DSI DATA SCIENCE AND INFORMETRICS
VOL.2, NO.1, FEB. 2022

RESEARCH ARTICLES

Research topics in the international research collaboration measurement domain

Ba Xuan Nguyen[a,d], Jesse David Dinneen[b], Markus Luczak–Roesch[a,c]

a. School of Information Management, Victoria University of Wellington, Wellington, New Zealand
b. Berlin School of Library and Information Science, Humboldt–Universität zu Berlin, Berlin, Germany
c. Te Pūnaha Matatini, Aotearoa New Zealand´s Centre of Research Excellence for Complex Systems, New Zealand
d. Posts and Telecommunications Institute of Technology, Ho Chi Minh City, Vietnam

ABSTRACT

Here we report on a study that combined a scoping review with co-occurrence analysis to assess the current state of publications and research topics in the area of international research collaboration measurement (IRCM). Our study found that IRCM studies have been published in source titles of diverse subject areas and that there are two core research topics that have been commonly discussed across different subject areas in the IRCM domain - scientific productivity measurement and scientific impact measurement. The appearance of papers about IRCM in venues beyond those concerned only with bibliometric measures indicates the broad importance of IRCM for diverse research subjects, and that studies of IRC within particular fields should draw on diverse venues to provide a holistic and interdisciplinary picture of IRCM.

KEYWORDS

International research collaboration; Measurement; Research topic; Scoping review; Co-occurrence analysis

1 Introduction

International research collaboration (IRC), which refers to scientific activities between individuals from different countries, has become a promising topic in recent years. Because IRC produces benefits for countries, such as improved scientific impact and global competitiveness in innovation (Sharma & Thomas, 2008), most governments are interested in supporting IRC with appropriate research strategies (Katz & Martin, 1997). As a result, measuring IRC activities is a central area of bibliometrics and informetrics research (e.g. to assess the outcomes of supporting IRC), and IRC measurement (IRCM) in particular has received interest from researchers (Beaver & Rosen, 1978, 1979) and policy makers (Abramo, 2011). The collective work and findings pertaining to IRCM have, however, not yet been studied comprehensively (Chen et al., 2019). Notably, the interdisciplinary nature of IRCM studies suggests they can be published in scientific journals of different fields or journals covering different themes, depending on the needs and interests of different audiences. Furthermore, as IRCM is an emerging research domain, there are multiple benefits to analysing its publications and

subject classifications. For example, identifying the status of IRCM publications and venues has practical benefits for bibliometric researchers, such as indicating to researchers new to IRCM where to find literature or where they can reach a desired audience. Finally, characterising emerging research topics contributes to research foundations (e.g. making sense of the collective methods or findings). Another example is that identifying research topics in the IRCM domain also benefits policy-makers because they need to promote promising research topics (Xu et al., 2019).

2 Methodology and Data

We performed a scoping review, which is an appropriate method to create an initial, high-level overview into the nature and extent of research evidence in a research area (Grant & Booth, 2009). This is a necessary first step to guide further in-depth literature reviews. The present study applied the methodological framework proposed by Arksey and O'Malley (2005) for scoping reviews.

In addition, subject classification (i.e., the organisation of publications into categories of research topics) is necessary to detect emerging research trends (Jeangirard, 2021) of IRCM domain. Among subject classification techniques, co-occurrence analysis is a technique that extracts research themes from texts and detects the themes relationships directly from content, without relying on any a priori definition of these research themes (Sedighi, 2016). This technique was applied in this study to construct a visual map of distinct groups of closely connected publications in the IRCM domain so that different groups of research topics could be easily identified.

The scoping review and the co-occurrence analysis were combined in this study in seven steps (the first five steps for the scoping review and the last two steps for co-occurrence analysis) as follows:

Step 1 - Defining the research questions: this study aims to answer the following research questions (RQs):

RQ1. What are the common venues for discussing IRCM, and what are their subject areas (i.e. fields)?

RQ2. What are the primary research topics (i.e. specific research areas) in the IRCM domain?

