formulation logic or framework (e.g., lacks dimensions of the question).	3
Uses RQ framework (e.g., PICO) but lacks an explanation of elements used in the RQ.	4
Explicit RQ framework (e.g., PICO). Explains each element of the RQ framework.	5
Research Question Strength (RQS)	**Scale**
Poorly defined, broad, or flawed. Low impact.	1
Clear but lacks specificity and originality.	2
Clear, specific, and relevant but may lack feasibility.	3
Clear, specific, relevant, feasible and based on sound methodology.	4
Clear, specific, relevant, feasible, narrow and precise. High impact.	5
Question' Evidence Aim (QEA)	**Scale**
The study does not explain what evidence will be critically appraised with proposed RQ.	1
The study uses outcomes, findings, or results for proposing the RQ.	2
The study uses descriptive evidence from retrieved studies for proposing the RQ.	3
The study critically appraises and synthesises evidence for allocating gaps for proposing the RQ.	4
The study uses and aims for evidence (e.g., meta-analysis) behind the proposed RQ.	5
Question' Data Aim (QDA)	**Scale**
The RQ does not address what data is used to support evidence from primary studies.	1
The RQ focuses on the meta-data as evidence from primary studies.	2
The RQ relies on qualitative data (e.g., methodology, findings, outcomes) as evidence.	3
The RQ uses specific data to compare and synthesise evidence across studies.	4
The RQ gathers and aims to critically inspect data from each included study.	5

Table 2. Cogency Criterion Items

	Search Strategy (SST)	**Scale**
Cogency	The search strategy is explained only narratively.	1
	The search strategy lacks replicable eligibility criteria for the retrieval of studies.	2
	The search strategy is explained and shows how articles (studies) are collected (e.g., PRISMA).	3
	Explains the search strategy through all information sources (index bases, search engines).	4
	Uses a standardised checklist and presents arguments for each step of the search strategy.	5
	Selection Process Reasoning (SPR)	**Scale**
	Does not provide clear reasoning for the inclusion of studies.	1
	Explains the study selection process (e.g., inclusion/exclusion criteria).	2
	Provides a detailed list of eligibility criteria (e.g., loosely related, partially related).	3
	Provides reasoning behind each eligibility criterion.	4
	Explains step by ellaborating and rationalising the study selection process.	5
	Information Literature Sources (ILS)	**Scale**
	Uses only a particular database/indexbase for the search (e.g., SCOPUS).	1
	Uses multiple indexbases/databases (e.g., WoS, SCOPUS, ScienceDirect, IEEE).	2
	Extends previous databases/index bases with search engines (e.g., Scholar, BASE).	3
	Extends with publishers and libraries, snowballing and contacting authors, organisations, etc.	4
	Extends by contacting authors, organisations, and registries (e.g., Cochrane, PROSPERO).	5
	Type of Publications (TPUB)	**Scale**

Uses only peer-review articles from journal articles.	1
Uses peer-review journal articles and extends them with conference papers and book chapters.	2
Extends previous items with reports and theses.	3
Extends with grey literature (preprints, magazines, non-peer review studies).	4
Extends with datasets, patents, standards, project reports, technical reports, etc.	5
Quality Assessment Tools (QAT)	**Scale**
Does not use quality assessment tool for retrieved primary studies.	1
Uses subjective quality assessment in the review.	2
Formulates quality assessment strategy for estimating the quality of retrieved studies.	3
Uses specific QAT for estimating the quality and reporting of retrieved studies.	4
Uses a detailed QAT checklist for assessing each retrieved study's reporting quality.	5
Data Quality Assessment (DQA)	**Scale**
Does not use data quality assessment within obtained primary source articles.	1
Implicitly address some data quality dimensions (e.g., discusses missing data).	2
Explicitly addresses specific data quality dimensions (e.g., accuracy of data, bias in data, timeliness).	3
Explicitly assess data quality within obtained articles that can cause bias in final results.	4
Explicitly assess and critically appraises five main data quality dimensions (e.g., USAID DQA).	5

Table 3. Extent criterion items

	Selection Process Transparency (SPT)	**Scale**
Extent	The search strategy states the selection process narratively (e.g., lacks tables; flow diagrams).	1
	The selection process is stated narratively but provides inclusion and exclusion criteria.	2
	The selection process is depicted through tables and diagrams replicating the results.	3
	Provides exact search strings, isolation criteria, and results that can be replicated.	4
	Provides transparent results in the attachment or as supplementary material.	5
	Transparency of Articles included (ToA)	**Scale**
	It does not show explicitly what studies are used for systematic review.	1
	Studies are depicted in the SLR by references or tabular; however, still need full representation.	2
	Shows included studies in the review through tables or in the attachment (e.g., author, title, journal).	3
	Explicitly describes primary articles' main (meta-records) through tables.	4
	Provides a complete list of meta-records and processed data in the attachment or supplementary material.	5
	Transparency of Evidence (ToE)	**Scale**
	Either omitted or vague association of evidence without referencing (e.g., replication fail).	1
	Provides highlights about the study (e.g., outcomes, findings, results) as a response to the RQ(s).	2
	Provides parts of included studies' evidence (e.g., methods, variables, samples) to respond to the RQ(s).	3
	Control measures (e.g., metrics, cost functions, error estimates) are included in response to the RQ(s).	4
	Explicitly describes evidence (e.g., in the attachment or supplementary) from included articles.	5
	Transparency of Data (ToD)	**Scale**
	The review does not show transparent data obtained from processed studies.	1
	The review provides only the meta-data of studies included in the review.	2
	The review provides data across included studies.	3
	Explicitly describes obtained data from primary studies (e.g., tables, spreadsheets).	4
	Extends previous by including additional data insights (e.g., descriptives, performance metrics).	5
	Quality Assessment Tool Results (QATR)	**Scale**
	Does not provide transparently results of quality assessment.	1
	Implicitly states or gives partial results of quality assessment.	2
	Overall results of quality assessment of included studies are provided.	3
	The results for every study in the review are provided.	4
	Extends previous by providing interrater agreement and scores for every item of used QAT.	5

Table 4. List of features (factors) used for the analysis

Category	Subcategory	Coding	Ref
Author(s)	Authors characteristics	1 = h index 0-10 2 = h index 11-20 3 = h index 21-30 4 = h index 31-40 5 = h index >41 Reworked to include raw h index from authors.	(Grover et al. 2014; Judge et al. 2007; Xie, Gong, Li, et al. 2019)
	Author characteristc	Max h index	(Vanclay 2013; Xie, Gong, Li, et al. 2019)
	Authors characteristic	Average *h* index per author	(Xie, Gong, Li, et al. 2019)
	Affiliation	1 = Rank 1-200 2 = Rank 210-400 3 = Rank 401-600 4 = Rank 601-800 5 = Rank >800 Reworked ranking based on Shanghai's list of University rankings.	Improved from: (Grover et al. 2014; Judge et al. 2007)
	Number of authors	1 = one author 2 = two authors 3 = three authors 4 = four authors 5 = five or more authors	(Grover et al. 2014; Vanclay 2013)
	Research discipline	Not coded. (engineer-indexed SCOPUS SLRs)	-
	Type of Issue	1 = Regular Issue 0 = Special Issue	(Judge et al. 2007)
Title	Length words	Number of words	-
	Length characters spaces	Characters with spaces	-
	Length characters without	Characters without spaces	-
Abstract	Length	Number of words. (Not used in this study).	(Xie, Gong, Li, et al. 2019)
	Language	Only English included.	(Kousha and Thelwall 2023)
	Keywords	Number of keywords. (Not used in this study).	(Xie, Gong, Li, et al. 2019)
Manuscript	Search coverage	"*Manual analysis of whether the authors document and report the search coverage*". Reworked in this study to RCE.	(Vanclay 2013; Wagner et al. 2016)
	Synthesizing, Identifying Resarch Gaps or Developing Research Agenda	"*Association for Information Systems and identify research gaps or develop a research agenda*". Reworked in this study to RCE.	(Wagner et al. 2016)

2.2 Study samples

The gathering of enginering-type SLR studies was performed in two runs. The first sample search was performed on the 12th of October 2022. The keywords set for the search on SCOPUS are "*systematic literature review*" and are limited to the ABS-TITL-KEY (Abstract-Title-Keywords). The SCOPUS search string is given as: *TITLE-ABS-KEY ("systematic literature review") AND (LIMIT-TO (SUBJAREA, "ENGI")) AND (LIMIT-TO(PUBYEAR, 2020)*. The obtained result was 884 studies. Aligning with earlier research on the impact of year-related bias and to increase the homogenity of data sample (Kousha and Thelwall 2023) we used single year publications. The sorting was performed based on citation ("Cited by (highest)") to isolate 240 papers in interdisciplinary engineering studies (Top 120 and Bottom 120). The selection of papers is perfomed based on the IES (Isolation-Exclusion-Selection) criteria (Table 5). Based on the IES criteria 70 papers were removed from the study. The final corpus included 170 SLR articles (103 journal articles, 65 conference papers, and two book chapters).

Table 5. IES criteria for obtaining studies.

I/E/S Criteria	Sub-criteria	Criteria Explanation
Isolation	Full-text papers	Papers are complete (e.g., not editorials, short communications).
	Language	No language restriction.
	Time frame	Papers that were published in 2020.
Exclusion	Engineering studies	Papers are not suitable and are outside of the engineering sphere.
	Only Top/Bot 120 papers.	Top 120 and Bot 120 articles based on the citation count.
	Not SLR	Papers that still appear in the search but are not SLR studies.
	Access Policy	Papers not accessible (e.g., green/open access) by the institution.
	Reviewer decision	Papers are misclassified as engineering (e.g., medical, genetics).
Selection	Closely-Related	SLRs that are suitable for review by the expert panel.

Under the assumption that book chapters and conferences, due to lower visibility and citation metrics (e.g., JIF), may cause bias and heterogeneity (Wagner et al. 2021), another sample is used consisting only of the "Big 3" journals who published most engineering-based SLRs in SCOPUS (MDPI Sustainability, IEEE Access, and Elseviers' Journal of Cleaner Production). The search string for the second search on SCOPUS is given as: *TITLE-ABS-KEY ("systematic literature review") AND (LIMIT-TO (PUBYEAR, 2021)) AND (LIMIT-TO (SUBJAREA, "ENGI")) AND (LIMIT-TO (EXACTSRCTITLE, "IEEE Access") OR LIMIT-TO (EXACTSRCTITLE, "Sustainability Switzerland") OR LIMIT-TO (EXACTSRCTITLE, "Journal Of Cleaner Production"))*. Based on the eligibility criteria (Table 6), 227 out of 255 SLR articles were included – Sustainability 78, IEEE Access 81 and Journal of Cleaner Production 68 studies. Finally, the complete list of both samples of SLRs are given in the supplementary files, including both included and excluded SLRs.

Table 6. IES criteria for obtaining studies.

I/E/S Criteria	Sub-criteria	Criteria Explanation
Isolation	Full-text papers	Papers are complete (e.g., not editorials, short communications).
Exclusion	Language	No language restriction.
	Time frame	Papers that were published in 2021.
	Engineering studies	Papers are not suitable and are outside of the engineering sphere.
	Big3-Journal Articles	The SLR studies were not published in „Big 3".
	Not SLR	Papers that still appear in the search but are not SLR studies.
	Access Policy	Papers not accessible (e.g., green/open access) by the institution.
	Reviewer decision	Papers are misclassified as engineering (e.g., medical, genetics).
Selection	Closely-Related	SLRs that are suitable for review by the expert panel.

2.3 Partial Least Square Discriminant Analysis (PLS-DA)

The PLS-DA is extremely popular in classification (Perk et al. 2011) and feature selection (Yan et al. 2017) in various applications. The PLS-DA takes the relation of selected variables and constructs a new set of features in respect to the projection (loading) into lower dimensional space vectors, which are often noted as LV (Latent Variables) or PC (Principal Components). Unlike unconstrained PCA (Principal Component Analysis) technique, which constructs a set of features by means of linear transformation, which best explains the variance within used data, the PLS-DA is constrained, meaning that it projects LVs (PCs) with repect to the class label Y (Orošnjak et al. 2023), which in this case are classification labels – "top-ranked" and "bottom-ranked" SLRs. However, the model is often misused since it is prone to overfitting (Ruiz-Perez et al. 2020); thus, cross-validation is needed to avoid misinterpretation (Westerhuis et al. 2008).

The overfitting is a significant challenge for evaluating proposed RCE in SLRs, due to nuanced nature of citation context. PLS-DA's constraint in projecting LVs in respect to class labels, coupled with ten-fold cross-validation, will ensure that proposed model specifically addresses the concern and ensure robustness. Therefore, the study performs ten-fold cross-validation in order to avoid bias, i.e., overfitting. Detailed analytical description of PLS-DA is provided in Appendix 3. Finally, we peform permutation test, a critical part of our the analysis, to validate and capture the genuine structure of RCE in the context of SLRs.

Considering the nature of scientometric data structures, which usually involves complex and multidimensional datasets, such as meta-, methodological- and content-based data, PLS-DA is particularly adept to handling complex structures. Its ability to construct new sets of features, i.e., LVs (PCs), in relation to the proposed top- and bottom-ranked labels makes it an ideal tool for the analysis. Next, the specific choice of PLS-DA is also in its ability to handle multicollinearity, high-dimensionality and a small sample size relative to predictors. Also, the use of VIP (Variance Importance in Projection) score of PLS-DA aligns well with the need for precise feature selection, which in this case helps identify and place emphasis on those features that are most impactful between the top-ranked and bottom-ranked SLRs, ultimately making it a crucial aspect of our study.

2.4 Network analysis

Lately, many researchers started utilising Network Analysis (NA) over traditional traditional Factor Analysis (FA). There are several reasons for such decision. Firstly, the NA enables visual representation of complex non-linear relationship among variables, thereby providing insights that traditional FA may not adequately capture. This visualisation also helps in identifying and avoiding spurious correlations, which enhances the robustness of the analysis. Next, the NA can identify clusters or communities of closely related nodes (variables) within the network, ultimately understanding the modular structure of the data. Also, centrality measures helps NA identify the most influental nodes (variables) providing an aditional layer of analysis that FA may not offer. Finally, NA can handle multidimensional data, including ordinal, nominal and scale data, providing greater flexibility in analysis. The NA model, consisting of composed items is constructed to assess specific underlying dimensions (Briganti et al. 2019). As such, in social and health sciences scales are constructed with several similar items that are used to measure a specific construct. This is a challenge for network models because the meaning of connections between nodes (items) changes. Usually, Gaussian graphical models are used to build the network, where nodes represent variables and edges represent conditional (in)dependence depicted by statistical estimation between variables, such as partial correlation coefficients (after conditioning on all other nodes). Simply put, an association between items reflect the shared variance and a common cause is plausible.

The Fruchterman-Reingold (Fruchterman and Reingold 1991) algorithm is used for positioning the nodes in the NA graph. The FR algorithm can be understood as a force-directed measure that is visually represented by nodes (vertices) and edges (lines) as strings for the visual representation of data using specific estimators (Leme et al. 2020). The EBICglasso (Extended Bayesian Information Criterion graphical least absolute shrinkage and selection operator) estimator is used for the network type. The specific choice of the EBICglasso network (Gamma parameter 0.5 (Foygel and Drton 2010)) is suitable for both continuous and ordinal data. It is most prevalent in today's scientific practice because it reduces the complexity of connections for explaining the covariance between variables (Leme et al. 2020). The network interpretation is performed based on the global centrality measures represented by the centrality plot (JASP 2018). The centrality indices include *Betweenness* (the number of times a node lies within the shortest path between other nodes (Kalantari et al. 2022)), *Closeness* (the average shortest path between a single node with other nodes in the network), and *Strength* (absolute values of connections between different nodes (Robinaugh et al. 2016)). Since EBICglasso model requires an estimate of the variance-covariance matrix that then returns parsimonious network, the computation of data is conducted by polychoric correlation because items consist of ordinal data. Using an R environment this can be conducted from the *qgraph* by *cor.auto* fuction that returns polychoric correlation matrix. Additionally, instead of using R the results can be replicated in open-source software JASP since it is built on R and includes mentioned packages. The whole dataset and parameters are given in supplementary csv files to assure reliability of results. Inspired by previous similar work on scientometrics using parsimonious models with the sphere of factors of scientific impact of reviews (Wagner et al. 2021), we believe that results will resonante with insightful findings.

Finally, to assure validity of obtained results, we used bootstrapping for estimating edge and node sensitivity, i.e., network stability. The network stability estimates edges between nodes in terms of strength. For estimating the confidence intervals, 1000 bootstrapps were performed. For the network stability we used stability coefficients by measuring maximum proportion of cases that can be dropped (%sample) in order to maintain 0.7 correlation with the original sample. Since we are using both in-network and out-network analysis for measuring items influencing the performance of the network and citation impact, respectively, we additionally performed multiple comparison testing with Bonferroni correction outside of the PLS-DA analysis. For the testing between top- and bottom-ranked SLRs we used item and composite score of RCE criterions (factors) in respect to the citation impact using two samples, $n_1 = 170$ and $n_2 = 227$ studies.

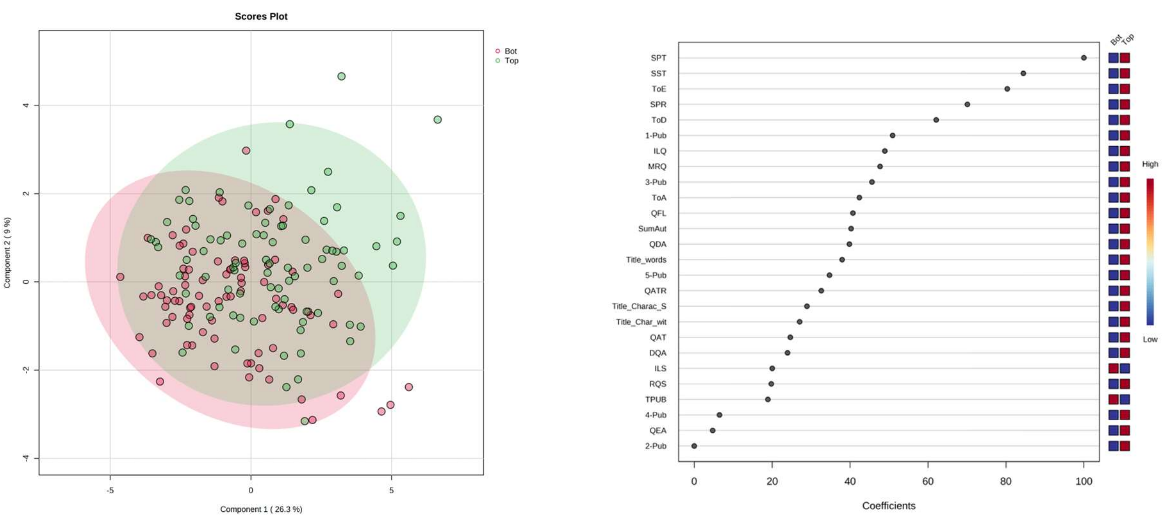
3 Research results

3.1 Analysis of the first sample of SLRs

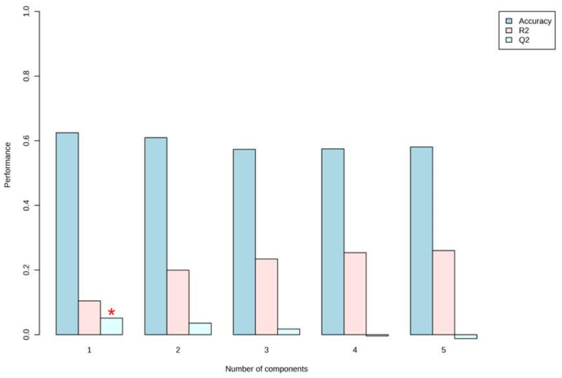
From descriptive statistics (Appendix 2, Table A1) data shows that items mostly vary around the sample mean with 2.38 ± 0.32 score across all items. Considering the score of QFL of 1.94±0.31 most authors fail to provide an explicit RQ, but instead start the SLR process without RQ or implicit statements (e.g., aims of the study). Looking at $QFL_{st.dev}$ = 1.02 (CI [0.86, 1.15]) and with Kurtosis = 2.04 and Skewness = 1.44, suggests bias in taildness and symmetry of data, respectively. Additionally, the cases of DQA and QATR items suggest bias in estimation since most of the respondents vary around score of 1. Namely, median of DQA and QATR is 1, suggesting that most of the responses show a lack of data quality assessment and transparency of evidence in retrieved studies. Comparing descriptive results between classes, data shows significant deviations between the two, for instance JIF_{Top} = 8.255 and JIF_{Bot} = 1.013; 1-Pub_{Top} = 30.085 and 1-Pub_{Bot} = 16.788; 2-Pub_{Top} = 59.273 and 2-Pub_{Bot} = 55.580; ILQ_{Top} = 3.787 and ILQ_{Bot} = 2.906; MRQ_{Top} = 3.170 and MRQ_{Bot} = 2.447; and other features indicate higher performance considering top-ranked SLRs.

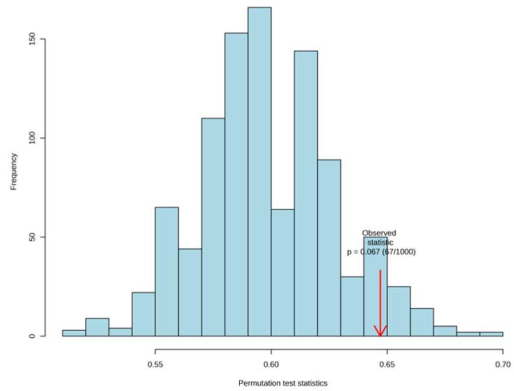
Investigating the effects of bibliometric features the evidence suggest high association with CiteScore (0.718; $p < 0.01$) and JIF (0.730; $p < 0.01$). Considering the variables of title characters with or without spaces and title words there is no statistically significant difference. Performing both Pearson's and Spearman's correlation after log normalisation of citation score due to high skewness we obtain following results. In both cases the analysis show that all except QFL, ILS and TPUB items report statistically significant results. Using partial correlation by controlling the effect of impact factor the statistically significant results include SST, SPR, SPT and ToE, suggesting the selected sample is heavily biased due to the presence of the effect of high quality journal articles versus book chapters and conference proceeding papers.

After conducting PLS-DA the results show that even with 38% of the explained variance of the first two components (Figure 1a) the classification accuracy is 80%, and after including additional component the classification was 90% (Figure 1c). The results after 1000 permutation test and 10-fold cross validation show relatively strong observed statistic $p < 0.001$ (Figure 1d). Finally, the features contributing the most to the separation of classess include SPT; SST; ToE; SPR; ToD; 1-Pub, ILQ, MRQ and 3-Pub (Figure 1b), respectively with threshold of 40.0 of VIP coefficient score.



(a) (b)





(c) (d)

Figure 1. Results of the PLS-DA on the first sample of SLRs using 10-fold cross validation: (a) scree plot of first two components; (b) VIP scores of features; (c) classification accuracy; (d) results of 1000 permutation test.

The results of PLS-DA suggest heavily unreliable and inconsistent results. Namely, negative Q^2 score suggest that model can be either overfitted or not predictive at all, which is all observed by the stability permutation test. Acknowledging the setback, we excluded the first sample from further analysis and strictly target journals with the most amount of engineering-based SLRs published indexed by SCOPUS – Elsevier Journal of Cleaner Production, MDPI Sustainability and IEEE Access.

3.2 Analysis of the "Big 3" journals

3.2.1 Descriptive statistics and analysis of control variables

From the descriptive statistics the results show the following. The average citation per paper is 18.52 (95%CI [15.599, 21.441]), with maximum of 155. The JIF (i.e., WoS-IF) and CiteScore were considered significant in this case and used to control on other variables. The University ranking according to Shanghai list shows mean of 3.485 (95% CI [3.264, 3.705]) and show no correlation with PCN (Figure A4) nor log transformed PCN (Figure A5) which was reported in previous studies. The average number of authors per paper 3.58 (95% CI [3.432, 3.731]), which does not show significant correlation with citations. Considering title charactersitics, i.e., number of letters with spaces, without spaces, and title words, there was no statistically significant results. The number of publications, all but except first author showed no statistically significant correlation with citation impact.

Table 7. Descriptive statistics of the Big 3 journals

	n	Mean	95%Upper	95%Bottom	St.dev	Skew	Kurt	Min	Max
PCN	227	18.520	21.441	15.599	22.333	2.575	8.727	0.000	155.000
CiteScore	227	8.842	9.446	8.238	4.616	0.803	-1.258	5.000	15.800
WoS-IF	227	5.895	6.339	5.450	3.396	0.873	-1.237	3.480	11.070
University_Rank	227	3.485	3.705	3.264	1.686	-0.339	-1.683	1.000	5.000
1st_H-index	227	7.211	8.459	5.963	9.542	5.151	43.468	1.000	102.000
2nd_H-index	222	13.775	15.390	12.159	12.213	2.030	5.615	1.000	73.000
3rd_H-index	181	15.088	17.004	13.173	13.059	2.823	15.705	0.000	111.000
4th_H-index	120	12.283	14.360	10.207	11.489	1.820	4.503	1.000	66.000
5th_H-index	64	16.734	20.244	13.224	14.051	1.641	3.705	1.000	73.000
Average_H	227	11.817	12.755	10.878	7.176	1.533	3.584	1.000	46.670
Max_H-index	227	22.198	24.265	20.132	15.799	2.107	7.197	1.000	111.000
1-Pub	227	19.876	24.667	15.086	36.631	5.038	36.491	0.940	360.000
2-Pub	220	54.677	64.453	44.901	73.572	4.102	26.399	1.000	684.000
3-Pub	181	69.188	83.226	55.149	95.716	3.728	21.033	1.000	801.000
4-Pub	122	54.189	68.823	39.554	81.648	2.904	9.191	1.000	456.000

5-Pub	63	85.159	111.342	58.975	103.965	1.778	2.681	1.000	440.000
Title_Charac_Space	227	98.449	102.020	94.879	27.300	0.521	-0.175	44.000	181.000
Title_Charac_without_Space	227	86.595	89.707	83.482	23.799	0.516	-0.159	38.000	159.000
Title_words	227	12.855	13.353	12.357	3.807	0.535	-0.214	6.000	23.000
SumAut	227	3.581	3.731	3.432	1.147	-0.264	-0.990	1.000	5.000

Considering the presence of the statistically significant effect we report Pearsons' correlation score, *p* value and Fisher's *z* effect size. Considering the Average_H index of authors there is a significant correlation ($r = 0.294$; $p < 0.001$; $z = 0.303$) with citation impact (Figure A5). The Max_H-index result report the presence of correlation ($r = 0.263$; $p < 0.001$; $z = 0.269$) with citation, in addition to the effect with 1st_H-index ($r = 0.295$; $p < 0.001$; $z = 0.304$) and 3rd_H-index ($r = 0.228$; $p < 0.002$; $z = 0.232$) of authors. The data behind statistical analysis shows that third author' H index is the most influential with mean of 15.088 (95% CI [13.173, 17.004]) whereas H index of first author report mean of 7.211 (95% CI [5.693, 8.459]). Finally, the number of publication of the first author shows the presence of correlation ($r = 0.181$; $p < 0.006$; $z = 0.183$). The results show consistence with previous studies on the impact of author-level indicators and citation impact (Grover et al. 2014; Judge et al. 2007; Vanclay 2013; Xie, Gong, Li, et al. 2019).

3.2.2 Descriptives and partial correlation analysis of RCE items

A descriptive analysis RCE scores (Table 8) are in the following. Observing the overall mean scores of Rationale = 2.44, Cogency = 2.32, and Extent = 2.41, the evidence suggests slight increase in comparison to the first sample with mean average increase of 5-7% in respect to the first sample of SLRs. The mean scores of ILQ = 3.01±0.35 and SST = 3.12±0.24 suggest higher quality in interpreting the rationale and increased rigor in methodological assessment, respectively. Also, the SPR = 2.92±0.29 suggests that studies provide more insightful description and rationale behind selection of studies. The DQA and QATR items suggest bias in estimation since most of the respondents vary between score 1.00 and 2.00. Namely, the median of DQA and QATR is one, suggesting that most SLRs lack data quality assessment and transparency in reporting. Meanwhile, QATR item shows an increase in Skewness = 2.35 and Kurtosis = 4.17, indicating a slight deviation in terms of asymmetry, whereas 183/227 SLRs scored one. Although it is a usual practice in medical and health sciences to present the results of QAT, within engineering-based SLRs such practice is still not fully adopted.

Table 8. Descriptive statistics of RCE items of the "Big 3" SLRs (n = 227).

	Mean	Std. Dev.	$95\%CI_{Upper\text{-}Mean}$	$95\%CI_{Lower\text{-}Mean}$	Std. Error	$95\%CI_{Upper\text{-}Stdev}$	$95\%CI_{Lower\text{-}Stdev}$
ILQ	3.009	1.350	3.184	2.833	0.090	1.425	1.266
MRQ	2.211	0.959	2.336	2.087	0.064	1.060	0.848
QFL	2.762	0.905	2.880	2.644	0.060	1.001	0.798
RQS	2.586	0.976	2.713	2.459	0.065	1.053	0.888
QEA	2.608	0.973	2.735	2.481	0.065	1.042	0.903
QDA	2.401	1.044	2.537	2.265	0.069	1.130	0.958
SST	3.119	0.940	3.241	2.997	0.062	1.019	0.864
SPR	2.925	1.097	3.068	2.782	0.073	1.173	1.011
ILS	2.339	0.849	2.450	2.229	0.056	0.904	0.785
TPUB	1.960	0.975	2.087	1.834	0.065	1.058	0.880
QAT	1.925	1.140	2.073	1.777	0.076	1.240	1.016
DQA	1.648	0.877	1.762	1.533	0.058	0.980	0.773
SPT	2.749	0.984	2.877	2.621	0.065	1.070	0.895
ToA	2.885	1.253	3.049	2.722	0.083	1.325	1.176

ToE	2.621	1.100	2.764	2.478	0.073	1.197	0.992
ToD	2.295	1.131	2.442	2.148	0.075	1.228	1.011
QATR	1.485	1.142	1.633	1.336	0.076	1.307	0.937

Based on the obtained results from Pearsons (Figure A6) and Spearman (Figure A7) correlation analysis of RCE items and citation impact, the evidence suggests the following. The list of statistically significant association with citation impact include: ILQ (r = 0.535; p < 0.001; z = 0.598); MRQ (r = 0.141; p < 0.017; z = 0.142); RQS (r = 0.168; p < 0.005; z = 0.170); QEA (r = 0.225; p < 0.001; z = 0.229); QDA (r = 0.198; p < 0.001; z = 0.200); SST (r = 0.118; p < 0.038; z = 0.119); DQA (r = 0.241; p < 0.001; z = 0.246); ToA (r = 0.241; p < 0.001; z = 0.246); ToE (r = 0.157; p < 0.009; z = 0.159); ToD (r = 0.199; p < 0.001; z = 0.201). The results are consistent with Spearman correlation results (Figure A5). However, considering the risk of confounding effect, we used control variables to partial out the effect, the results of Partial Pearson's correlation (Figure A8) show that only ILQ (r = 0.343; p < 0.001; z = 0.358) showed moderate effect (z > 0.3), while other variables MRQ (r = 0.134; p < 0.039; z = 0.134); QEA (r = 0.173; p < 0.011; z = 0.174); QDA (r = 0.171; p < 0.012; z = 0.173); DQA (r = 0.269; p < 0.001; z = 0.276); SPT (r = 0.136; p < 0.037; z = 0.137); ToD (r = 0.166; p < 0.014; z = 0.168), maintained low effect. The results of Partial Spearman's correlation (Figure A9) show consistent and slightly inflated results, where ILQ (r = 0.405; p < 0.001; z = 0.429) showed moderate effect, followed by MRQ (r = 0.172; p < 0.011; z = 0.174); QEA (r = 0.248; p < 0.001; z = 0.253); QDA (r = 0.247; p < 0.001; z = 0.252); DQA (r = 0.251; p < 0.001; z = 0.257); ToD (r = 0.212; p < 0.002; z = 0.215), with low effect. In the following, we continue to the PLS-DA analysis of the "Big 3" for allocating VIP features of top-cited SLRs.

3.2.3 PLS-DA Analysis of the "Big 3"

From the PLS-DA analysis (Figure 2), the results of the "Big 3" SLRs suggest the following. The first two components of the PLS-DA explain 26.2% of the variance (Figure 2a). The results of VIP score (Figure 2b) show that by far ILQ is the most important feature with VIP_{ILQ} = 3.37; followed by VIP_{QEA} = 1.52, VIP_{ToD} = 1.48, VIP_{ToE} = 1.47, VIP_{DQA} = 1.41, VIP_{QDA} = 1.35, $VIP_{1st\text{-}H\text{-}index}$ = 1.24, VIP_{ToA} = 1.24, $VIP_{Max\_H\text{-}index}$ = 1.06, $VIP_{3rd\_H\text{-}index}$ = 1.05, and others below VIP < 1.0. It should be noted that VIP scores are displayed for the first component, while the full list of scores is given in Appendix 2. Classification results show 72.33% accuracy (R^2 = 0.3594; Q^2 = 0.2448) just from the two components (Figure 2c). Finally, after conducting permutation test (n = 1000 permutations) and fitting the model (Figure 2d), the permutation results of show reliable predictions varying around 70% accuracy even after conducting 1000 permutations, suggesting that the model predictive capabilities are not merely a result of a random chance.

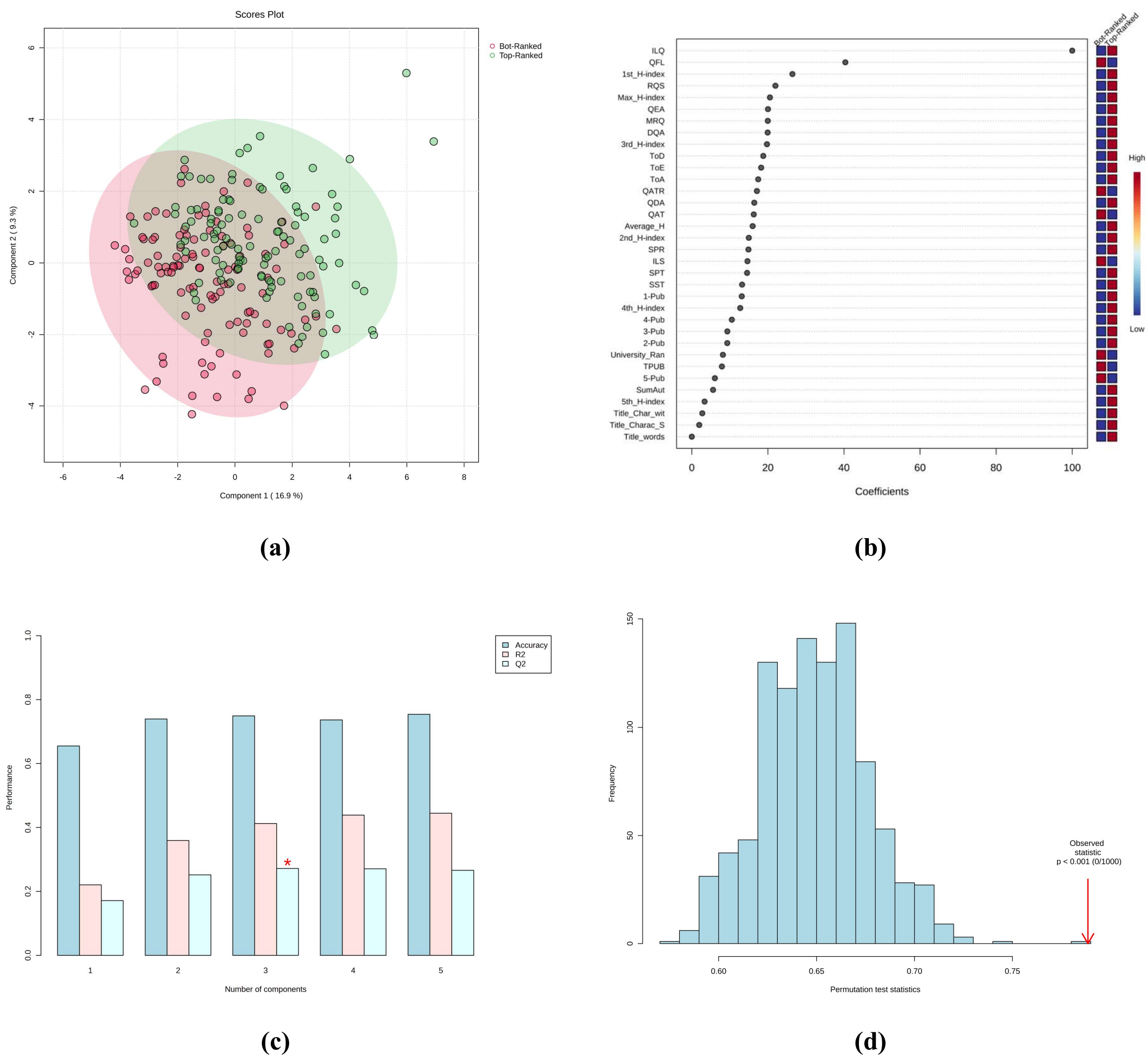


Figure 2. PLS-DA results of the Big 3 journals: (a) scree plot of first two components; (b) VIP coefficient scores of selected features; (c) classification accuracy score; (d) model validation by permutation test based on prediction accuracy ($p < 0.001$ (0/1000)).

3.2.4 Testing the significance of RCE between top- and bottom-ranked SLRs

In order to validate the internal consistency of composite RCE scores we performed Cronbach's alpha (Table 9). The results show an average score of Rationale = 0.88 (CI [0.86, 0.91]), Cogency = 0.82 (CI [0.77, 0.86]) and Extent = 0.85 (CI [0.81, 0.88]) criterion items used. The results show that items are statistically related.

Table 9. Cronbach's alpha of RCE items

Estimate	Rationale	Cogency	Extent
Point estimate	0.886	0.824	0.853
95% CI lower bound	0.857	0.773	0.814
95% CI upper bound	0.910	0.864	0.884

For checking the assumptions needed for conducting test statistic of RCE, we depict RCE composite scores in Figure 3 with boxplots and Q-Q plots. Visual inspections suggest the existence of normality between the split. Next, assumptions check using Brown-Forsythe and Levene's test show that the p-value < 0.05 is not violated. Although outliers can be excluded from the analysis; however, the Shapiro-Wilk test shows that cases of Cogency ($W = 0.969$, $p < 0.01$) and Extent ($W = 0.969$, $p < 0.01$) violate the assumption of normality.

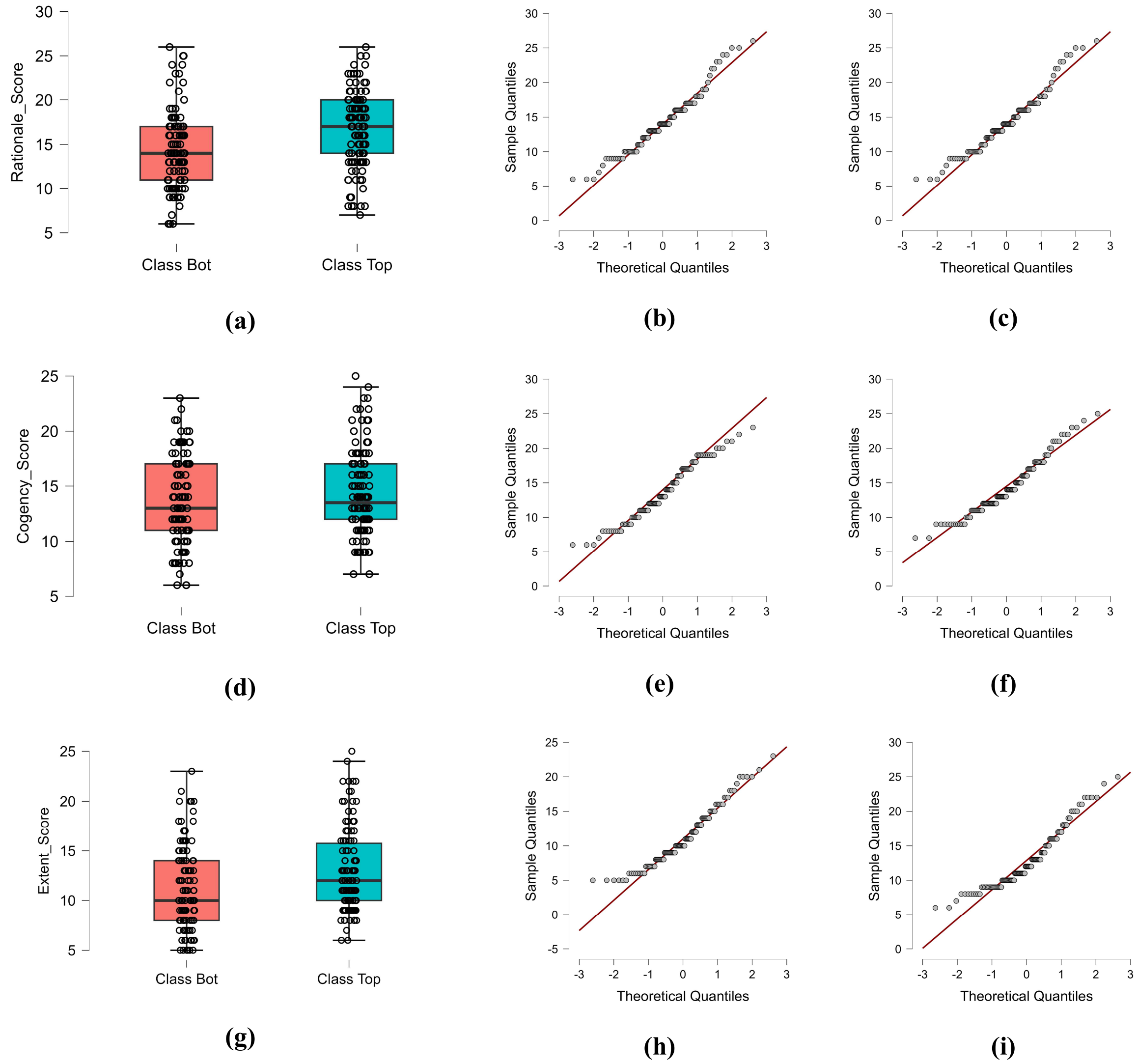


Figure 3. RCE boxplots of Rationale split (a), Cogency split (d), Extent split (g) and Q-Q plots of Rationale bot (b) and top (c) SLRs, Cogency Q-Q plots of the bot (e) and top (f) SLRs, and Extent Q-Q plots bot (h) and top (i) SLRs.

Given that the obtained sample size violates the normality assumption, the Mann-Whitney *U* test is used. Next, since we are dealing with multiple comparison problem, the Bonferroni correction is again used to control the error rate, thus reducing the inflation of results. The obtained results are given in Table 10. Considering Rationale, the results show that all except QFL are statistically significant (3/6 items after Bonferroni), with the most significant effect size in ILQ. Considering Cogency items, only DQA show significance, while composite scores and items do not show the presence of difference. Finally, the Extent score suggests significant difference, whereas items ToE and ToD show significant median difference between top- and bot-ranked SLRs.

Finally, the VS-MPR depicts maximum possible odds in favour of alternative hypotheses considering Rationale score with RBC 0.359 (CI [0.486, 0.245]), while ILQ (RBC = 0.616 with CI [0.701, 0.532]) and ToD (RBC = 0.303 with CI [0.433, 0.185]), in favour of top-ranked SLRs. Comparing the score to the first case study, the evidence suggests a difference between QFL, SPR, and DQA. However, comparing the case studies' split, the evidence suggests that ILQ and QEA (consequently the Rationale) suggest substantial differences between samples. The VS-MPR and RBC effect size also support this conclusion. Thus, it can be seen that the overall score of top-cited published SLRs provides a more compelling narrative, i.e., informal logic for starting the review, while maintaining the specificity and providing compelling evidence before proposing the RQ.