Step 2 - Identifying relevant studies: For the purpose of addressing the two research questions stated above, the key concepts are "international research collaboration" and "measurement". The Web of Science and Scopus have been considered the leading databases for scientometric studies, of which Scopus covers the larger number of journals (Mongeon & Paul-Hus, 2016). Therefore, Scopus was chosen for this search for its better coverage, but the limitations associated with this choice are discussed below. Two keywords were initially used together in the literature search —"international research collaboration" and "measurement" — and their different alternatives (e.g., "international scientific collaboration" and "measure") that were found in the results were consistently added to the collection of keywords and subsequently searched. In the final iteration of searching, 151 publications were identified in this step.

Step 3 - Study Selection: Given the goal of the present study, our inclusion criteria were that papers be in English and be clearly relevant to IRCM. Therefore, from the 151 results, only English-language papers were reviewed (n=140), and irrelevant papers (e.g. about other kinds of industrial or medical measurement) were excluded (n=75). 65 publications were thus kept for analysis (the full bibliographic list is available online)[1].

Step 4 - Charting the data: relevant bibliographic details were extracted for analysis, in-

1 https://github.com/baxuan/IRCM/tree/master/Research–Topics

cluding author names, title, year of publication, source title (i.e. publication title), abstracts, and keywords.

Step 5 - Collating, summarising, and reporting the results: The extracted information was used to identify the common venues for publishing IRCM studies, which were then categorised by applying the journal classification from the SCImago Journal and Country Rank (SJR) database, which is a publicly available portal based on the Scopus database. The classification of journals from SCImago Journal Rank was used to identify the subject areas of each source title (presented in Table 1).

Step 6 - Analysing keyword co-occurrence from IRCM publications' keywords: A keyword co-occurrence map was created using VOSviewer, a popular software tool that applies a clustering algorithm for subject classification (Gómez-Núñez et al., 2014). Primary research topics were identified by looking for themes in the keyword co-occurrence map (i.e., publications' keywords grouped into the same cluster were more likely to reflect the same subjects). We used the threshold level (i.e., the number of publications in which a keyword occurs at least once) at 8 because this threshold could show clear clusters of the IRCM domain. Of the 606 keywords from 65 relevant publications, 17 met the threshold level 8. These 17 publications' keywords identify the primary research topics in the IRCM domains.

Step 7 - Analysing term co-occurrence from IRCM publications' titles and their abstracts: Keyword co-occurrence analysis was applied to the text of the titles and abstracts of the relevant IRCM publications to identify the common terms. The present study set the threshold level (minimum number of occurrences of a term) at 7. Of the 2112 terms from 65 relevant publications, 57 terms met the threshold of 7 (i.e., the number of publications in which a term occurs at least once). To keep the map legible (i.e. to avoid many overlapping terms) we applied the VOSviewer default selection to display only 60% of the most relevant terms (i. e. 34 terms).

The steps and relevant data implemented in this study are described in the following process:

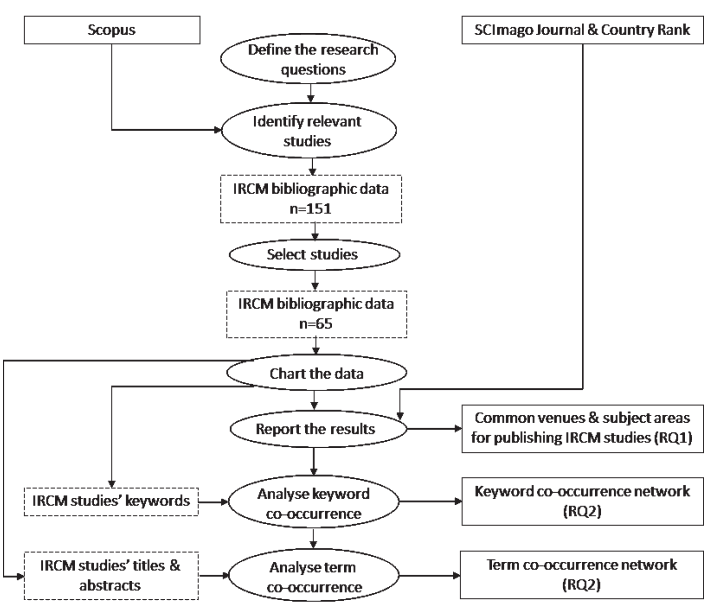

Figure 1 A summary of the data collection and processes of the scoping review and co-occurrence analysis