Table 10. Statistical testing results of the "Big 3"

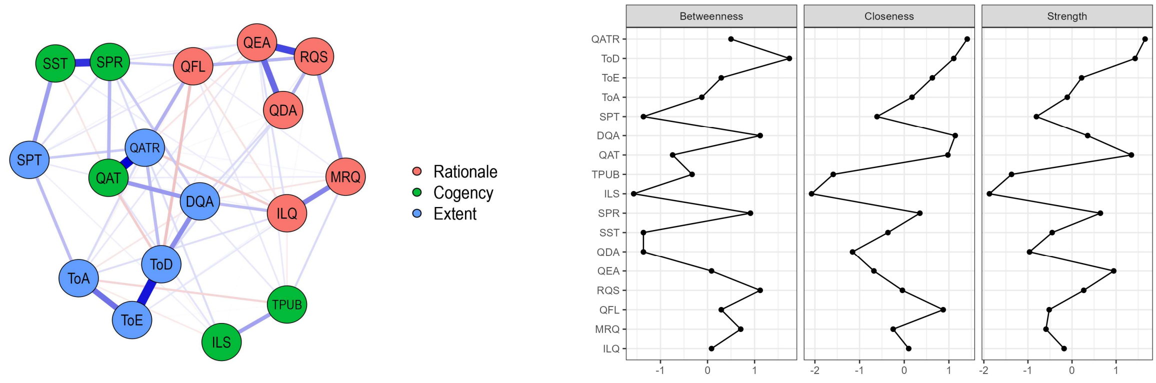
Item	U	p	VS-MPR[a]	RBC[b]	95% CI RBC_{Lower}	95% CI RBC_{Upper}
ILQ	2468.5	< 0.001	8.19E+13*	-0.616	-0.701	-0.532
MRQ	5443.5	0.014	6.041	-0.154	-0.296	-0.028
QFL	7359.5	0.992	1.000	0.144	-0.005	0.265
RQS	5302	0.008	9.477	-0.176	-0.317	-0.051
QEA	4644.5	< 0.001	524.834*	-0.278	-0.410	-0.158
QDA	4688.5	< 0.001	335.915*	-0.271	-0.404	-0.150
Rationale Score	4119.5	< 0.001	19611.461*	-0.359	-0.486	-0.245
SST	5710.5	0.061	2.146	-0.112	-0.257	0.014
SPR	5919	0.141	1.332	-0.08	-0.226	0.047
ILS	6986	0.884	1.000	0.086	-0.064	0.210
TPUB	6557.5	0.607	1.000	0.02	-0.130	0.145
QAT	6221.5	0.324	1.007	-0.033	-0.181	0.094
DQA	4705.5	< 0.001	799.132*	-0.268	-0.402	-0.148
Cogency Score	5899	0.14	1.335	-0.083	-0.229	0.043
SPT	5592.5	0.037	3.024	-0.13	-0.275	-0.005
ToA	5127	0.003	21.022	-0.203	-0.342	-0.079
ToE	4892.5	< 0.001	92.325*	-0.239	-0.375	-0.117
ToD	4480.5	< 0.001	2126.1*	-0.303	-0.433	-0.185
QATR	6484.5	0.563	1.000	0.008	-0.141	0.134
Extent Score	4960	0.001	39.555*	-0.229	-0.366	-0.106

[a]Vovk-Sellke Maximum *p*-Ratio is given as a two-sided *p*-value explaining the maximum odds favor the H_1 over H_0.
[b]RBC = Rank-Biserial-Correlation for measuring the effect size of Mann-Whitney U test.
*The statistically singificant value according to the Bonferroni correction where adjustmed *p* value is < 0.0025 ($\alpha < 0.05$; $m = 20$).

Finally, the results obtained are used to complement discussion of RCE scores after performing network analysis in the following subsection. The reason for conducting network analysis is to establish the robustness of the analysis considering the relation between nodes (features) and edges (partial correlations). This way we will provide visual representation of the behavior of SLRs considering the context of the analysis.

3.2.5 Network Analysis

The performed NA of association between nodes and edges is provided in Figure 4a, while the centrality plot of indices is given in Figure 4b. The network consists of 73 non-zero out of 136 possible edges resulting in 0.463 network sparsity. The association between of TPUB and ILS nodes suggests that these nodes are statistically independent when conditioning on all other variables (partial correlation is close to zero). Next, there is a strong link between QAT and QATR, suggesting that if an SLR applies the QAT tool, it will provide the analysis and results transparently. The strong link is also present between QEA, QDA and RQS suggesting that these variables are conditionally dependent. This also holds for the links SST and SPR, and ToA, ToE and ToD. The centrality plot (Figure 4b) shows that the most central node is ToD, followed by DQA and RQS, highlighting the importance of transparency and data quality. For the closeness metric, the QATR, ToD, DQA, QAT and QFL have the highest closeness among network items. Overall, looking at the strength of indices, the data suggests that QATR, ToD and QAT have the highest impact on the network structure. An interesting remark is the shared negative connection between QFL and ToD. Previous work on sharing negative connections within NA with regularised partial correlations is limited. Thus, the negative edge between QFL and ToD calls for an explanation. This can suggest that SLR studies that detect specific mechanisms, factors, interventions, etc., may lack explicit data within synthesised studies, controlling for all other associations within the network.



(a) **(b)**

Figure 4. The RCE network structure (a) and centrality plot (b) of the “Big 3” items

As with any CI, non-overlapping CIs indicate statistical differences in a given test, hence observing the CIs of edge-weights QAT-QATR, ToD-ToE, SST-SPR, RQS-QEA, QEA-QDA, and ToA-ToE are the six most important bootstrapped edge-weights (Figure 5). Next, we want to determine the stability of centrality indices by estimating the network based on the case-dropping bootstrap. The estimated network model, when the average correlation does not significantly change after 10% (suggested by (Epskamp et al. 2018)), can be considered error-free (Figure 6). Considering the correlation stability coefficient, *CS*-coefficient in short, the CS (cor = 0.7) represents the maximum proportion of cases that can be dropped to retain 95% probability of 0.7 or higher correlation between original centrality indices and centralities in networks based on subsets. The CS coefficient should not be below 0.25 and preferably above 0.5 to interpret centrality differences. A maximum drop proportion for retaining a 0.7 correlation with 95% CI shows the node strength is 0.674, which is higher than the recommended threshold of 0.5, considering the network to be stable (Epskamp et al. 2018).

The node and edge difference tests are performed with non-parametric bootstrapping. The results (Figure 7) show significant differences, primarily considering strength, where difference mainly cover QAT, QATR, ToD, QEA, and SPR within retrieved studies. The edge-difference test (Figure 8) obliges with an edge-weight difference (Figure 9), suggesting similar differences between edge weights.

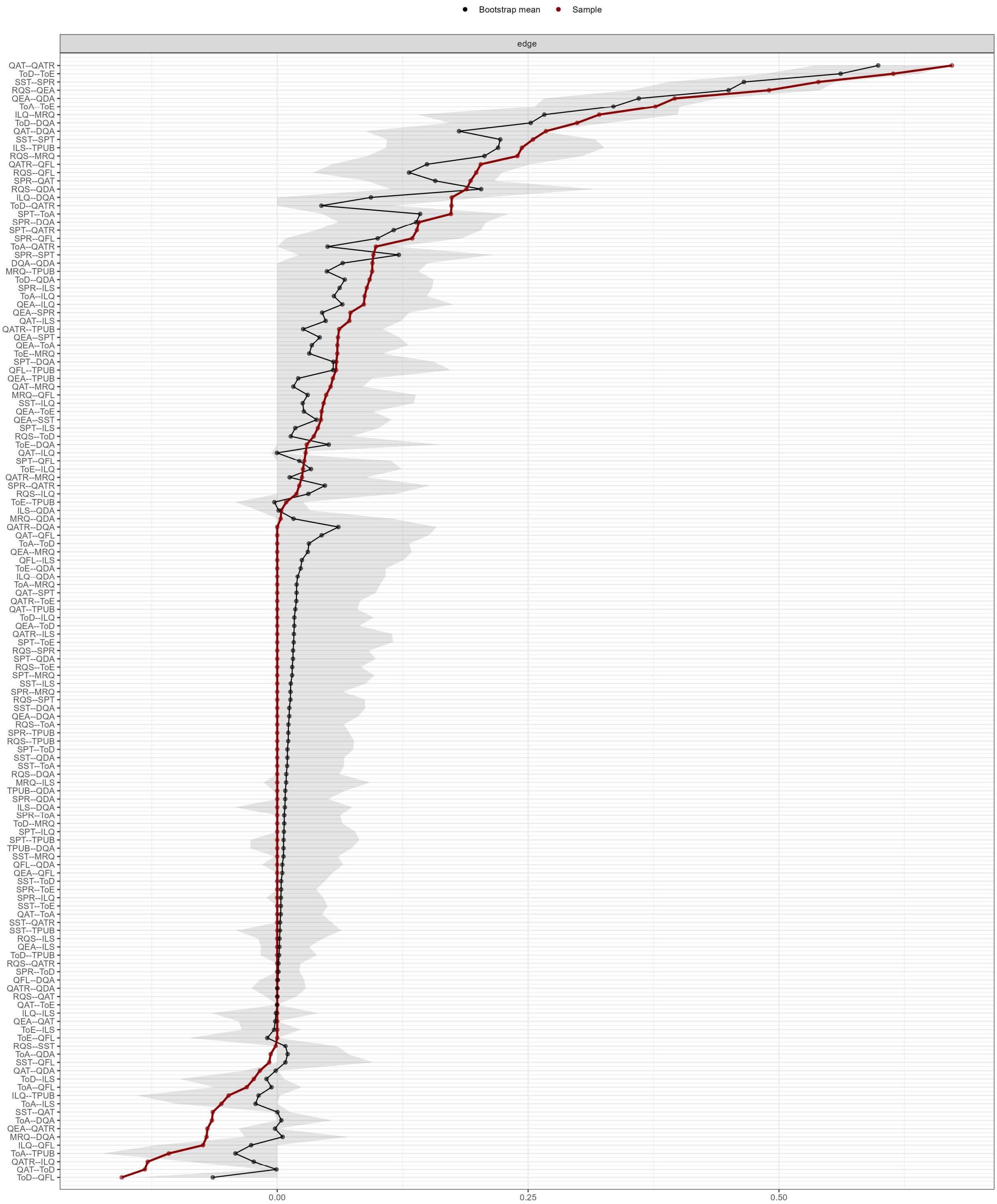


Figure 5. Bootstrapped edge weights with 95% confidence intervals (grey area). The edge weights are sorted from highest to lowest on the *x*-axis, and edges are given on the *y*-axis.

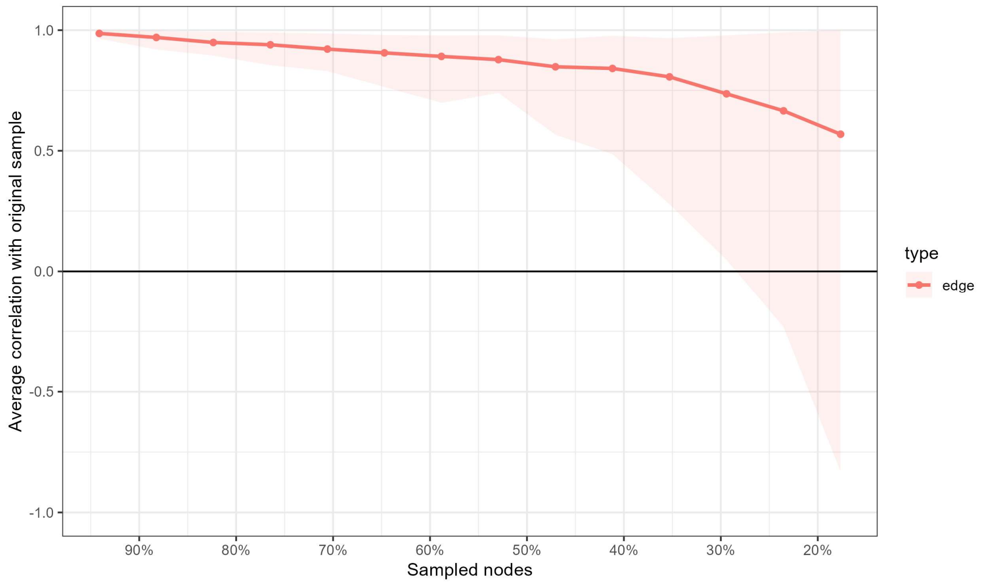


Figure 6. Centrality stability of the RCE network depicting edge values in the original and estimated network with fewer sampled nodes. Line defines mean score of the edge and estimated range is presented with 2.5 to 97.5 quantiles.

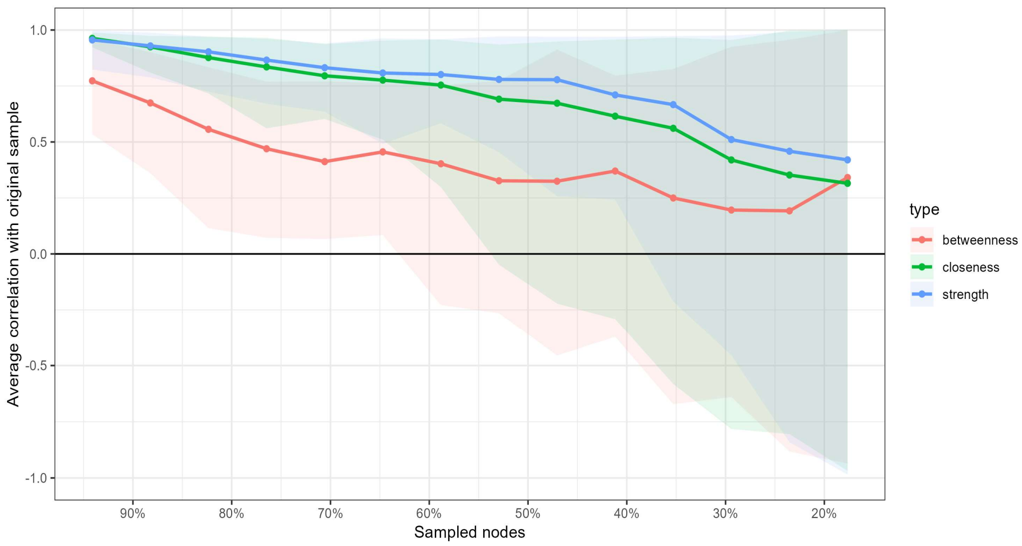


Figure 7. Centrality stability of the RCE network depicted with subset bootstrapping. The centrality stability demonstrates the correlation between the centrality values in the original network and networks with % of sampled nodes. The centrality indices show betweenness mean (red line), closeness mean (green line), strength mean (blue line) and estimated range from 2.5 to the 97.5 quantiles.

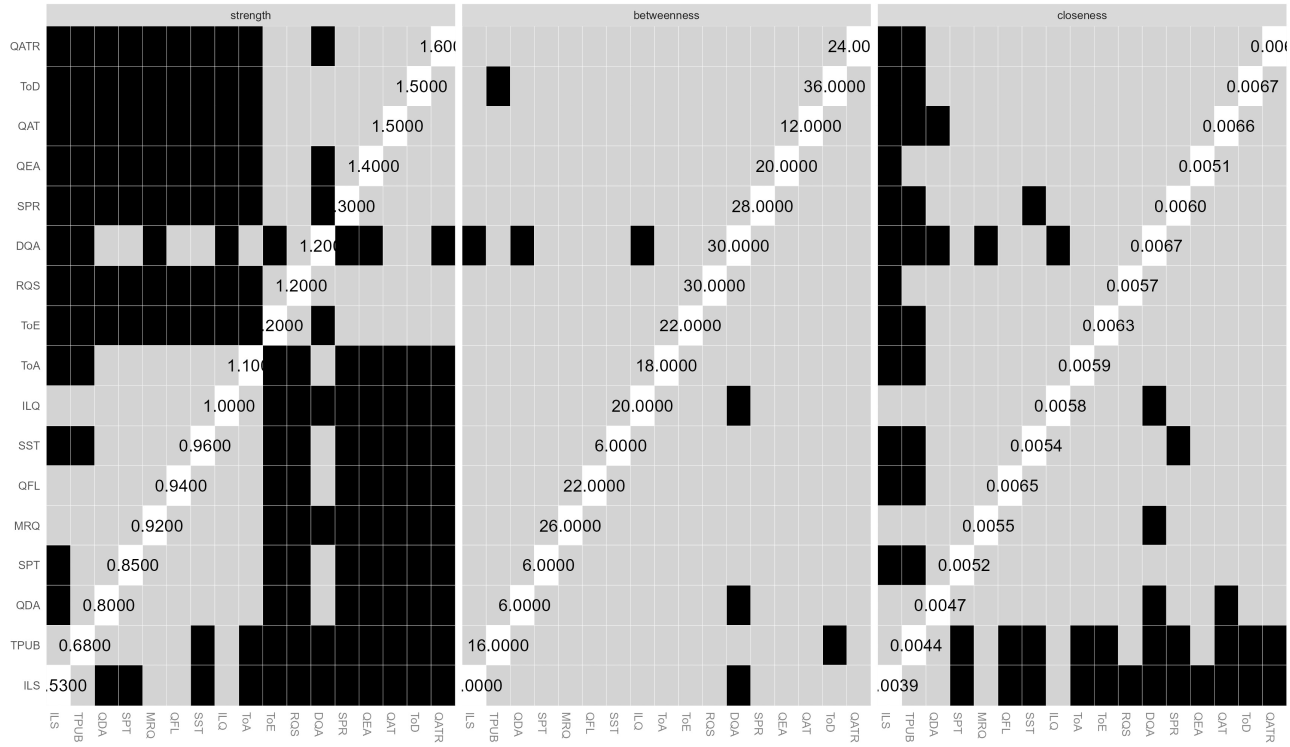


Figure 8. Centrality difference tests considering centrality indices. Statistical difference is represented by the black box while non-statistically significant is depicted by the grey box.

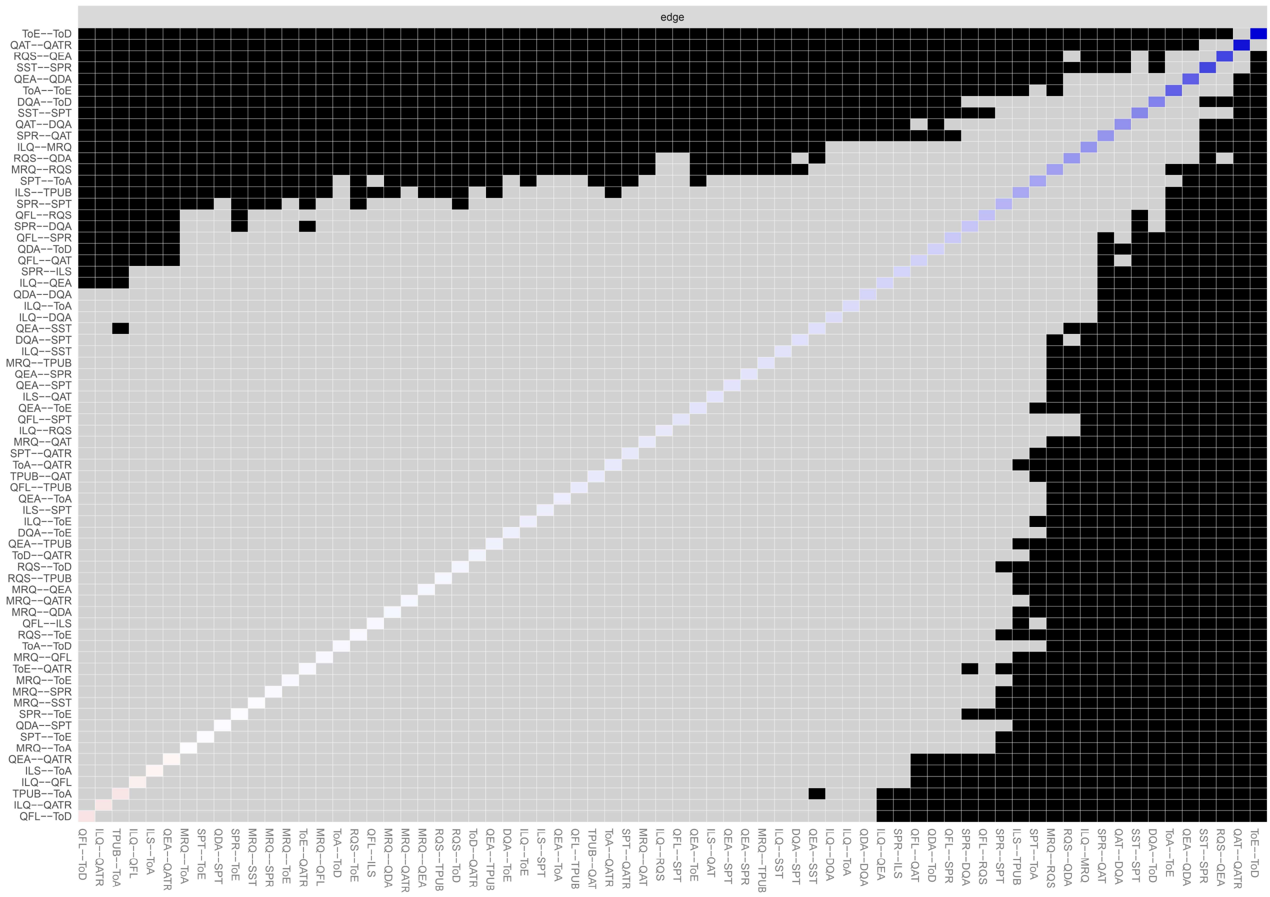


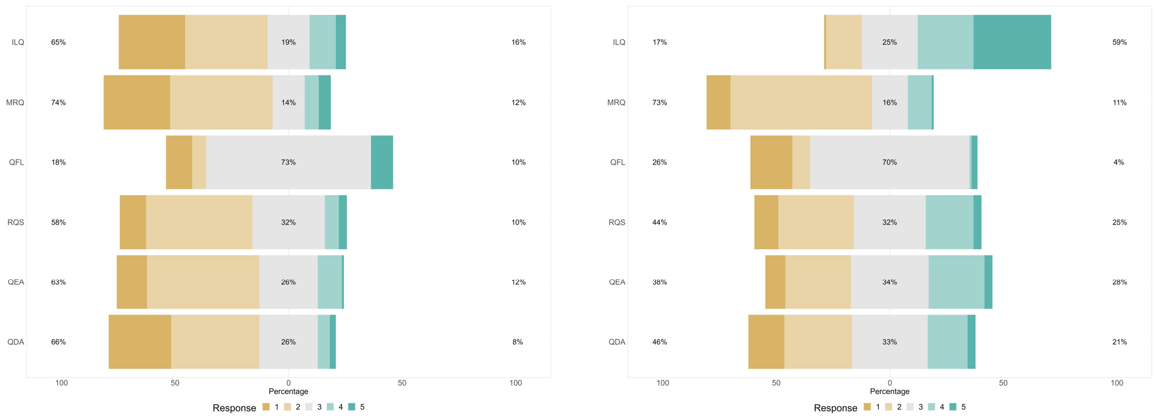
Figure 9. Bootstrapped difference test (α = 0.05) between edge-weights with non-zero nodes.

4 Discussion

4.1 Features associated with citation impact

As indicated by the previous research on scientometrics (Blümel and Schniedermann 2020; Y.-S. Ho and Shekofteh 2021; Wagner et al. 2016, 2021; Xie, Gong, Li, et al. 2019) there is an observed effect between author-level and paper-level bibliometric data. Namely, consistent with previous findings regarding authors' h-index, average h-index and max h-index, alongside the number of publications of the first author, there is an effect size within SLR studies considering correlation between author-level metrics and citation impact. Although the observed effect indeed points a higher theoretical probability than an SLR will gain more citations with the presence of author with high *h* index, this is not always the case. Some papers with poor *h* index received higher citation count. We suspect that the potential reason for this is that nowadays many early career researchers (e.g., PhD students) engage in the production of an SLR study with an interesting topic. The number of authors per paper and university ranking did not show any significant results considering citation impact.

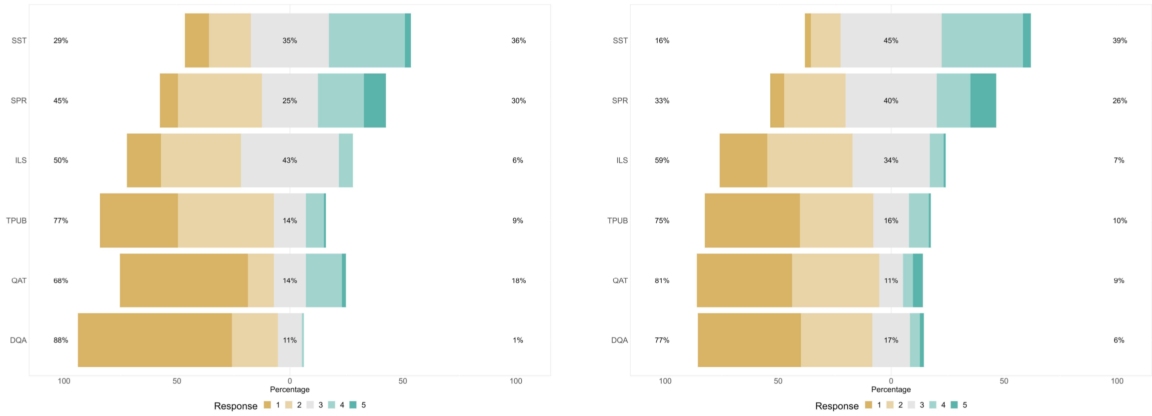
Considering RCE items, we delve into Rationale criterion. Namely, after partialing out the effect of control variables the results point out that ILQ score has the highest difference between SLRs varying between 3.754 (95% CI[3.547, 3.962]) of top-ranked compared to 2.257 (95% CI[2.046, 2.468]) bottom-ranked SLRs. The likert plot (Figure 10) shows that outside of ILQ, little variation is noticed for other RCE items, where MRQ show high frequency of responses where SLR adressess motivation for the study but not elaborating on the each specific RQ proposed. The significant differences of QEA and QDA, shows that there is higher frequency of responses of top-ranked SLR where authors critically appraise evidence alongside with data to allocate gaps with proposed RQs. Namely, considering RQs of top-ranked SLRs, there is higher tendency that the study will critically apraise, synthesise and compare evidence (and data) across studies, leading to the underlying reason of potential citation impact of QEA and QDA on citation rise.



(a) (b)

Figure 10. Likert plot of Rationale items considering (a) bottom-ranked SLRs and (b) top-ranked SLRs

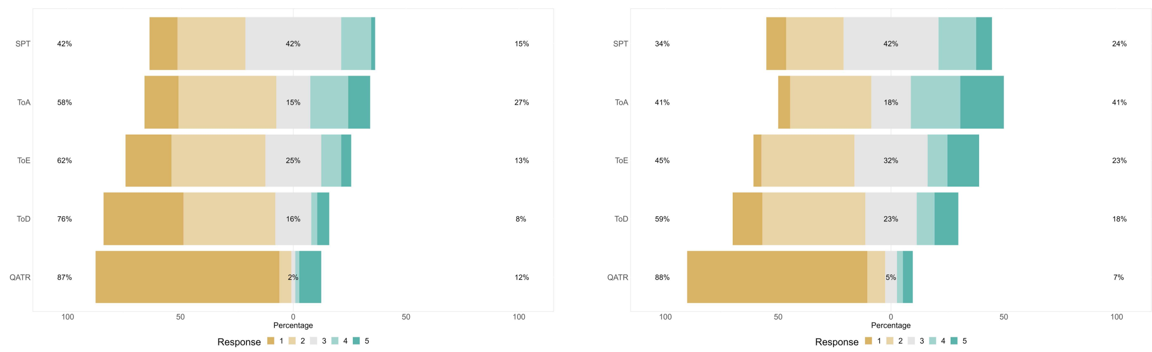
Considering Cogency criterion, the evidence suggests that only DQA shows the existence of significiant assocation with the citation impact. Namely, there is a slight increase in frequency of responses (Figure 11) considering SLRs that, instead of excluding data quality assessment, either implicitly address some data quality dimensions (e.g., discuss missing data) for explaining the outcome or bias in the results of interest, or, explicitly assess and critically appraise data quality dimensions suggesting the underlying reason for previous findings and biases, which leads to paper potentially providing research gaps and consequently being more cited.



(a) **(b)**

Figure 11. Likert plot of Cogency items considering (a) bottom-ranked SLRs and (b) top-ranked SLRs

Considering Extent criterion, the partial correlation analysis shows statistically significant association of SPT and ToD with citation impact. However, although previous results were consistent using both Pearson's and Spearman's correlation; however, the Spearman's correlation did not report the association of SPT with citation impact. Namely, the SPT results suggest a tendency of top-ranked SLRs to report explicit search strings used for the search, eligibility criteria and obtained results that can be replicated (in the paper or even in the attachment), instead of providing a narrative description of selection process (e.g., search strings, selection criteria), which potentially leads to higher association with citation impact. The effect of correlation between score of ToD and citation impact shows much higher frequency of responses where SLRs do not provide transparent data obtained from retrieved studies, in contrast with top-ranked where authors tend to provide (e.g., through tables) and even describe (e.g., provide descriptive statistics) from retrived studies. Also, the presence of association between ToD and citation impact is due to the fact that these SLRs are machine and deep learning studies and quantitative in nature, while bottom-ranked SLRs with score of 1 are mostly qualitative. Hence, since the effect of the presence of association between ToD and citation impact is mostly associated with the topic of explicit data processing in machine and deep learning instead of qualitative analysis, the association and causality is inconclusive and is rather due to the topic of interest.

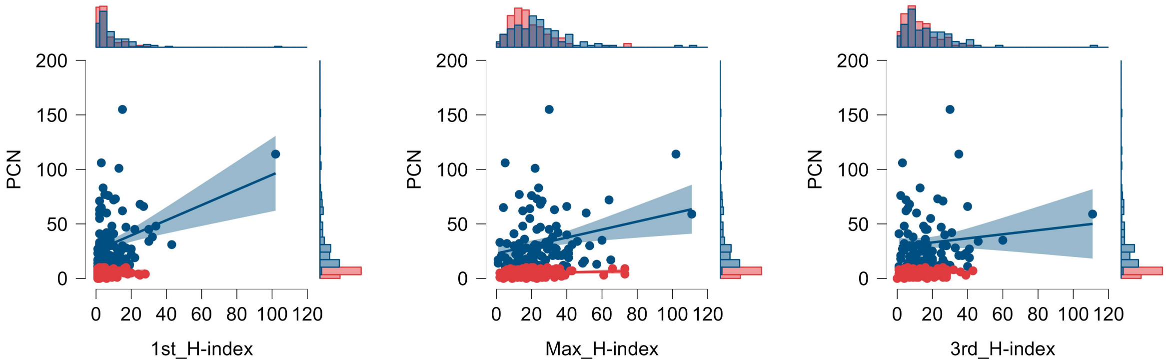


(a) **(b)**

Figure 12. Likert plot of Extent items considering (a) bottom-ranked SLRs and (b) top-ranked SLRs

4.2 PLS-DA analysis results

Looking at the overall PLS-DA analysis considering both bibliometric and RCE variables the evidence suggests that the most impactful bibliometric indices include first author h-index, max h-index and third author h-index. The results seem to be consistent with previous findings in scientometrics. Argubly, first authors with low h-index that publish impactful work, they are, presumably, under the mentorship with an experienced co-author.



(a) (b) (c)

Figure 13. Bibliometric features influencing citation impact: (a) first author h-index; (b) max h-index of authors from all included authors on the paper; (c) third authors h-index.

Based on the analysis of the overall VIP score, we have depicted six most important features of top- and bottom-ranked SLRs, namely ILQ, QFL, RQS, QEA, MRQ, and DQA (Figure 14). Depicting RCE features against citations (Figure 14), we observe the following. The informal logic behind the research question (i.e., ILQ) shows significant association between citation at top-ranked SLRs, which is also confirmed by the analysis of partial correlation. However, although the PLS-DA algorithm extracted features based on the contribution for classification of items, there is an obvious setback considering the distribution of citation score. Namely, observing distributions of RQS, QEA, MRQ and DQA, the data shows the tendency between these items and theoretical probability of receiving higher citations.

Although previous items with higher score tend to contribute to the separation of top-ranked SLRs, with question formulation logic (i.e., QFL) item there is a higher frequency of responses at higher scores, which is contrary to the assumption that higher score of question formulation logic will lead to a higher theoretical probability of receiving higher citations. In contrast, using an explicit question framework (e.g., PICO) tend to contribute more to the separation of bottom-ranked SLRs and lower citation count.

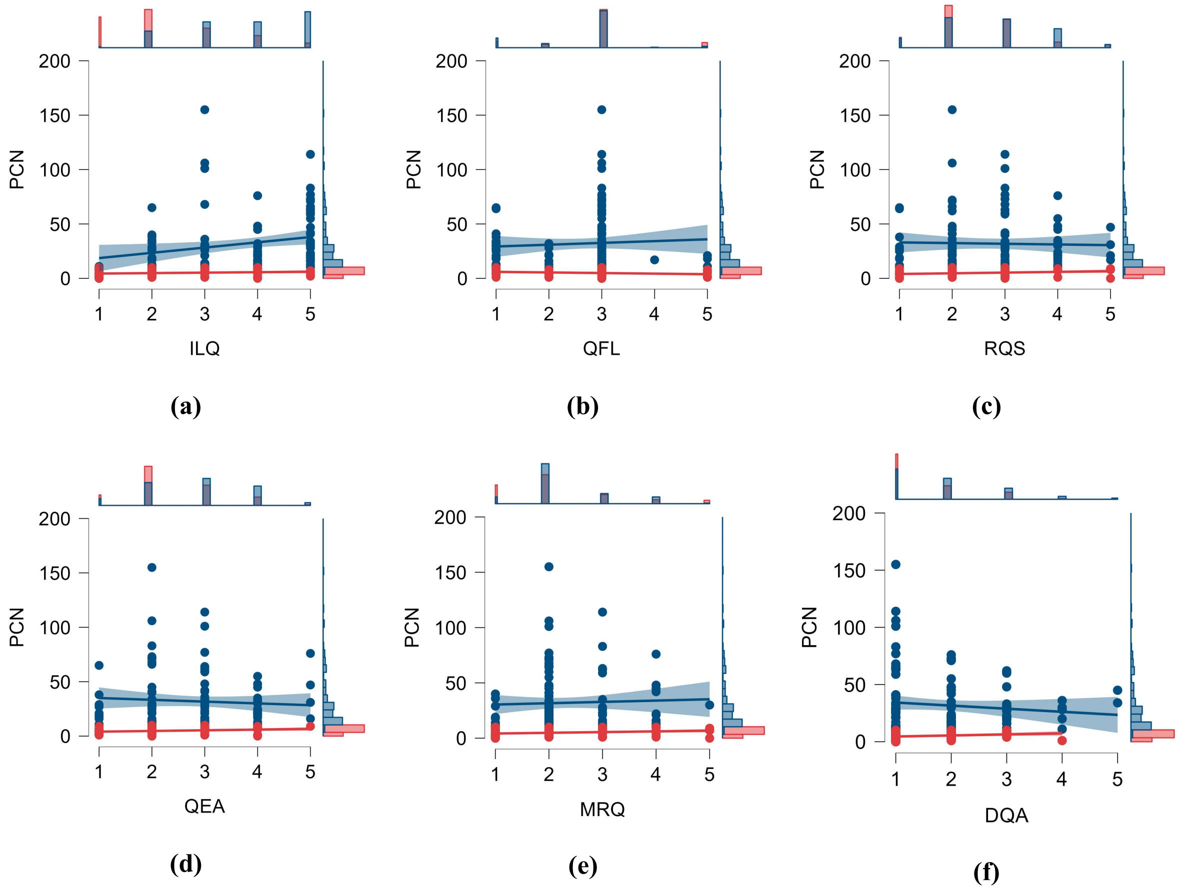


Figure 14. RCE features influencing citation impact of top- and bottom-ranked SLRs: (a) informal logic; (b) question formulation logic; (c) research question strength; (d) question evidence aim; (e) motivation for research question; (f) data quality assessment.

4.3 Network analysis results

Conducted NA of the "Big 3" journals (n = 227) provided exciting remarks. Namely, type of publications (i.e., TPUB) and information literature sources (i.e., ILS) seem to be conditionally independent within the RCE network, suggesting low association and impact between nodes. This can suggests that changes in the selection of studies or type of literature sources does not cause changes in other nodes in the network. Presumably, the independence can be attributed to the fact that an extensive amount of evidence does not suggest changes in the network structure. Looking at the strength centrality, the QATR, ToD, QAT, QEA, and SPR seem to be the most impactful features in the network, which is also supported by bootstrapped centrality difference test. The impact of transparency of data presumably illustrates the validity and the depth of the analysis, which drags other transparency items (e.g., transparency of articles – ToA, transparency of evidence – ToE). This can also be said for QAT item because such studies move away from tokenistic citation towards actual evaluation and assessment of retrieved studies. In addition, the SPR follows a similar pattern since it moves from general to step-by-step description of the study selection process, which ultimately impacts SST and SPT. Thus, the more in-depth description and rationale behind study selection is provided the more it resonantes with validity, which ultimately impacts the network structure.

Contrary to initial assumptions positioning QATR as the most influential node, the evidence suggests that SLRs incorporating QAT do not necessarily yield a higher incidence of reported QAT score. Although difficult to come to a reasonable conclusion from just the NA analysis, we suspect another underlying variable explains the QATR strength. To further inspect this, we performed heatmap analyses examinig the relationship within RCE network and citation numbers (Figure 15) with respect to QATR scores. The results substantiate the pivotal role of QATR revealing that across categories elevated RCE scores were generally observed in top-ranked SLRs. Namely, from the each composite scores (Figure 15a, b, c), the results show that under every class category, the RCE scores (13/15) favoured the top-ranked SLRs considering the QATR level. Intriguingly, the QATR in both case studies yielded non-significant difference ($p = 0.061$, $p = 0.563$). This lends support to the fact that the increase in QATR conditionally augments score of other nodes (items), thereby elevating RCE composite scores and, by extension, the citation impact. In essence, this could mean that authors who provide more detailed and transparent quality assessment results might be perceived as more trustworthy, potentially impacting the reliability of the study. The lack of significant difference in QATR suggests consistency in the way QATR are reported.

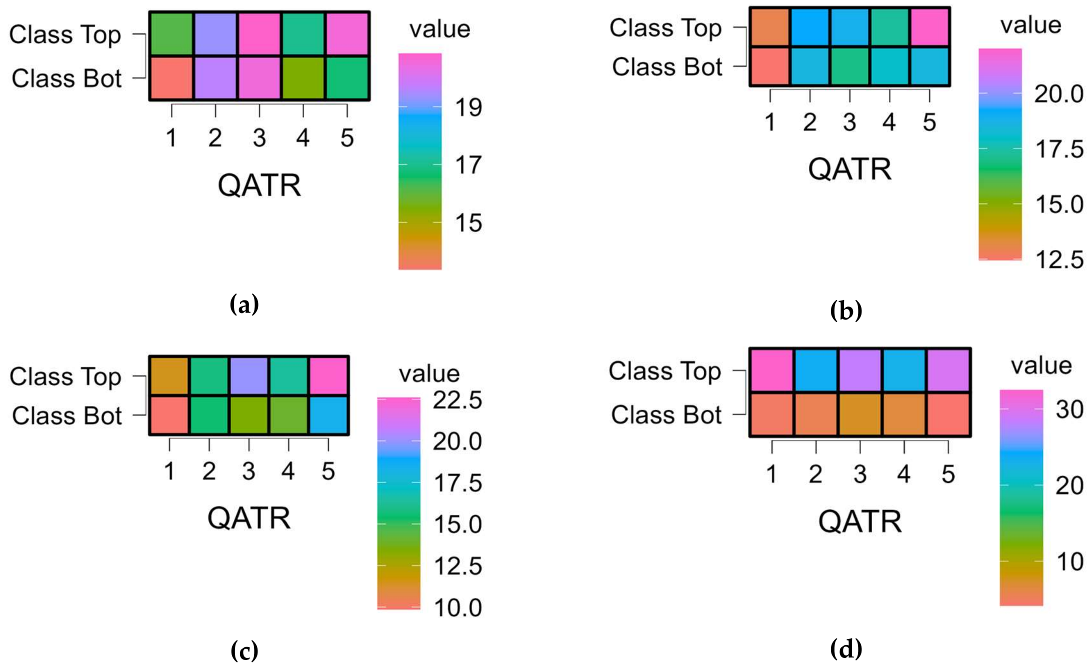


Figure 15. Analysis of QATR by split of (a) Rationale composite scores; (b) Cogency composite scores; (c) Extent composite scores; (d) PCN of included SLRs.

Combining the analysis of PLS-DA and NA we derive several remarks: (a) the transparency and quality of reporting is crucial; (b) the tools and methodologies employed in assessing the quality of retrieved evidence is vital; and (c) the theoretical probability of research impact with strong informal logic and by aligning RQ to the evidence it seeks to appraise is pivotal. Thus, we believe that the proposed scales defined captures a continuum from the absence or low-quality research impact to a comprehensive and transparent presence. Overall, the findings advance our understanding of determinants of both bibliometric and content quality metrics in SLR studies.

4.4 Limitations

In accordance with prior scientometric studies, the present study is not without its limitations. Namely, as the focus is confined to engineering domain, specifically selected journals that published the most SLR studies, it is difficult to generalise the conclusions to other disciplines outside of engineering-based domain. Next, the engineering research within SLRs varies and transcends to different domains (e.g., industrial, biotechnology, sustainability, logistics, supply chain), which may also be subjected to the impact of the findings. Additionally, our study omits various dimensions of research impact (e.g., altmetrics) focusing solely on citations, which may impose bias (Eysenbach 2011) offering a one-dimensional view of citation count (Adler et al. 2009). Next, features like writing style, topic trend, etc., are not considered, which in fact causes bias since these variables report the presence of the effect on citation count. This is observed in studies dealing with advanced technological issues on the topic of Industry 4.0, Deep Learning, and Smart systems which tend to attract more citations. Another bibliometric factor that make SLRs more likely to be cited include factors such whether the corresponding author is from developed country; whether the study received external funding, and international collaboration (Minh et al. 2023).

Furthermore, the study acknowledges low test power, stemming from small sample, consequently restricting generalisation of findings. This can suggest that some effects may be missed that are actually significant. While traditional methodologies in scientometrics typically employ generalized linear models, this study utilised PLS-DA. Namely, the rationale for this is that PLS-DA can handle collinearity of predictor variables. Next, the PLS-DA can handle high-dimensional data or cases where predictors exceeds the number of observations, which in return mitigates risk of overfitting. This is a common pitfal of traditional regression models. The PLS-DA focuses on maximizing the covariance between dependent and independent variables thereby offering more robust prediction results of class labels, i.e., categorical outcomes. Also, the PLS-DA facilitates VIP score that is known a variable selection process that identifies important features in the datase, which is not inherently a process of regression models. However, major limitation is that PLS-DA is supervised model that requires a priori categorical labels, whereas regression models do not. This imposes less flexibility of PLS-DA in situations of unsupervised scenarios, thus the selection of class labels based on top- and bottom-ranked split may also impose bias in the estimation of feature selection VIP score.