3 Results and Discussion

For RQ1 ("What are the common venues for discussing IRCM, and what are their subject areas?"), there were 56 source titles initially identified. The most commonly chosen source titles publishing IRCM discussions are represented in Table 1. Other details (the subject areas of each source title, year range and number of the IRCM publications identified with each source title) are also represented in this table. The present study showed that Scientometrics is the most common venue for IRCM publications. The study also showed that the ICRM publications have a wide range of readers and the most common subject areas are social sciences, computer science, and decision sciences.

Table 1 The most chosen source titles for publishing IRCM discussions (only those having more than one IRCM publications were represented in this table.)

Source Title	SCImago subject areas	Number of publications
Scientometrics	Computer Science; Social Sciences	4
ISSI Conference Proceedings	Business, Management and Accounting; Computer Science; Decision Sciences; Mathematics	2
Global Knowledge Memory And Communication	Social Sciences	2
International Journal Of Environmental Research And Public Health	Environmental Science Medicine	2
Journal Of Clinical Gastroenterology	Medicine	2
Library Philosophy And Practice	Arts and Humanities; Social Sciences	2
Research Policy	Business, Management and Accounting; Decision Sciences; Engineering	2

For the topic of IRCM in particular, this study showed that the most common subject areas for IRCM publications are Social Sciences (i.e., some IRCM studies are classified into the Library and Information Sciences, which is a category of Social Sciences), Computer Science, Decision Sciences, Medicine, and Business, Management and Accounting. While Social Sciences and Computer Science are suitable subject areas for bibliometric researchers, the Decision Sciences area is familiar to policy makers. Therefore, IRCM studies have been published in a wide range of appropriate journals for their target readers, including both researchers and policy makers. One implication is that the authors of future IRCM studies should choose the journals with focus and high rank on Social Sciences and Computer Science (or Decision Sciences) if they want bibliometric researchers (or policy makers) to read their results. Other source titles categorised as multidisciplinary journals are also good choices to meet the needs of both readers in particular subject areas (e.g., 'Medicine') and bibliometric researchers and policy makers. The diverse journals in Table 1 also implies that the contents of many IRCM studies have been limited to specific disciplines (i.e., 'subject areas'). As different disciplines have their own focuses, the IRCM studies in each discipline lose sight of what is mentioned in other disciplines. The benefits of having an overview of the IRCM picture, or thinking across boundaries of academic disciplines, may suggest the need for more interdis-

ciplinary studies of IRCM.

For RQ2 ("What are the primary research topics in the IRCM domain?"), two clusters of publications' keywords were identified. The results obtained using the full counting method or the fractional counting method are similar. Figure 2 displayed the publications' keywords in these different clusters with different colours.