From the performed NA, the results can be misleading in certain situations. Namely, there is a debate whether intepretation of centrality indices and nodes in a network can be taken solely on their scalability in the NA structure. Epskamp et al. (2018) provides interesting arguments given the analogy of the regression problem when adding or removing predictors will actually change the regression coefficients leading to unstable results, which is the case in the NA. The underlying reason is because the methodology of NA is in its infancy and is still not been fully worked out. Nevertheless, we used EBICglasso that is built on a regularized Gaussian graphical model to strengthen the robustness of identified relationships and make sure that results are not due to random noise. The fact that there is no direct relationship between higher QATR scores and citation impact is interesting. While QATR might be the most impactful in terms of strength centrality, it does not correlate with citation. Hence, pressumbly there are a lot of underlying latent variables that bypass QATR item – topic's popularity, study novelty, applicability, or even authors' reputation, among others. Lastly, the study did not explore the nature of citations, including self-citations and reciprocal citations, adding another layer of complexity to understanding drivers of citation impact.

4.5 Implications

In contemplating motives and drivers for the mass production of SLRs, we exercise caution when postulating on the causality between the intent to consolidate knowledge on a given topic and elevate citation impact. Nonetheless, we believe that saturation of SLRs and "*reviews of reviews of reviews*" (McColl 2022) have caught the attention of editors and policymakers in medical and health domains suggesting that these types of studies are starting to contribute to the research waste (Roberts and Ker 2015). As much as challenging and provoking is to deliver statements that align with “publish or perish” ethos presumably driven by citation metric game, we seek to contribute constructively by introducing criteria that delineates high-quality research. In this regard, our findings complement and enrich those of Wagner et al. (2021), particularly in engineering domain. Although their study demonstrates the importance of transparent methodology and research agenda, our study aligns with findings but also extends and allocates features contributing the research impact measured by means of citation rates.

The proposed RCE targets not only early-career researchers and junior scientists engaging in the production of the SLR, but also to editorial boards and policymakers in delineating high-quality SLRs. The implications we account for are not only to decrease desk-rejection rates but also to increase the validity and reliability of these studies in the long-run. Finally, in support of our arguments regarding the decline in quality and abuse of SLR, an SLR study conducted by Misra and Agarwal (2018) who synthesising *retracted SLRs* (n = 85) associated with fraudulent reasons (e.g., methodological errors, fabricated peer review process, omitted studies), perfectly illustrates substance behind arguments we intent to make.

The strong differentiation exhibited by informal logic behind research question (ILQ) underscores the importance of sound and thorough argumentation scheme. This could imply that the SLR quality and its stakeholders are heavily reliant on the depth and rigor with which the authors' necessity and relevance are established. Moreover, the association of QDA with citation rates highlights the shift from descriptive towards empirical data for substantiating conclusions. This again emphasises the need for an SLR to be explicit in reporting how data is handled and managed. Consequently, it equips stakeholders with greater understanding whether findings are valid and reliable or biased and mishandled. This is demonstrated in SLR studies who synthesise and provide detailed insights of data, and consequently, appear to be more influential.

4.6 Concluding remarks

Systematic Literature Reviews are vital in scoping knowledge, guiding research, aiding understanding and informing public across diverse academic discplines. The quality and impact of these reviews are multi-faceted, influenced by various factors. To allocate these factors, we propose a list of items grouped into Rationale, Cogency and Extent (RCE) criterion. The purpose of this criterion is to offer a comprehensive insight into the variables affecting top-tier engineering-based SLRs. Employing Partial Least Sqaure Discriminant Analysis (PLS-DA), we conclude that informal logic for starting a review is the most impactful feature associated with citation impact. Incorporating both bibliometric and RCE items, the results obtained show that journals’ metrics, authors’ metrics and RCE composite scores show statistically significant disparities between top- and bottom-ranked studies. We also employ Network Analysis (NA) to evaluate the robustness of variables, or nodes, within RCE network, finding that the use and reporting of quality assessment stand out as most influental nodes by means of strength centrality.

Dual application of PLS-DA and NA aims to allocate features predictive of citation impact and those enhancing methodological rigor of SLRs. Although the PLS-DA findings were consistent with previous findings, offering additional insights about RCE scores, the interpretation of NA findings provide conflicting remarks. Namely, the elevated strength centrality of the QATR item underscores its role in fostering transparency and methodological rigour, indirectly affecting other nodes (variables). This leads us to hypothesise that QATR may function as moderating variable given its lack of direct statistical significance.

Lastly, the interplay between QEA, QDA, and ILQ indicate that the cornerstone of high-quality SLRs lies in the clear articulation and justification of its RQ, its alignment with the evidence, and its specific use of data to address the problem at hand. SLRs that offer a sound rationale, challenge extant knowledge, and critically appraise existing evidence are more likely to be contributing valuable knowledge, thus garnering higher citation rates. That is why we argue that these three factors collectively might serve as a blueprint for crafting impactful systematic reviews.

5 Appendix

Appendix 1. Analysis of SCOPUS SLRs across different sources

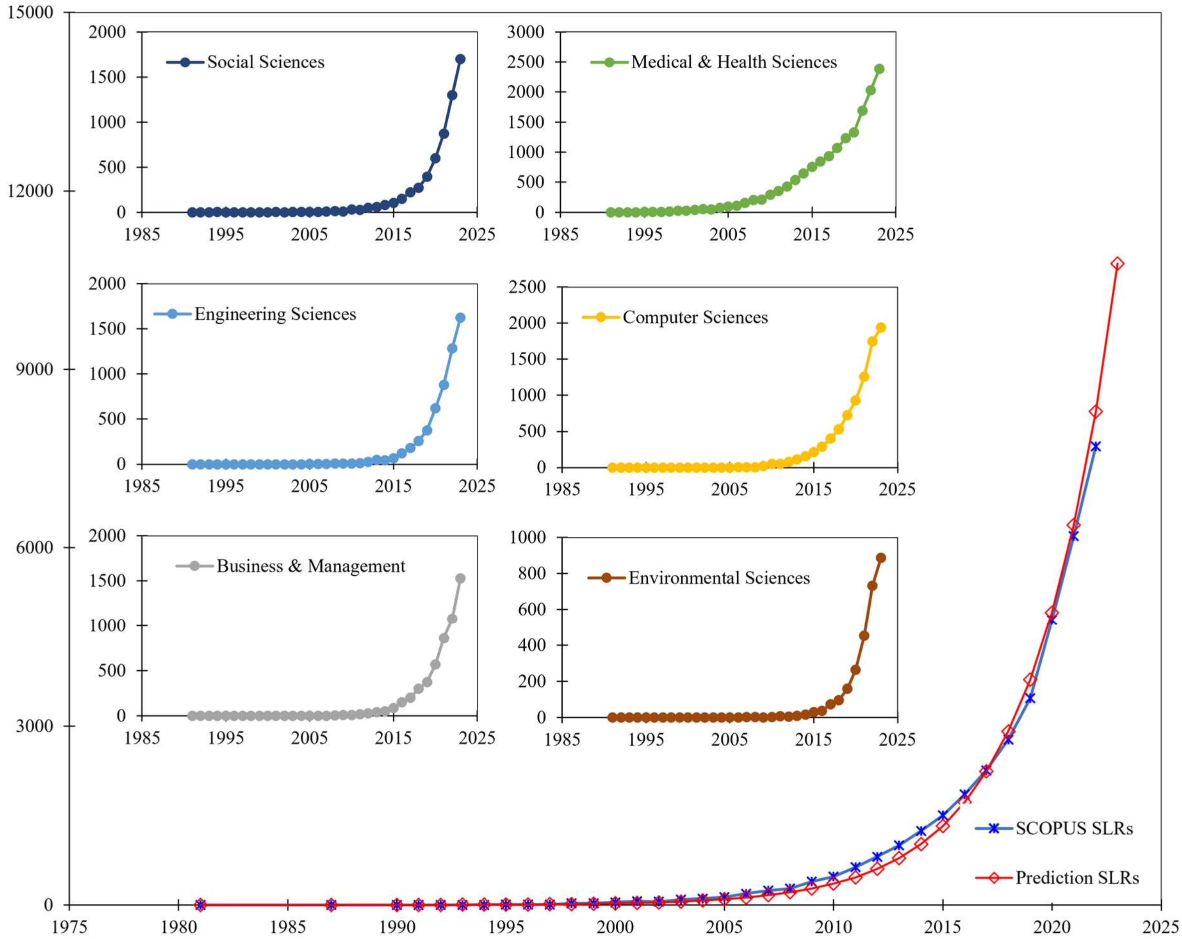


Figure A1. Systematic Literature Review studies across different sources

Appendix 2. Descriptive statistics and analysis of the first sample

Table A1. Descriptive statistics of the first sample

	Mean	Std. Dev.	95%CI$_{\text{Upper-Mean}}$	95%CI$_{\text{Lower-Mean}}$	Std. Error	95%CI$_{\text{Upper-Stdev}}$	95%CI$_{\text{Lower- Stdev}}$
ILQ	3.135	1.254	3.324	2.947	0.096	1.334	1.156
MRQ	2.612	1.089	2.775	2.448	0.084	1.195	0.981
QFL	1.947	1.016	2.100	1.794	0.078	1.151	0.861
RQS	2.500	1.004	2.651	2.349	0.077	1.088	0.905
QEA	2.635	0.989	2.784	2.487	0.076	1.081	0.890
QDA	2.300	0.99	2.449	2.151	0.076	1.083	0.889
SST	2.965	1.235	3.150	2.779	0.095	1.329	1.140
SPR	2.647	1.051	2.805	2.489	0.081	1.134	0.949
ILS	2.506	0.999	2.656	2.356	0.077	1.083	0.904
TPUB	2.082	0.951	2.225	1.939	0.073	1.061	0.833
QAT	2.006	1.128	2.175	1.836	0.087	1.238	0.991
DQA	1.359	0.675	1.460	1.257	0.052	0.762	0.568
SPT	2.741	1.193	2.921	2.562	0.092	1.285	1.089
ToA	2.735	1.276	2.927	2.544	0.098	1.367	1.167
ToE	2.429	1.025	2.584	2.275	0.079	1.125	0.918
ToD	2.176	1.122	2.345	2.008	0.086	1.235	1.006
QATR	1.671	0.972	1.817	1.525	0.075	1.100	0.833

Table A2. Statistical testing results

	U	p	VS-MPR[a]	RBC[b]	95% CI RBC$_{\text{Lower}}$	95% CI RBC$_{\text{Upper}}$
ILQ	2872.5	0.018	5.060	-0.205	-0.364	-0.034
MRQ	3157.5	0.139	1.342	-0.126	-0.292	0.047
QFL	2981.5	0.033	3.272	-0.175	-0.337	-0.002
RQS	3264	0.257	1.054	-0.096	-0.264	0.077
QEA	3453.5	0.604	1.000	-0.044	-0.215	0.129
QDA	2898.5	0.02	4.719	-0.198	-0.358	-0.026
Rationale Score	2898	0.026	3.907	-0.198	-0.358	-0.026
SST	2492.5	< 0.01	134.959*	-0.310	-0.458	-0.145
SPR	2814.5	0.01	8.149	-0.221	-0.379	-0.050
ILS	3923.5	0.312	1.012	0.086	-0.088	0.255
TPUB	3838.5	0.448	1.000	0.063	-0.111	0.232
QAT	3235	0.212	1.119	-0.104	-0.272	0.069
DQA	3262	0.152	1.286	-0.097	-0.265	0.077
Cogency Score	2806.5	0.012	7.039	-0.223	-0.381	-0.053
SPT	2310	< 0.01	1194.988*	-0.361	-0.502	-0.201
ToA	2936	0.03	3.462	-0.187	-0.348	-0.015
ToE	2456.5	< 0.01	262.776*	-0.320	-0.467	-0.156
ToD	2830.5	0.011	7.453	-0.216	-0.375	-0.046
QATR	3081	0.061	2.166	-0.147	-0.312	0.026
Extent Score	2435	< 0.01	187.213*	-0.326	-0.472	-0.163

[a]Vovk-Sellke Maximum *p*-Ratio is given as a two-sided *p*-value explaining the maximum odds favor the H_1 over H_0.
[b]RBC = Rank-Biserial-Correlation for measuring the effect size of Mann-Whitney U test.
*The statistically singificant value according to the Bonferroni correction where adjustmed *p* value is < 0.0025 ($\alpha < 0.05$; $m = 20$).

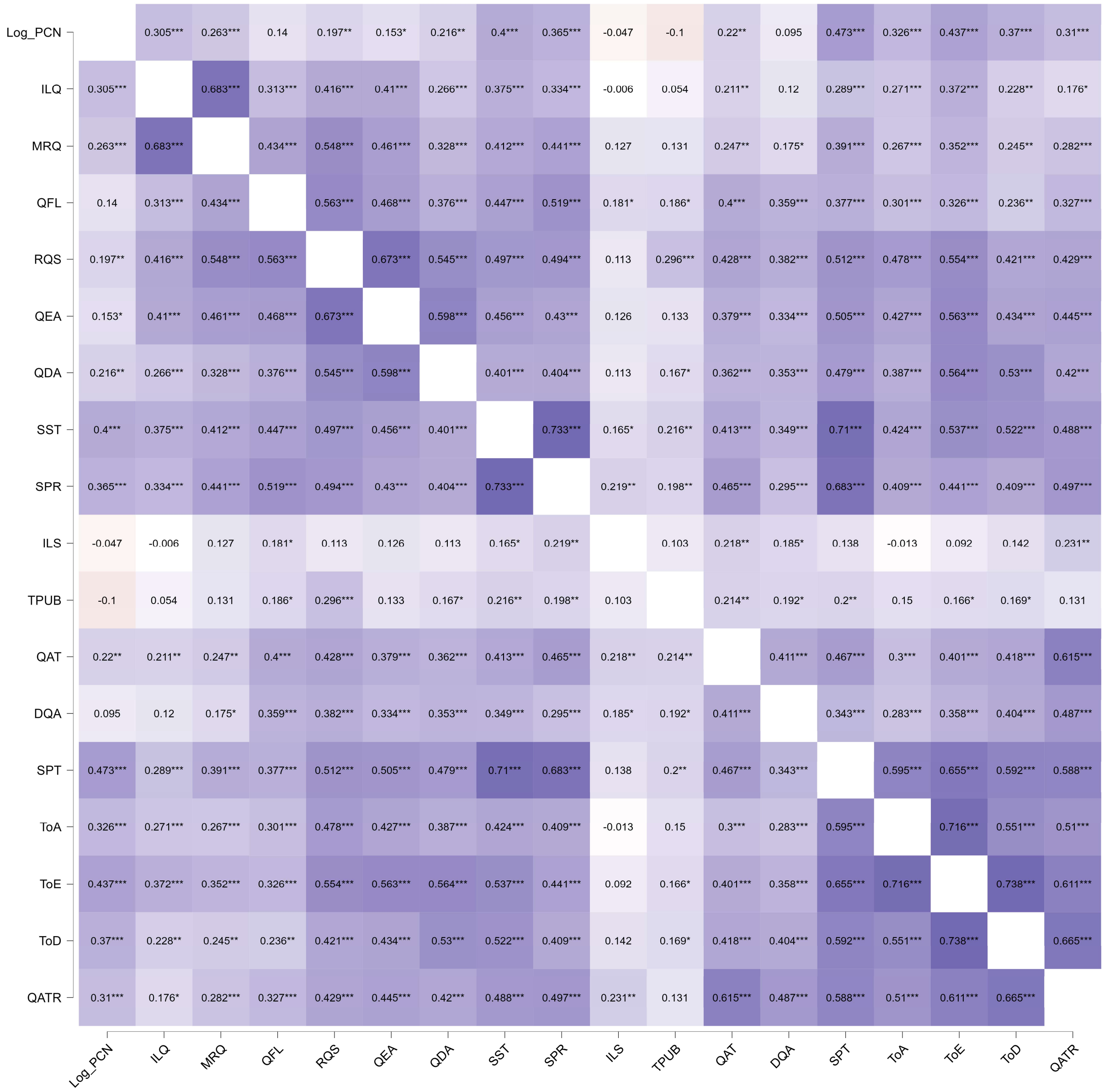


Figure A2. Pearson's correlation with flagged significant correlations (**p*<0.05; ***p*<0.01; ****p*<0.001)

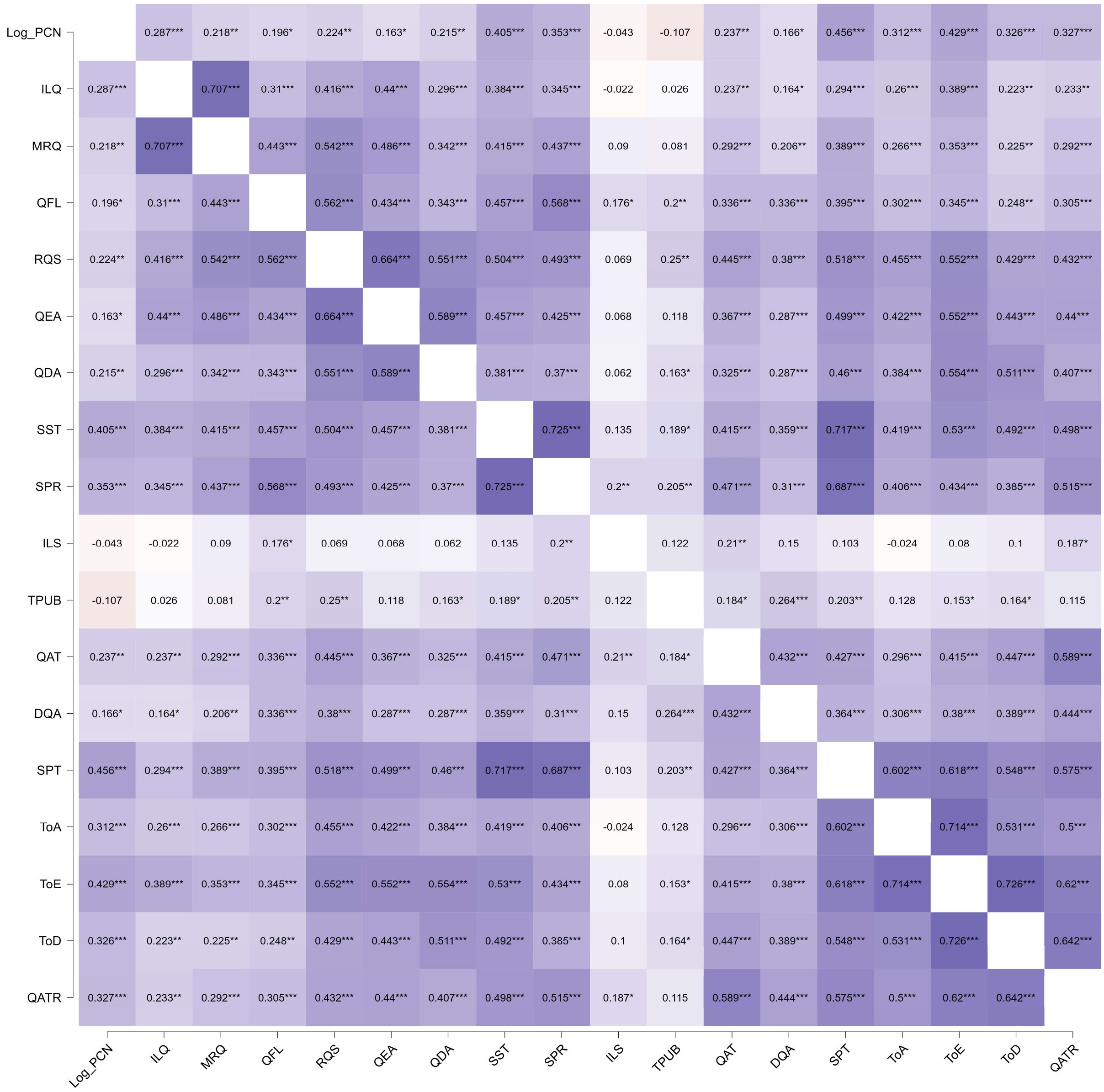


Figure A3. Spearman's correlation with flagged significant correlations (*p<0.05; **p<0.01; ***p<0.001)

Appendix 3. Descriptive and correlation analysis of the “Big 3” sample

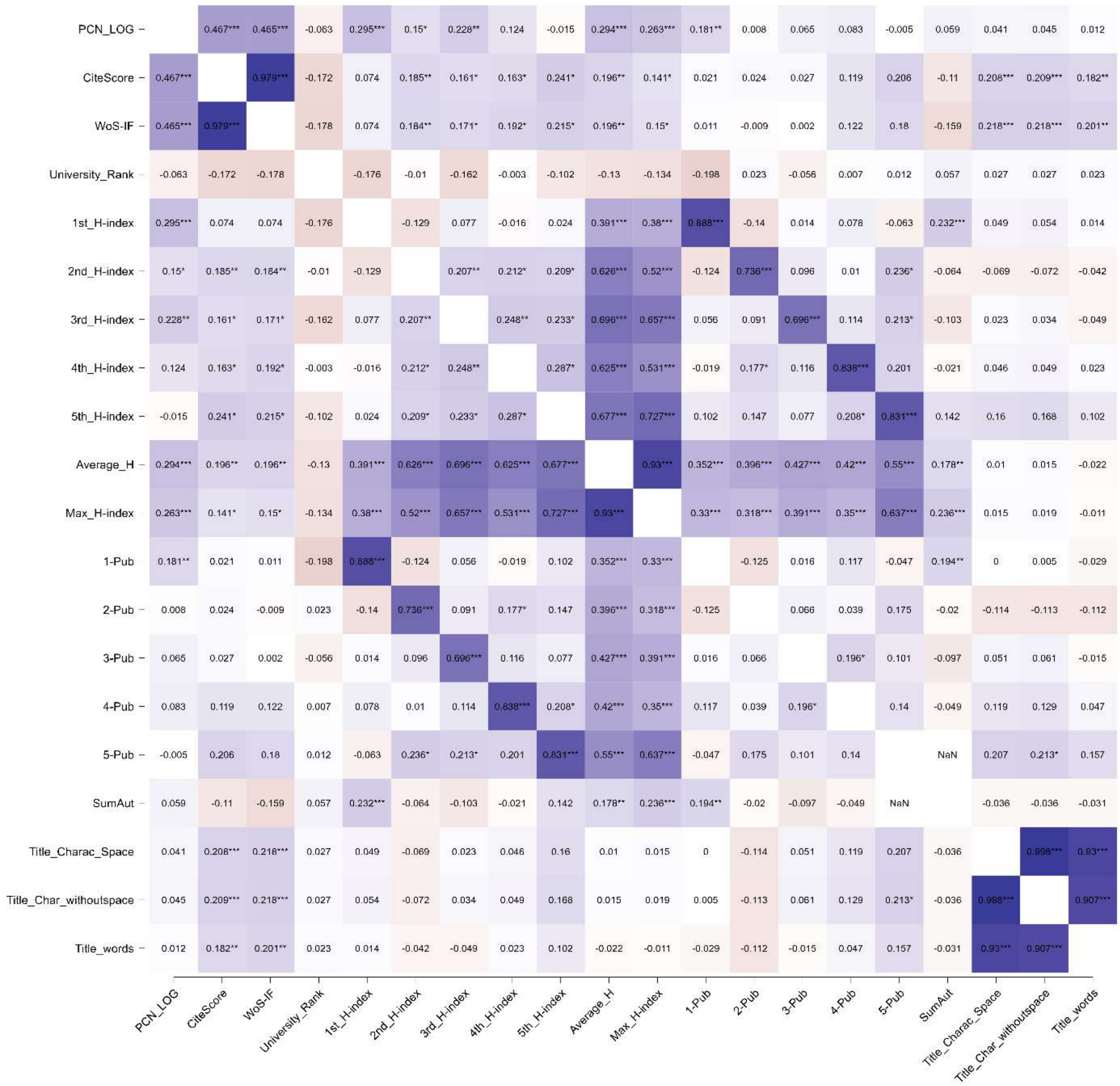


Figure A4. Pearsons’ correlation heatmap of variables associated with citations (**p*<0.05; ***p*<0.01; ****p*<0.001)

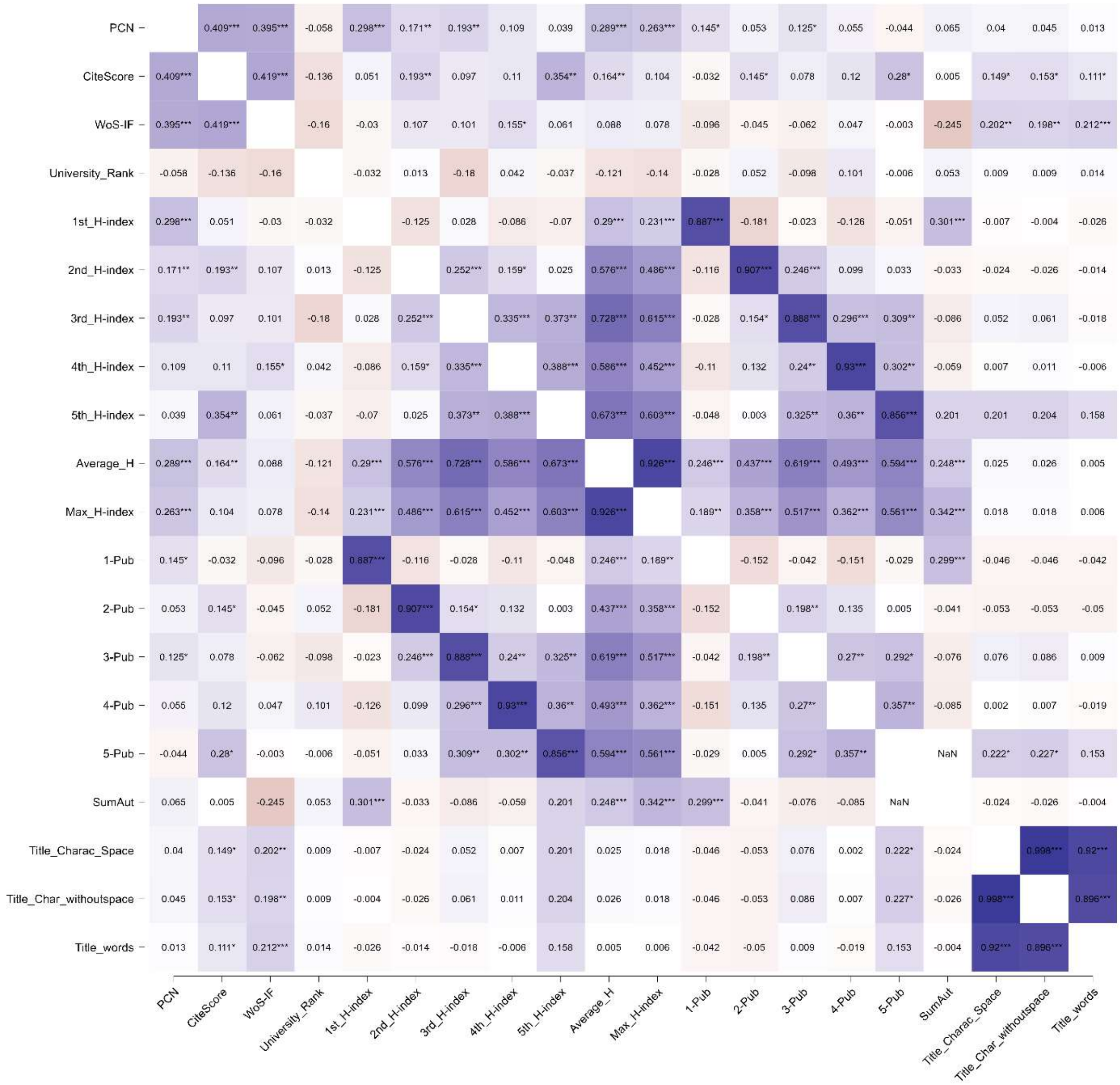


Figure A5 Spearman correlation heatmap of variables associated with citations (*p<0.05; **p<0.01; ***p<0.001)

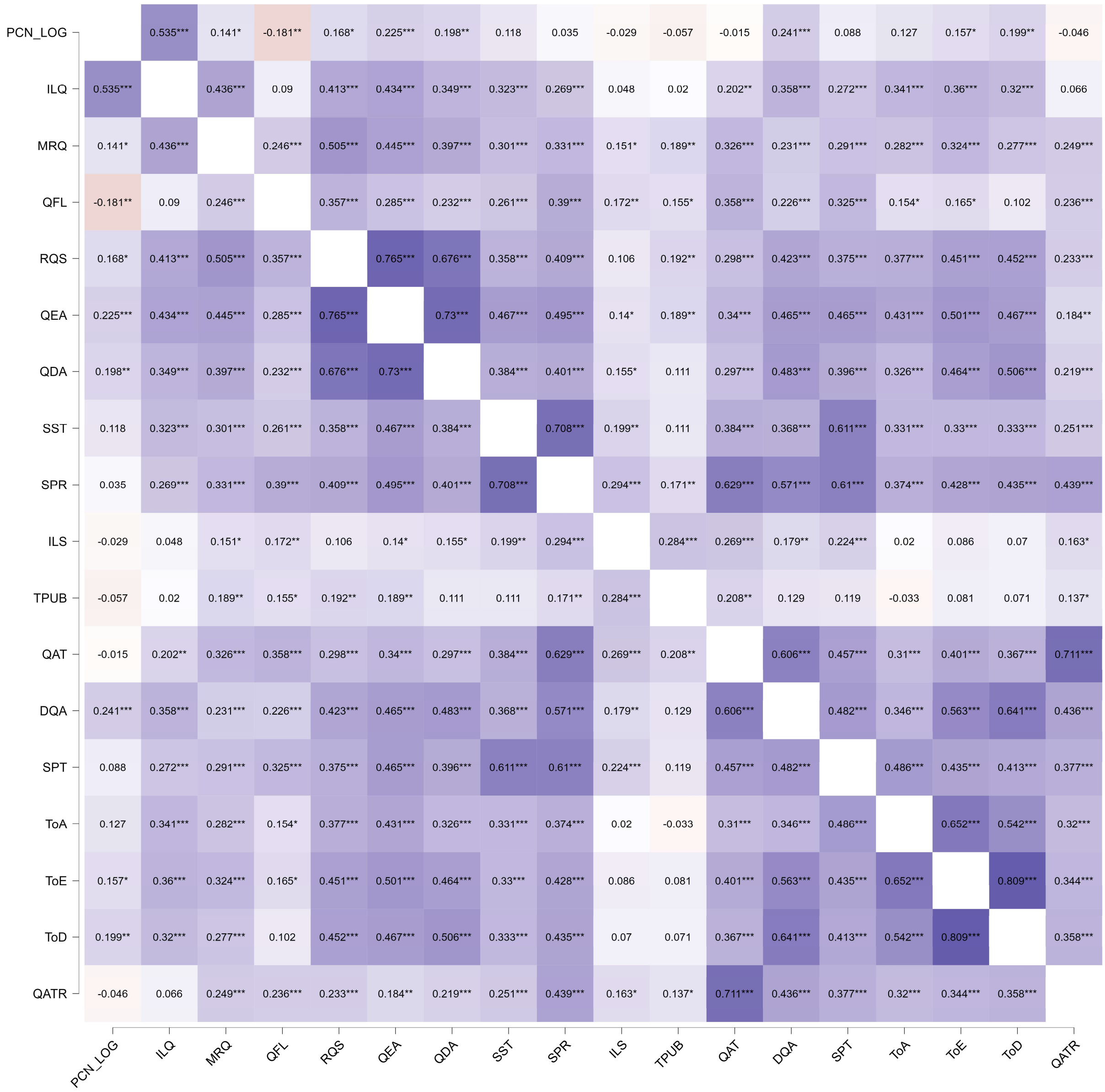


Figure A6. Pearson's correlation between RCE items and log-citations (**p*<0.05; ***p*<0.01; ****p*<0.001)

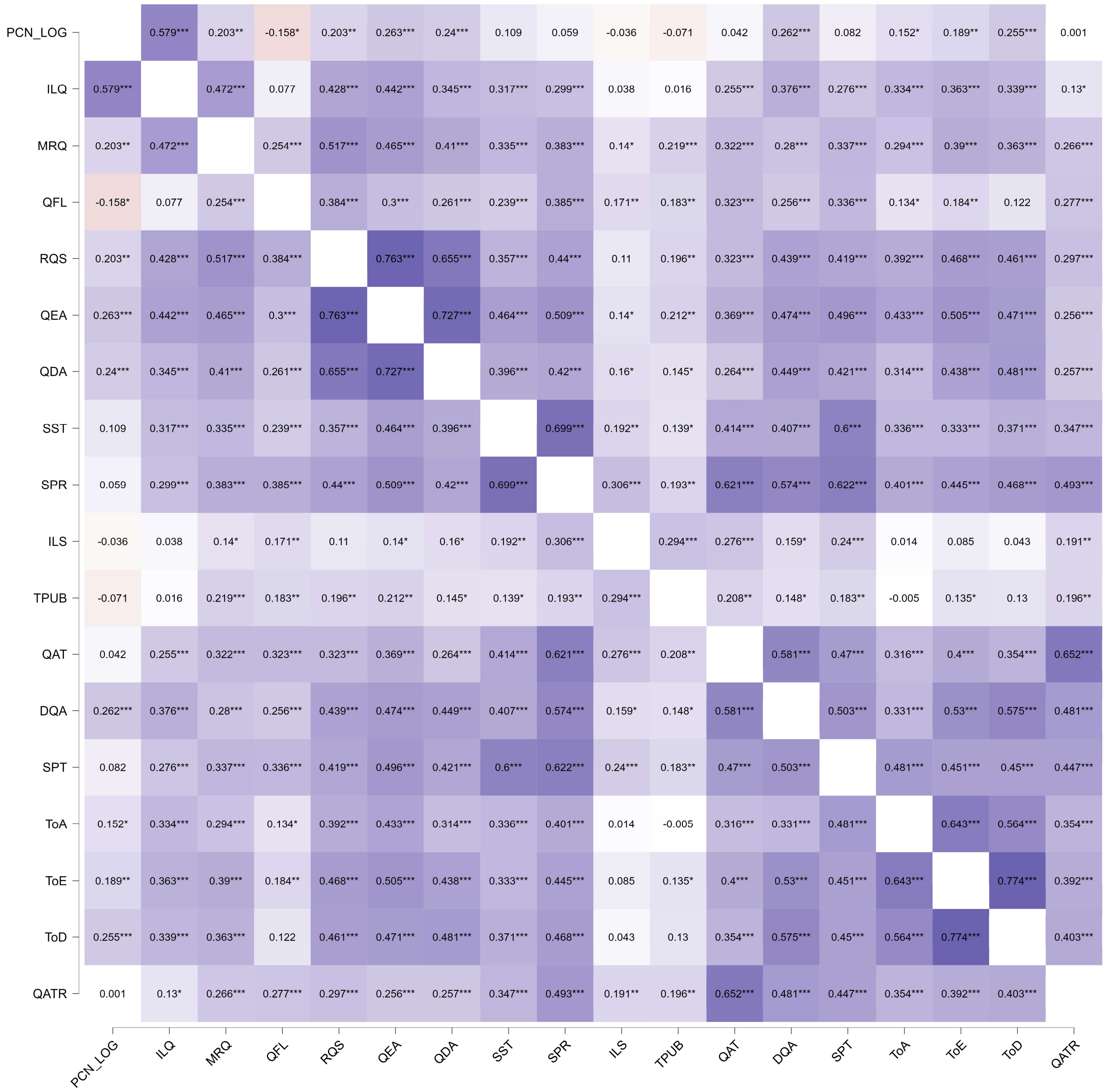


Figure A7. Spearman's correlation between RCE items and citations (**p*<0.05; ***p*<0.01; ****p*<0.001)

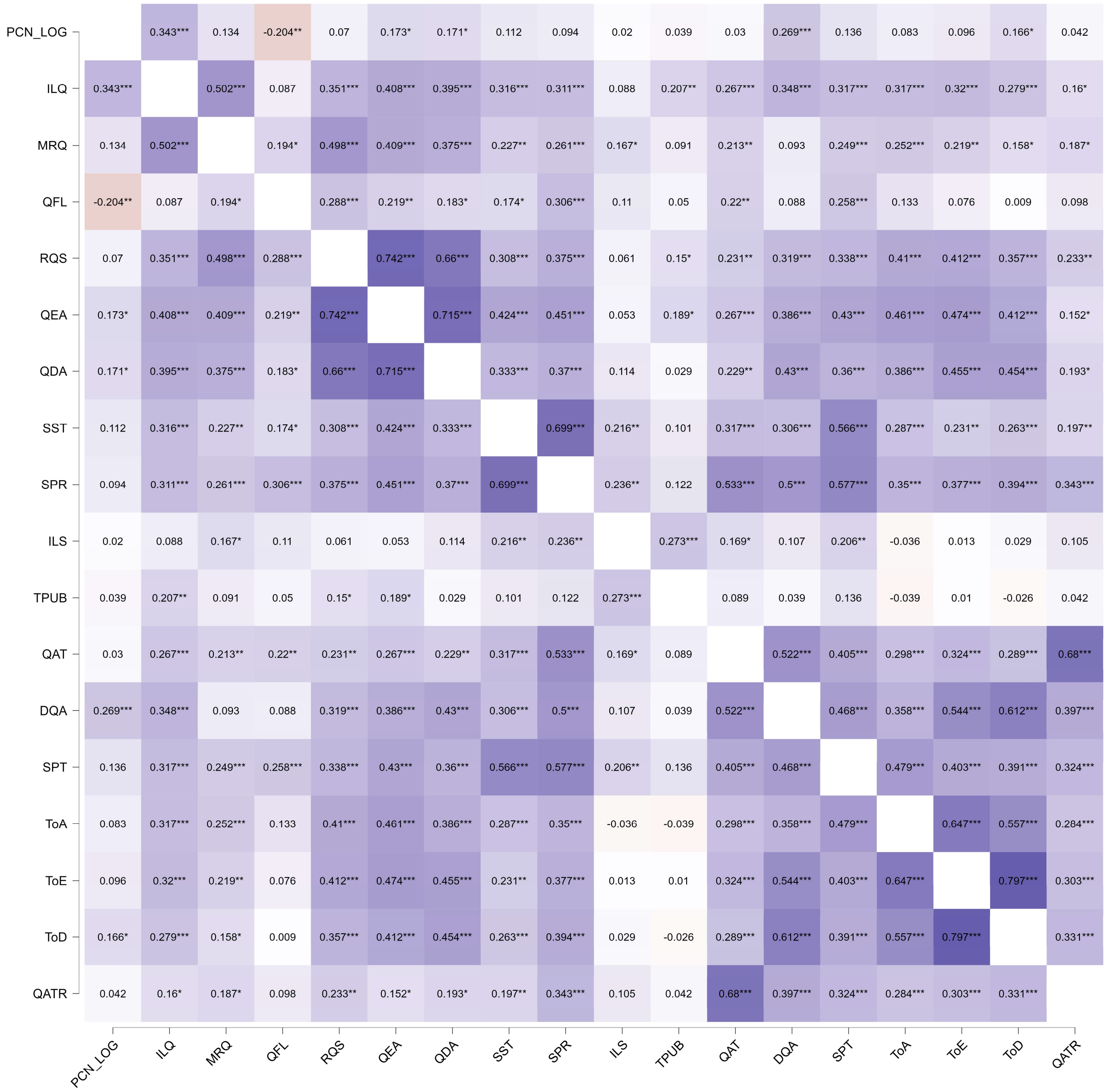


Figure A8. Partial Pearson's correlation heatmap of RCE items and citations ((*p<0.05; **p<0.01; ***p<0.001)

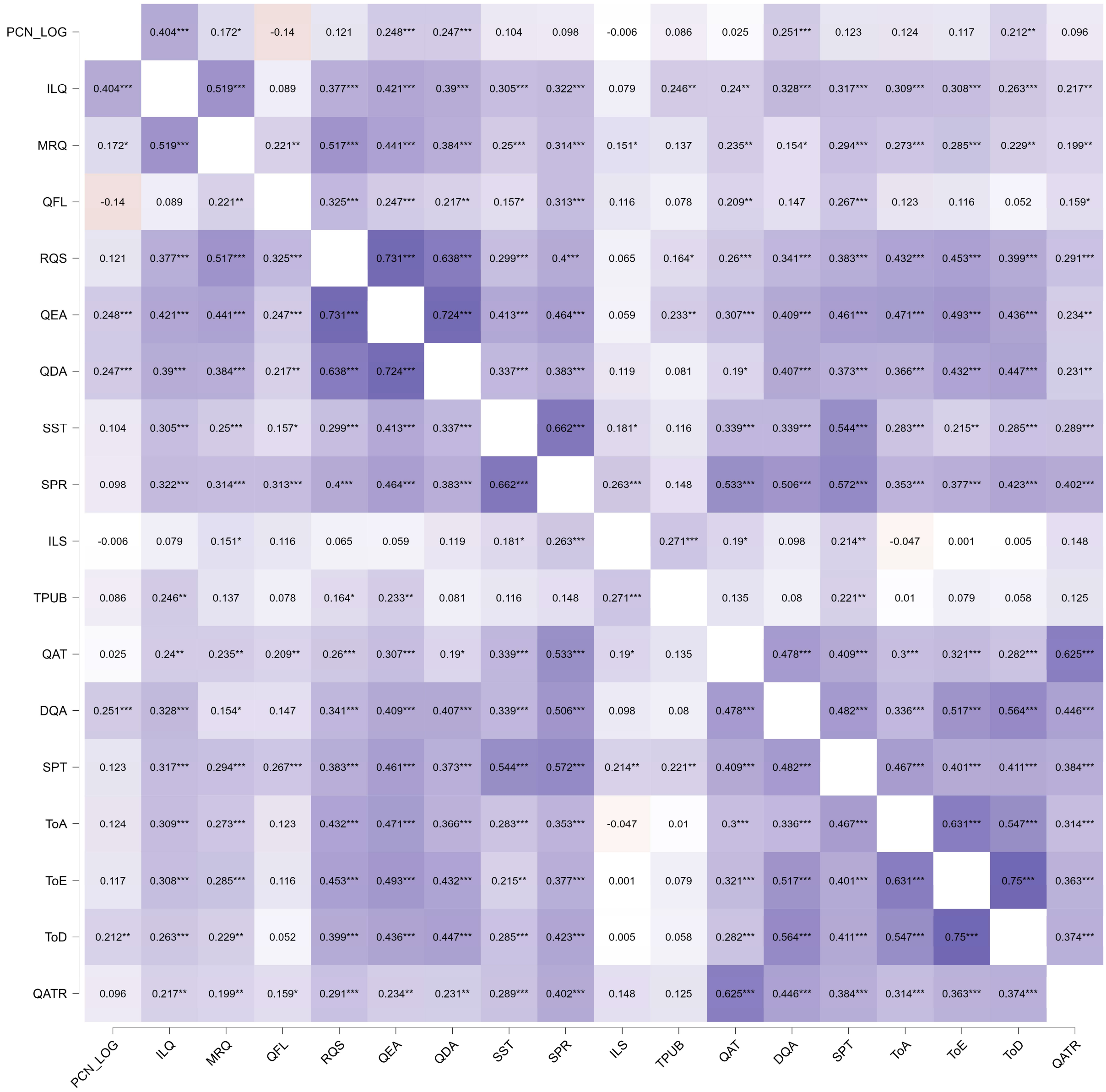


Figure A9. Partial Spearman's correlation heatmap of RCE items and citations (*p<0.05; **p<0.01; ***p<0.001)

Appendix 3. PLS-DA analytical description and results

From an analytical perspective, the PLS-DA transforms dataset X into lower dimension matrix A, where X is $n{\times}m$ matrix, and A is transformed X matrix into a lower dimension of $m{\times}d$-dimensional vectors C, with error matrix E, such that $C = XA + E$. The transformed C contains rows corresponding to the transformed vectors, while the E matrix contains information for the next PC. The difference between PCA's PCs is that in PC1 (or LV1), the PCA preserves the most variance of the original dataset X, while PLS-DA preserves in PC1 (LV1) as much variance but to a target or a class label. Eigenvectors of the covariance matrix C give PCs as:

$$C = \frac{1}{n-1} X^T C_n X \ , \tag{1}$$

where C_n is the $n{\times}n$ centre matrix. The loadings ($L_1 \ldots L_n$) are given, for eigenvectors $e_1 \ldots e_n$ and eigenvalues $\lambda_1 \ldots \lambda_n$ of C such that:

$$L_i = \sqrt{\lambda_i} e_i \ , \text{ for } i = 1,\ldots n, \tag{2}$$

while for PLS-DA, the C is formulated as:

$$C = \frac{1}{(n-1)^2} X^T C_n y y^T C_n X \ , \tag{3}$$

and through an iterative process, we get loading vectors a_i after k iterations, such that:

$$\max_{(a_k, b_k)} cov(X_k a_k, y_k b_k), \tag{4}$$

where b_k is the loading of each label y_k, $X_1 = X$, and X_k and y_k are error matrices after the transformation of previous *k-1* components. The misinterpretation lies in believing that the model should be "better" with higher explained variance, which in the case of PLS-DA, can cause bias. Overall, it can be considered that PLS-DA is a supervised version of PCA, such that PLS is just ordinary PLS regression with a special dummy *y*-variable where PC1 is used to best separate the classes, while PC1 in PCA is used that contains the most variance in a given dataset. For a detailed mathematical explanation, the reader is referred to (Ruiz-Perez et al. 2020).