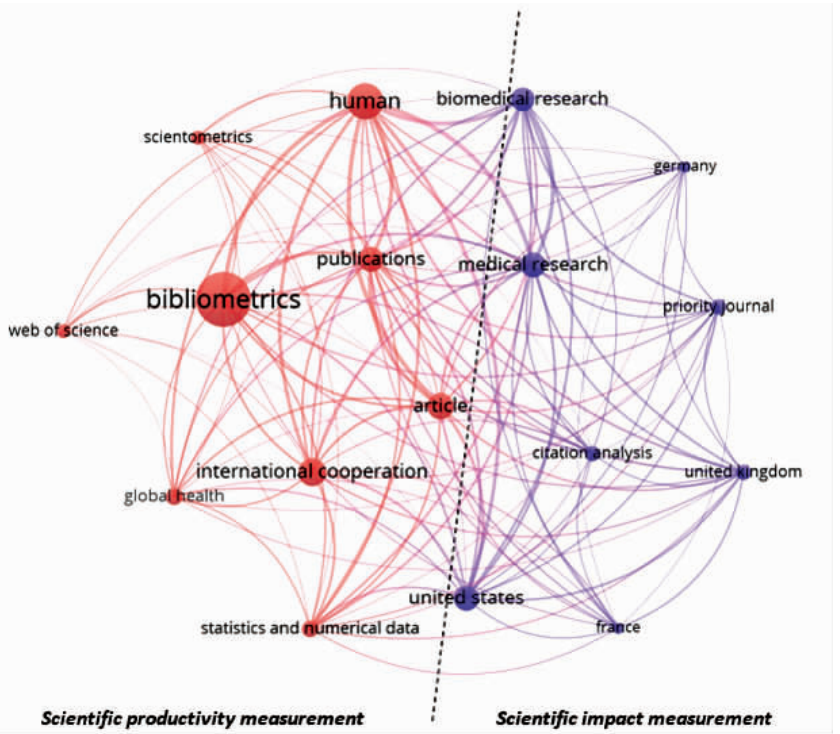

Figure 2 Two clusters of IRCM publications' keywords.

In the above figure, the size of the node reflects the co-occurrence of a given word (i.e., the higher the number, the larger the size of the node). The figure shows that the most frequently used keywords (including their alternatives) are "bibliometrics", "human", "article", "publications" "scientometrics", "web of science", "citation analysis", "global health", "statistics and numerical data", "priority journal", some countries' names ("united states", "united kingdom", "Germany", "France") and some fields of study ("biomedical research", medical research").

Analysing publications' keywords in each cluster, the common themes suggested two research topics of IRCM publications:

The first research topic is scientific productivity measurement. The publications involved in this cluster mainly mentioned ways to measure scientific productivity through scientific outputs (e.g., publications).

The second research topic is scientific impact measurement. The publications involved in this cluster mainly mentioned ways to measure the impact of scientific activities. The scientific impact is commonly measured by citation based metrics such as h-index or citation counts.

The present study also identified the most common terms that were used in the IRCM

publications. These terms were extracted from the IRCM publications' abstracts. Figure 3 showed the co-occurrence map including 2 clusters of the most common terms.

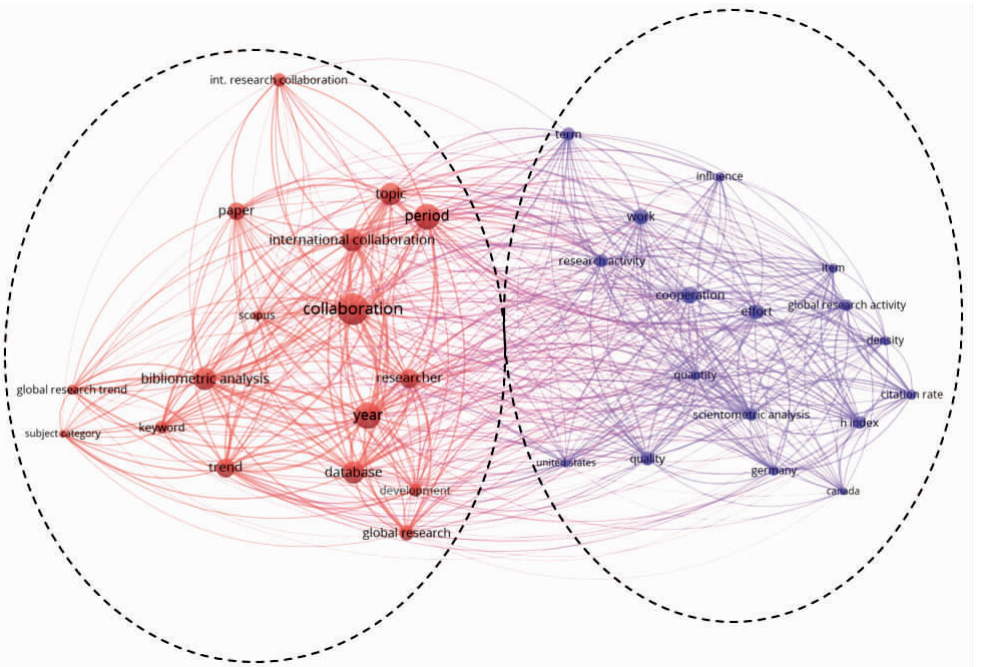

Figure 3 Common terms used in IRCM publications.