For extracting relevant features that are most impactful on the classification between top- and bottom-ranked SLRs, we used VIP (Variance Importance in Projection) score to estimate most important features. The metric is employed in PLS-DA to ascertain the significance of each feature in the model, effectively serving as criterion for feature selection. The VIP score quantifies the contribution of features by considering influence across all PCs (or LVs). The VIP scores are particularly useful in high-dimensional datasets in that it offers a succinct yet powerful mesasure for feature selection, allowing for more parsimonious models without sacrificing predictive accuracy. The calculation of VIP score is given as:

$$\mathrm{VIP}_j = \sqrt{\frac{\sum_{k=1}^{A} w_{jk}{}^2 \cdot SSY_k}{\sum_{k=1}^{A} SSY_k}} \ , \tag{5}$$

where A is the number of components (i.e., latent variables) used in PLS-DA, the w_{ij} is the weight of the j-th feature for the k-th component (i.e., latent variable), while $\mathrm{SSY_k}$ is the explained sum of squares for the k-th component of the response variable Y. It is worth noting that w_{jk} should be normalised such that sum of $w_{jk}^2 = 1$ for each component k.

For establishing model significance, the permutation test is performed. The permutation test for the PLS-DA serves as a non-parametric technique to validate PLS-DA model and its statistical significance. Specifically, the labels of the model are permuted multiple times, in this particular case 1000 permutations is performed, and PLS-DA model is fitted to each of these permuted datasets. After conducting permutations, the original model is compared to the distribution of the models derived from permuted data, typically by observing metrics of R^2 and Q^2. The rationale behind the use of permutation test is to establish a null distribution against which the actual model can be compared. If the model's metric significantly outperform metrics from permuted datasets, one can reasonably assume that the model is not fitting to noise but captures a genuine structure within data. The underlying reason why permutation test is used is to check whether model's predictive performance is statistically significant or is a result of overfitting, which is often the case in the use of PLS-DA.

Finally, model significance measures for *Accuracy* is given as:

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)}, \quad (6)$$

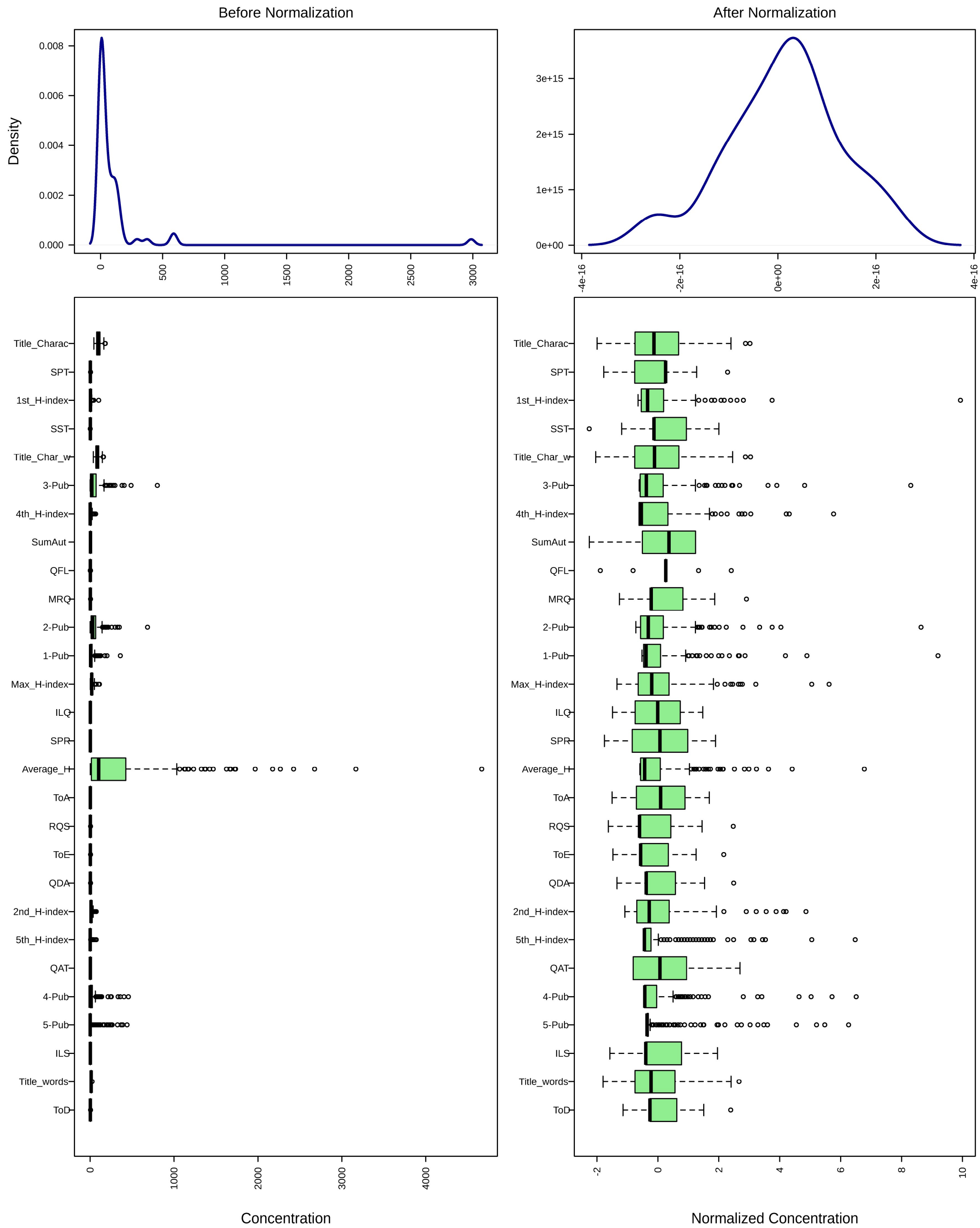
as the number of correct predictions (TP – True positive, TN – True negative) over total number of predictions (TP, TN, FP – False Positive, FN – False Negative). The coefficient of determination R^2 of PLS-DA:

$$\mathrm{R}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \quad (7)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value of the model, $\bar{y}$ is the mean of the observed values. And Q^2 is determined as:

$$\mathrm{Q}^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_{i,-i})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}, \quad (8)$$

here, the $\hat{y}_{i,-i}$ is the predicted value of y_i obtained by a model fit without i-th observation (in this case ten-fold-cross-validation method is performed). The realisation of PLS-DA is performed in MetaboAnalyst 5.0 web-based platform. The PC characteristics used are Intel Core™ i5-10400 CPU at 2.9 GHz, 16GB RAM, nVidia GeForce GT 1030 Graphical Card. In the following, the normalisation of features (Figure A4) and samples (Figure A5) is provided, including the description of the results obtained in PLS-DA.



818

819 Figure A9. Results of features normalisation by sample mean and standard deviation

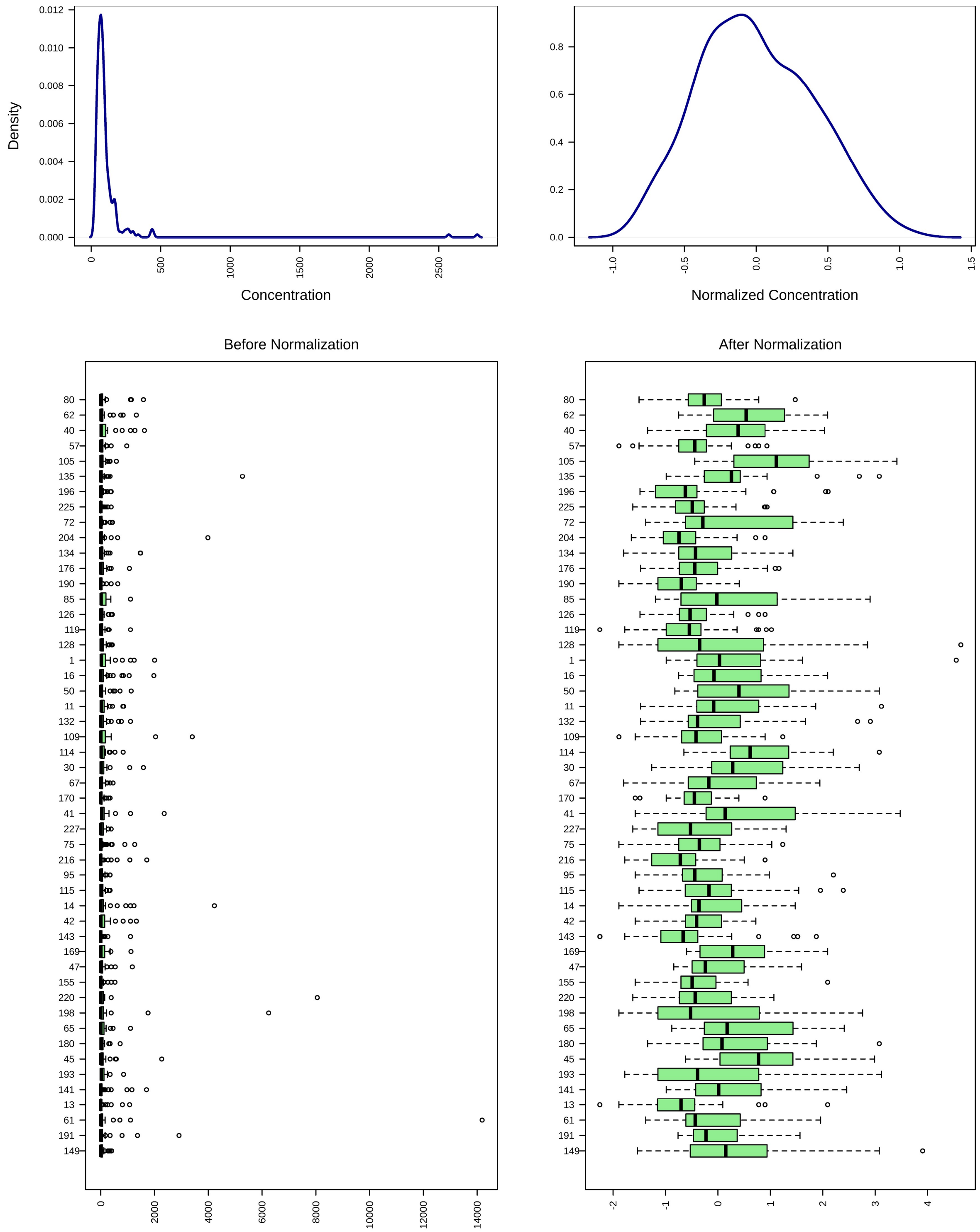


Figure A10. Normalisation of samples by means and standard deviation

Table A2. PLS-DA loadings of components

	C1	C2	C3	C4	C5	C6	C7	C8
Title_Charac_Space	0.0711	0.0281	-0.1304	-0.0102	0.2483	-0.5187	0.6801	-0.3687
Title_Char_withoutspace	0.0759	0.0300	-0.1308	-0.0097	0.2502	-0.5048	0.6769	-0.3638
Title_words	0.0356	0.0138	-0.1178	-0.0120	0.2161	-0.5641	0.6451	-0.3700
University_Rank	-0.0440	-0.1416	0.1269	0.1448	-0.0108	-0.0857	0.1446	0.0772
1st_H-index	0.1609	0.1548	-0.2430	0.0675	0.0515	0.2039	-0.1651	-0.0818
2nd_H-index	0.0637	0.0875	-0.1644	0.1821	0.1726	-0.1776	-0.1375	0.3981
3rd_H-index	0.1790	0.0921	-0.3613	0.1765	0.1902	0.0750	-0.1062	0.2156
4th_H-index	0.1392	0.0911	-0.3028	0.2506	-0.3822	0.1742	0.1490	-0.0772
5th_H-index	0.0581	-0.0771	-0.2413	0.4687	-0.3890	-0.0156	0.0720	-0.1455
Average_H	0.1251	0.1629	-0.2415	-0.1145	0.2221	-0.2814	-0.2316	0.3065
Max_H-index	0.1865	0.1599	-0.4743	0.2656	0.0604	-0.0590	-0.1909	0.1396
1-Pub	0.1434	0.1063	-0.2473	0.0091	0.0169	0.1885	-0.2577	-0.2300
2-Pub	0.0438	-0.0547	-0.2113	0.0463	0.1246	-0.0828	-0.0593	0.6468
3-Pub	0.1217	-0.0531	-0.3132	0.1289	0.1934	-0.0948	-0.1564	0.1003
4-Pub	0.1520	0.0369	-0.3015	0.1245	-0.3412	0.1616	0.1480	-0.1353
5-Pub	0.0198	-0.0782	-0.2415	0.4568	-0.3492	-0.0867	0.1312	-0.0137
SumAut	0.1208	-0.0956	-0.1884	0.4468	-0.2999	0.1818	0.0546	-0.0551
ILQ	0.3504	0.1750	0.2637	0.1434	0.1945	-0.1642	-0.0009	-0.0110
MRQ	0.2447	-0.1786	-0.0593	-0.1300	0.1199	-0.3325	-0.1445	0.1400
QFL	0.0770	-0.3854	-0.1020	-0.0160	0.2820	0.2079	0.0528	-0.0727
RQS	0.3153	-0.1909	0.0015	-0.1981	0.1406	0.0706	-0.0441	-0.2077
QEA	0.3427	-0.1684	0.1011	-0.1421	0.1079	0.1003	0.0768	-0.0700
QDA	0.3242	-0.1421	0.0539	-0.1827	0.0636	0.3212	-0.0312	-0.1026
SST	0.2443	-0.2423	0.1732	0.0193	-0.0316	-0.3159	-0.0364	0.2218
SPR	0.2370	-0.3874	0.1876	0.0177	-0.0418	-0.2112	-0.0585	0.1450
ILS	0.0199	-0.2523	0.1199	0.0407	0.1351	-0.1648	-0.1468	0.0684
TPUB	0.0522	-0.2013	-0.0400	0.1739	0.3374	0.0232	-0.1316	-0.2426
QAT	0.1836	-0.4246	0.1289	0.1247	0.0132	-0.0725	-0.1274	-0.1522
DQA	0.3220	-0.2085	0.1256	0.0475	-0.0377	0.1465	-0.0264	-0.2082
SPT	0.2374	-0.2864	0.2608	-0.0343	-0.0055	-0.1283	0.0537	0.2125
ToA	0.2698	-0.1115	0.2155	-0.1816	-0.2230	-0.0815	0.0841	0.0917
ToE	0.3418	-0.1390	0.1165	-0.1653	-0.2026	0.1366	0.0961	0.0548
ToD	0.3449	-0.1281	0.0798	-0.1399	-0.2218	0.1968	0.0817	0.0150
QATR	0.1248	-0.3826	0.1050	0.1126	-0.0374	0.1465	0.0148	0.0876

Table A3. VIP Scores calculated from coefficients of features in PLS-DA

	C1	C2	C3	C4	C5	C6	C7	C8
ILQ	3.3704	2.9673	2.8823	2.8046	2.7872	2.7761	2.7699	2.7619
QEA	1.5234	1.2882	1.2108	1.1758	1.1772	1.1757	1.1736	1.1702
ToD	1.4838	1.2808	1.1986	1.1886	1.1824	1.1829	1.1803	1.1773
ToE	1.4695	1.2692	1.1855	1.1828	1.1763	1.1762	1.1739	1.1706
DQA	1.4149	1.2048	1.168	1.1333	1.1262	1.1209	1.1199	1.1182
QDA	1.3482	1.1921	1.1138	1.0907	1.0900	1.0895	1.0890	1.0879
1st_H-index	1.245	1.0245	0.9952	0.9932	0.9987	1.0067	1.0050	1.0101
ToA	1.2412	1.0298	0.9643	1.0100	1.0126	1.0079	1.0069	1.0042
Max_H-index	1.0587	0.8302	1.0268	1.0460	1.0413	1.0458	1.0468	1.0439
3rd_H-index	1.0546	0.8265	0.8503	0.9177	0.9267	0.9238	0.9218	0.9214
RQS	0.9443	1.1736	1.1224	1.1162	1.1097	1.1069	1.1080	1.1087
QFL	0.9424	1.5979	1.4921	1.4589	1.4688	1.4635	1.4607	1.4585
SST	0.8223	0.8895	0.8805	0.8563	0.8576	0.8613	0.8608	0.8623
SPT	0.7960	0.8651	0.9420	0.9150	0.9093	0.9051	0.9043	0.9024
4th_H-index	0.7643	0.6009	0.7002	0.7124	0.7251	0.7324	0.7311	0.7305
Average_H	0.7439	0.5831	0.8378	0.8231	0.8185	0.8285	0.8267	0.8335
1-Pub	0.7397	0.5880	0.7663	0.7430	0.7390	0.7381	0.7483	0.7473
2nd_H-index	0.6489	0.5818	0.54457	0.6122	0.6107	0.6130	0.6119	0.6167
4-Pub	0.5255	0.5521	0.75727	0.7379	0.7418	0.7530	0.7527	0.7507
MRQ	0.4965	1.0243	1.0127	1.0042	0.9982	1.0037	1.0021	1.0062
ILS	0.4837	0.7189	0.7504	0.7324	0.7284	0.7291	0.7275	0.7309
SumAut	0.4533	0.4377	0.4208	0.6465	0.6605	0.6599	0.6594	0.6615
Title_Char_without_sp	0.3649	0.2960	0.3463	0.3421	0.3511	0.3640	0.3913	0.4187
University_Rank	0.3569	0.2979	0.5326	0.5509	0.5476	0.5452	0.5447	0.5444
Title_Charac_Space	0.3271	0.2714	0.3300	0.3257	0.3343	0.3491	0.3788	0.4062
SPR	0.3192	1.111	1.0963	1.0632	1.0587	1.0552	1.0541	1.0552
QATR	0.2458	1.0101	1.0391	1.0309	1.0264	1.0285	1.0265	1.0240
2-Pub	0.1914	0.4790	0.5298	0.5377	0.5440	0.5421	0.5441	0.5525
5-Pub	0.1815	0.3340	0.3126	0.5767	0.6010	0.6016	0.6070	0.6073
3-Pub	0.1500	0.5819	0.6854	0.7062	0.7043	0.7066	0.7056	0.7048
TPUB	0.1362	0.4612	0.5307	0.6517	0.6682	0.6693	0.6715	0.6700
QAT	0.1156	1.2115	1.1886	1.1561	1.1522	1.1543	1.1539	1.1510
Title_words	0.0643	0.1542	0.2601	0.2539	0.2575	0.2832	0.3234	0.3445
5th_H-index	0.0052	0.3507	0.3368	0.5411	0.5994	0.5967	0.5954	0.5950

Table A4. Results of classification and permutation test statistic

	C1	C2	C3	C4	C5
Accuracy	0.6551	0.7233	0.7454	0.7449	0.7403
R^2	0.2203	0.3594	0.4122	0.4388	0.4444
Q^2	0.1676	0.2448	0.2638	0.2719	0.2754

Table A5. Descriptive statistics of RCE items of selected SLRs of the first sample (n = 170).

	Mean	Std. Dev.	$95\%CI_{Upper\text{-}Mean}$	$95\%CI_{Lower\text{-}Mean}$	Std. Error	$95\%CI_{Upper\text{-}Stdev}$	$95\%CI_{Lower\text{-}Stdev}$
ILQ	3.135	1.254	3.324	2.947	0.096	1.334	1.156
MRQ	2.612	1.089	2.775	2.448	0.084	1.195	0.981
QFL	1.947	1.016	2.100	1.794	0.078	1.151	0.861
RQS	2.500	1.004	2.651	2.349	0.077	1.088	0.905
QEA	2.635	0.989	2.784	2.487	0.076	1.081	0.890
QDA	2.300	0.99	2.449	2.151	0.076	1.083	0.889
SST	2.965	1.235	3.150	2.779	0.095	1.329	1.140
SPR	2.647	1.051	2.805	2.489	0.081	1.134	0.949
ILS	2.506	0.999	2.656	2.356	0.077	1.083	0.904
TPUB	2.082	0.951	2.225	1.939	0.073	1.061	0.833
QAT	2.006	1.128	2.175	1.836	0.087	1.238	0.991

DQA	1.359	0.675	1.460	1.257	0.052	0.762	0.568
SPT	2.741	1.193	2.921	2.562	0.092	1.285	1.089
ToA	2.735	1.276	2.927	2.544	0.098	1.367	1.167
ToE	2.429	1.025	2.584	2.275	0.079	1.125	0.918
ToD	2.176	1.122	2.345	2.008	0.086	1.235	1.006
QATR	1.671	0.972	1.817	1.525	0.086	1.235	1.006

References

Adler, R., Ewing, J., & Taylor, P. (2009). Citation Statistics. *Statistical Science*, *24*(1). https://doi.org/10.1214/09-STS285

Blümel, C., & Schniedermann, A. (2020). Studying review articles in scientometrics and beyond: a research agenda. *Scientometrics*, *124*(1), 711–728. https://doi.org/10.1007/s11192-020-03431-7

Booth, A. (2006). Clear and present questions: formulating questions for evidence based practice. *Library Hi Tech*, *24*(3), 355–368. https://doi.org/10.1108/07378830610692127

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic Approaches to a Successful Literature Review*. (A. Booth, A. Sutton, & D. Papaioannou, Eds.)*SAGE Publications Ltd* (3rd ed., Vol. 34). SAGE Publications Ltd.

Bornmann, L., & Daniel, H. (2008). What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation*, *64*(1), 45–80. https://doi.org/10.1108/00220410810844150

Bornmann, L., & Leydesdorff, L. (2017). Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data. *Journal of Informetrics*, *11*(1), 164–175. https://doi.org/10.1016/j.joi.2016.12.001

Bramer, W. M., Giustini, D., & Kramer, B. M. R. (2016). Comparing the coverage, recall, and precision of searches for 120 systematic reviews in Embase, MEDLINE, and Google Scholar: a prospective study. *Systematic Reviews*, *5*(1), 39. https://doi.org/10.1186/s13643-016-0215-7

Briganti, G., Fried, E. I., & Linkowski, P. (2019). Network analysis of Contingencies of Self-Worth Scale in 680 university students. *Psychiatry Research*, *272*(December 2018), 252–257. https://doi.org/10.1016/j.psychres.2018.12.080

CASP. (2018). Critical Appraisal Skills Programme Checklist: 10 questions to help you make sense of a Systematic Review. https://casp-uk.net/

Chawla, D. S. (2020). Science is getting harder to read. *Nature index*. https://www.nature.com/nature-index/news-blog/science-research-papers-getting-harder-to-read-acronyms-jargon

Chen, M. C., Chen, S. H., Cheng, C. D., Chung, C. H., Mau, L. P., Sung, C. E., et al. (2023). Mapping out the bibliometric characteristics of classic articles published in a Taiwanese academic journal in dentistry: A scopus-based analysis. *Journal of Dental Sciences*, (xxxx). https://doi.org/10.1016/j.jds.2023.03.015

Cheng, K. L., Dodson, T. B., Egbert, M. A., & Susarla, S. M. (2017). Which Factors Affect Citation Rates in the Oral and Maxillofacial Surgery Literature? *Journal of Oral and Maxillofacial Surgery*, *75*(7), 1313–1318. https://doi.org/10.1016/j.joms.2017.03.043

Coombes, P. (2023). Systematic Review Research in Marketing Scholarship: Optimizing Rigor. *International Journal of Market Research*. https://doi.org/10.1177/14707853231184729

de Almeida, E. C. E., & Guimarães, J. A. (2013). Brazil's growing production of scientific articles—how are we doing with review articles and other qualitative indicators? *Scientometrics*, *97*(2), 287–315. https://doi.org/10.1007/s11192-013-0967-y

Dybå, T., & Dingsøyr, T. (2008). Strength of Evidence in Systematic Reviews in Software Engineering. *ESEM'08: Proceedings of the 2008 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*, (January), 178–187. https://doi.org/10.1145/1414004.1414034

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, *50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Eysenbach, G. (2011). Can Tweets Predict Citations? Metrics of Social Impact Based on Twitter and Correlation with Traditional Metrics of Scientific Impact. *Journal of Medical Internet Research*, *13*(4), e123. https://doi.org/10.2196/jmir.2012

Faggion, C. M., Bakas, N. P., & Wasiak, J. (2017). A survey of prevalence of narrative and systematic reviews in five major medical journals. *BMC Medical Research Methodology*, *17*(1), 176. https://doi.org/10.1186/s12874-017-0453-y

Fitzgerald, B., Alan R. Dennis, A. R., An, J., Tsutsui, S., & Muchala, R. C. (2019). Information Systems Research: Thinking Outside the Basket and Beyond the Journal. *Communications of the Association for Information Systems*, 110–133. https://doi.org/10.17705/1CAIS.04507

Foygel, R., & Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, & A. Culotta (Eds.), *Advances in Neural Information Processing Systems* (Vol. 23). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2010/file/072b030ba126b2f4b2374f342be9ed44-Paper.pdf

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience*, *21*(11), 1129–1164. https://doi.org/10.1002/spe.4380211102

Furley, P., & Goldschmied, N. (2021). Systematic vs. Narrative Reviews in Sport and Exercise Psychology: Is Either Approach Superior to the Other? *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.685082

Garcia-Doval, I., van Zuuren, E. J., Bath-Hextall, F., & Ingram, J. R. (2017). Systematic reviews: let's keep them trustworthy. *British Journal of Dermatology*, *177*(4), 888–889. https://doi.org/10.1111/bjd.15826

Ghidalia, S., Narsis Labbani, O., Bertaux, A., & Christophe, N. (2023). Mixed artificial reasoning, closer to human? (Under Review). *Artificial Intelligence Review*, 1–50. https://doi.org/10.21203/rs.3.rs-1881512/v1

Glänzel, W., Debackere, K., Thijs, B., & Schubert, A. (2006). A concise review on the role of author self-citations in information science, bibliometrics and science policy. *Scientometrics*, *67*(2), 263–277. https://doi.org/10.1007/s11192-006-0098-9

Grover, V., Raman, R., & Stubblefield, A. (2014). What Affects Citation Counts in MIS Research Articles? An Empirical Investigation. *Communications of the Association for Information Systems*, *34*. https://doi.org/10.17705/1CAIS.03474

Han, Y., Chong, W. K., & Li, D. (2020). A systematic literature review of the capabilities and performance metrics of supply chain resilience. *International Journal of Production Research*, 4541–4566. https://doi.org/10.1080/00207543.2020.1785034

Ho, M. H.-C., Liu, J. S., & Chang, K. C.-T. (2017). To include or not: the role of review papers in citation-based analysis. *Scientometrics*, *110*(1), 65–76. https://doi.org/10.1007/s11192-016-2158-0

Ho, Y.-S., & Shekofteh, M. (2021). Performance of highly cited multiple sclerosis publications in the Science Citation Index expanded: A scientometric analysis. *Multiple Sclerosis and Related Disorders*, *54*, 103112. https://doi.org/10.1016/j.msard.2021.103112

Ioannidis, J. P. A. (2016, September 1). The Mass Production of Redundant, Misleading, and Conflicted Systematic Reviews and Meta-analyses. *Milbank Quarterly*. Blackwell Publishing Inc. https://doi.org/10.1111/1468-0009.12210

JASP. (2018). How to Perform a Network Analysis in JASP. https://jasp-stats.org/2018/03/20/perform-network-analysis-jasp/. Accessed 1 May 2023

Joanna Briggs Institute, & JBI. (2022). Checklist for systematic reviews and research syntheses. *Joanna Briggs Institute*.

Jokic, M., & Ball, R. (2006). *Qualität und Quantität wissenschaftlicher Veröffentlichungen: Bibliometrische Aspekte der Wissenschaftskommunikation* (1st ed., Vol. 15). Jülich: Forschungszentrum Jülich GmbH.

Judge, T. A., Cable, D. M., Colbert, A. E., & Rynes, S. L. (2007). What Causes a Management Article to Be Cited: Article, Author, or Journal? *The Academy of Management Journal*, *50*(3), 491–506.

Kalantari, E., Kouchaki, S., Miaskowski, C., Kober, K., & Barnaghi, P. (2022). Network analysis to identify symptoms clusters and temporal interconnections in oncology patients. *Scientific Reports*, *12*(1), 17052. https://doi.org/10.1038/s41598-022-21140-4

Karunananthan, S., Welch, V. A., Tugwell, P., & Cuervo, L. G. (2021). When is systematic review replication useful, and when is it wasteful? *Revista Panamericana de Salud Pública*, *45*, 1. https://doi.org/10.26633/RPSP.2021.11

Kitchenham, B., Charters, S., Budgen, D., Brereton, P., Turner, M., Linkman, S., et al. (2007). *Guideliness for Performing Systematic Literature Reviews in Software Engineering*. *EBSE Technical Report* (Vol. 2.3). Durham, UK.

Knottnerus, J. A., & Knottnerus, B. J. (2009). Let's make the studies within systematic reviews count. *The Lancet*, *373*(9675), 1605. https://doi.org/10.1016/S0140-6736(09)60906-5

Kousha, K., & Thelwall, M. (2023). Factors associating with or predicting more cited or higher quality journal articles: An Annual Review of Information Science and Technology (ARIST) paper. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24810

Kung, J., Chiappelli, F., Cajulis, O. O., Avezova, R., Kossan, G., Chew, L., & Maida, C. A. (2010). From Systematic Reviews to Clinical Recommendations for Evidence- Based Health Care: Validation of Revised Assessment of Multiple Systematic Reviews (R-AMSTAR) for Grading of Clinical Relevance. *The Open Dentistry Journal*, *4*(2), 84–91. https://doi.org/10.2174/1874210601004020084

Leme, D. E. da C., Alves, E. V. da C., Lemos, V. do C. O., & Fattori, A. (2020). Network Analysis: a Multivariate Statistical Approach for Health Science Research. *Geriatrics, Gerontology and Aging*, *14*(1), 43–51. https://doi.org/10.5327/z2447-212320201900073

Liskiewicz, T., Liskiewicz, G., & Paczesny, J. (2021). Factors affecting the citations of papers in tribology journals. *Scientometrics*, *126*(4), 3321–3336. https://doi.org/10.1007/s11192-021-03870-w

Liu, D., Jin, J., Tian, J., & Yang, K. (2015). Quality assessment and factor analysis of systematic reviews and meta-analyses of endoscopic ultrasound diagnosis. *PLoS ONE*, *10*(4), 1–13. https://doi.org/10.1371/journal.pone.0120911

Liu, J. S., & Kuan, C. (2016). A new approach for main path analysis: Decay in knowledge diffusion. *Journal of the Association for Information Science and Technology*, *67*(2), 465–476. https://doi.org/10.1002/asi.23384

Mäntylä, M., & Garousi, V. (2019). Citations in Software Engineering -- Paper-related, Journal-related, and Author-related Factors.

McColl, E. (2022). Systematic reviews of reviews of reviews. *British Dental Journal*, *233*(8), 586–586. https://doi.org/10.1038/s41415-022-5159-9

Memon, J., Sami, M., Khan, R. A., & Uddin, M. (2020). Handwritten Optical Character Recognition (OCR): A Comprehensive Systematic Literature Review (SLR). *IEEE Access*, *8*, 142642–142668. https://doi.org/10.1109/ACCESS.2020.3012542

Minh, L. H. N., Le, H.-H., Tawfik, G. M., Makram, O. M., Tieu, T., Tai, L. L. T., et al. (2023). Factors associated with successful publication for systematic review protocol registration: an analysis of 397 registered protocols. *Systematic Reviews*, *12*(1), 93. https://doi.org/10.1186/s13643-023-02210-8

Misra, D. P., & Agarwal, V. (2018). Systematic Reviews: Challenges for Their Justification, Related Comprehensive Searches, and Implications. *Journal of Korean Medical Science*, *33*(12). https://doi.org/10.3346/jkms.2018.33.e92

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., et al. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1. https://doi.org/10.1186/2046-4053-4-1

Montori, V. M., Wilczynski, N. L., Morgan, D., Haynes, B., Eady, A., Marks, S., et al. (2003). Systematic reviews: A cross-sectional study of location and citation counts. *BMC Medicine*, *1*, 1–7. https://doi.org/10.1186/1741-7015-1-2

Mulrow, C. D. (1994). Rationale for systematic reviews. *BMJ*, *309*(3), 597–599. https://doi.org/10.1136/bmj.309.6954.597

Munn, Z., Stern, C., Aromataris, E., Lockwood, C., & Jordan, Z. (2018). What kind of systematic review should I conduct? A proposed typology and guidance for systematic reviewers in the medical and health sciences. *BMC Medical Research Methodology*, *18*(1), 5. https://doi.org/10.1186/s12874-017-0468-4

Nishikawa-Pacher, A. (2022). Research Questions with PICO: A Universal Mnemonic. *Publications*, *10*(3), 1–10. https://doi.org/10.3390/publications10030021

Oelen, A., Jaradeh, M. Y., Stocker, M., & Auer, S. (2020). Generate FAIR Literature Surveys with Scholarly Knowledge Graphs. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020* (pp. 97–106). New York, NY, USA: ACM. https://doi.org/10.1145/3383583.3398520

Orošnjak, M., Brkljač, N., Šević, D., Čavić, M., Oros, D., & Penčić, M. (2023). From predictive to energy-based maintenance paradigm: Achieving cleaner production through functional-productiveness. *Journal of Cleaner Production*, *408*, 137177. https://doi.org/10.1016/j.jclepro.2023.137177

Page, M. J., & Moher, D. (2016). Mass production of systematic reviews and meta-analyses: an exercise in mega-silliness? *Milbank Quarterly*, *94*(3), 515–519. http://www.prisma-statement.org/

Paré, G., Tate, M., Johnstone, D., & Kitsiou, S. (2016). Contextualizing the twin concepts of systematicity and transparency in information systems literature reviews. *European Journal of Information Systems*, *25*(6), 493–508. https://doi.org/10.1057/s41303-016-0020-3

Patnode, C. D., Henderson, J. T., Thompson, J. H., Senger, C. A., Fortmann, S. P., & Whitlock, E. P. (2015). Behavioral Counseling and Pharmacotherapy Interventions for Tobacco Cessation in Adults, Including Pregnant Women: A Review of Reviews for the U.S. Preventive Services Task Force. *Annals of Internal Medicine*, *163*(8), 608–621. https://doi.org/10.7326/M15-0171

Perk, S., Teymour, F., & Cinar, A. (2011). Adaptive agent-based system for process fault diagnosis. *Industrial and Engineering Chemistry Research*, *50*(15), 9138–9155. https://doi.org/10.1021/ie102058d

Roberts, I., & Ker, K. (2015). How systematic reviews cause research waste. *The Lancet*, *386*(10003), 1536. https://doi.org/10.1016/S0140-6736(15)00489-4

Robinaugh, D. J., Millner, A. J., & McNally, R. J. (2016). Identifying highly influential nodes in the complicated grief network. *Journal of Abnormal Psychology*, *125*(6), 747–757. https://doi.org/10.1037/abn0000181

Rousseau, S., Catalano, G., & Daraio, C. (2021). Can we estimate a monetary value of scientific publications? *Research Policy*, *50*(1), 104116. https://doi.org/10.1016/J.RESPOL.2020.104116

Royle, P., Kandala, N.-B., Barnard, K., & Waugh, N. (2013). Bibliometrics of systematic reviews: analysis of citation rates and journal impact factors. *Systematic Reviews*, *2*(1), 74. https://doi.org/10.1186/2046-4053-2-74

Ruiz-Perez, D., Guan, H., Madhivanan, P., Mathee, K., & Narasimhan, G. (2020). So you think you can PLS-DA? *BMC Bioinformatics*, *21*(Suppl 1), 1–10. https://doi.org/10.1186/s12859-019-3310-7

Schniedermann, A. (2021). A comparison of systematic reviews and guideline-based systematic reviews in medical studies. *Scientometrics*, *126*(12), 9829–9846. https://doi.org/10.1007/s11192-021-04199-0

Shea, B. J., Reeves, B. C., Wells, G., Thuku, M., Hamel, C., Moran, J., et al. (2017). AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *BMJ (Online)*, *358*, 1–9. https://doi.org/10.1136/bmj.j4008

So, M., Kim, J., Choi, S., & Park, H. W. (2015). Factors affecting citation networks in science and technology: Focused on non-quality factors. *Quality and Quantity*, *49*(4), 1513–1530. https://doi.org/10.1007/s11135-014-0110-z

Soheili, F., Khasseh, A. A., Mokhtari, H., & Sadeghi, M. (2022). Factors Affecting the Number of Citations: A Mixed Method Study. *Journal of Scientometric Research*, *11*(1), 01–14. https://doi.org/10.5530/jscires.11.1.1

Tahamtan, I., Safipour Afshar, A., & Ahamdzadeh, K. (2016). Factors affecting number of citations: a comprehensive review of the literature. *Scientometrics*, *107*(3), 1195–1225. https://doi.org/10.1007/s11192-016-1889-2

Templier, M., & Paré, G. (2018). Transparency in literature reviews: an assessment of reporting practices across review types and genres in top IS journals. *European Journal of Information Systems*, *27*(5), 503–550. https://doi.org/10.1080/0960085X.2017.1398880

Triguero, I., García-Gil, D., Maillo, J., Luengo, J., García, S., & Herrera, F. (2019). Transforming big data into smart data: An insight on the use of the k-nearest neighbors algorithm to obtain quality data. *WIREs Data Mining and Knowledge Discovery*, *9*(2). https://doi.org/10.1002/widm.1289

Uthman, O. A., Okwundu, C. I., Wiysonge, C. S., Young, T., & Clarke, A. (2013). Citation Classics in Systematic Reviews and Meta-Analyses: Who Wrote the Top 100 Most Cited Articles? *PLoS ONE*, *8*(10), 1–11. https://doi.org/10.1371/journal.pone.0078517

Uttley, L., Quintana, D. S., Montgomery, P., Carroll, C., Page, M. J., Falzon, L., et al. (2023). The problems with systematic reviews: a living systematic review. *Journal of Clinical Epidemiology*, *156*, 30–41. https://doi.org/10.1016/j.jclinepi.2023.01.011

van der Braak, K., Ghannad, M., Orelio, C., Heus, P., Damen, J. A. A., Spijker, R., et al. (2022). The score after 10 years of registration of systematic review protocols. *Systematic Reviews*, *11*(1), 191. https://doi.org/10.1186/s13643-022-02053-9

Vanclay, J. K. (2013). Factors affecting citation rates in environmental science. *Journal of Informetrics*, *7*(2), 265–271. https://doi.org/10.1016/j.joi.2012.11.009

Wagner, G., Prester, J., Roche, M., Benlian, A., & Schryen, G. (2016). Factors Affecting the Scientific Impact of Literature Reviews: A Scientometric Study. In P. J. Agerfalk, N. Levina, & S. S. Kien (Eds.), *Thirty Seventh International Conference on Information Systems* (pp. 1–24). Dublin: Association for Information Systems.

Wagner, G., Prester, J., Roche, M. P., Schryen, G., Benlian, A., Paré, G., & Templier, M. (2021). Which factors affect the scientific impact of review papers in IS research? A scientometric study. *Information & Management*, *58*(3), 103427. https://doi.org/10.1016/j.im.2021.103427

Wang, M., Wang, Z., & Chen, G. (2019). Which can better predict the future success of articles? Bibliometric indices or alternative metrics. *Scientometrics*, *119*(3), 1575–1595. https://doi.org/10.1007/s11192-019-03052-9

Wang, M., Yu, G., An, S., & Yu, D. (2012). Discovery of factors influencing citation impact based on a soft fuzzy rough set model. *Scientometrics*, *93*(3), 635–644. https://doi.org/10.1007/s11192-012-0766-x

Wang, M., Yu, G., & Yu, D. (2011). Mining typical features for highly cited papers. *Scientometrics*, *87*(3), 695–706. https://doi.org/10.1007/s11192-011-0366-1

Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, *26*(2), 13–23.

Westerhuis, J. A., Hoefsloot, H. C. J., Smit, S., Vis, D. J., Smilde, A. K., Velzen, E. J. J., et al. (2008). Assessment of PLSDA cross validation. *Metabolomics*, *4*(1), 81–89. https://doi.org/10.1007/s11306-007-0099-6

Whiting, P., Savović, J., Higgins, J. P. T., Caldwell, D. M., Reeves, B. C., Shea, B., et al. (2016). ROBIS: A new tool to assess risk of bias in systematic reviews was developed. *Journal of Clinical Epidemiology*, *69*, 225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005

Wormald, R., & Evans, J. (2018). What Makes Systematic Reviews Systematic and Why are They the Highest Level of Evidence? *Ophthalmic Epidemiology*, *25*(1), 27–30. https://doi.org/10.1080/09286586.2017.1337913

Xie, J., Gong, K., Cheng, Y., & Ke, Q. (2019). The correlation between paper length and citations: a meta-analysis. *Scientometrics*, *118*(3), 763–786. https://doi.org/10.1007/s11192-019-03015-0

Xie, J., Gong, K., Li, J., Ke, Q., Kang, H., & Cheng, Y. (2019). A probe into 66 factors which are possibly associated with the number of citations an article received. *Scientometrics*, *119*(3), 1429–1454. https://doi.org/10.1007/s11192-019-03094-z

Yan, Z., Kuang, T.-H., & Yao, Y. (2017). Multivariate fault isolation of batch processes via variable selection in partial least squares discriminant analysis. *ISA Transactions*, *70*, 389–399. https://doi.org/10.1016/j.isatra.2017.06.014

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (2014). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics*, *101*(2), 1233–1252. https://doi.org/10.1007/s11192-014-1279-6

Yuan, Y., & Hunt, R. H. (2009). Systematic Reviews: The Good, the Bad and the Ugly. *The American Journal of Gastroenterology*, *104*(5), 1086–1092. https://doi.org/10.1038/ajg.2009.118

Zhou, Y., Zhang, H., Huang, X., Yang, S., Babar, M. A., & Tang, H. (2015). Quality assessment of systematic reviews in software engineering: A tertiary study. *ACM International Conference Proceeding Series*, *27-29-Apri*. https://doi.org/10.1145/2745802.2745815

Zong, Q., Xie, Y., & Liang, J. (2020). Does open peer review improve citation count? Evidence from a propensity score matching analysis of PeerJ. *Scientometrics*, *125*(1), 607–623. https://doi.org/10.1007/s11192-020-03545-y