The above figure shows some notable points. The terms in these two clusters seem to be matched with the two research areas identified previously: scientific productivity measurement and scientific impact measurement. However, "article" and "publications" are the common keywords in IRCM publications while "paper" is the corresponding term commonly used in their titles and abstracts. Another difference is that "web of science" seems to be mentioned more frequently in these publications' keywords and "scopus" is mentioned more in these publications' titles and abstracts.

The two clusters in the keyword co-occurrence map suggest that there are two primary research topics of IRCM publications. These two research topics can be categorised as scientific productivity measurement and scientific impact measurement. Further reading on 65 relevant publications shows that this classification holds true across disciplines because both research topics were covered in different subject areas (e.g., in medical research). A classification into two mentioned research topics is still simple but necessary to provide an overview of the available research evidence. This finding is similar to a previous suggestion that the two indicators of research collaborations, namely co-authorship and citations, reflect two main aspects of assessment: quantitative and qualitative (Schmoch & Schubert, 2008). However, the studies researching scientific impact measurement focus on the individual level and institutional level rather than the country level. One implication of this finding is that IRCM studies about the research collaboration between countries should review the literature of the cluster scientific productivity measurement specifically. By contrast, studies reviewing the whole IRCM literature could include the IRCM studies at all levels (i.e., individual level, insti-

tution level and country level). The findings in the present study can be used as the initial scoping work before embarking on further IRCM literature review, in the context that IRCM is still an underdeveloped domain and more exploratory studies should be carried out.

There are two final implications that can be inferred from the term co-occurrence map. Firstly, this map reflects the current state of IRCM studies' content. Therefore, suitable combinations of key terms should be used in searching the relevant publications. In particular, "international collaboration", "measure" and "publications" should be used together in searching the productivity of IRC, while "international collaboration", "measure" and "citation analysis" should be used together in searching publications studying the impact of IRC. Secondly, the co-occurrence maps also suggest some possible gaps in the literature of IRCM. For example, the co-occurrence maps show that "paper", "article" and "publication" are often used as the common form of research output. Regarding the results of research, there are various types of outcomes (e.g., patents, joint research grants) beyond joint research publications (Yuan et al., 2018) and different rewards of the collaborations (e.g., acknowledgements in PhD theses, research journals) as contributionship (Laudel, 2002), and books. These outputs, which may be included in the measurement of research collaboration, have been ignored in the IRCM publications surveyed. Consequently, the use of publications' co-authorship as the only indicator for IRCM does not fully reflect actual collaboration (Melin & Persson, 1996). In addition, the co-occurrence maps also show that Web of Science and Scopus are the common data sources that are used in IRCM publications. However, there are not only various methods to measure "research collaboration" (e.g., association strength, Salton index, inclusion index …) but also various data sources that can be used for IRCM (e.g., Google Scholar and nationally funded research projects...). The results may be different if these studies use different measures (Luukkonen et al., 1993) or use different data sources (De Stefano et al., 2013). Therefore, further studies about the roles of indicators and data, which are elements of IRC measurement other than measures, should be carried out in the future.

5 Conclusions and Future Work

The present study aimed to identify the common venues where IRCM has been discussed, and to identify research topics in the IRCM domain. Our investigation confirmed that IRCM is an interdisciplinary domain, since IRCM studies have been published in source titles of different subject areas. These subject areas are Social Sciences, Computer Science, Decision Sciences, Medicine, and Business, Management and Accounting. The spread across these areas implies that researchers apply theoretical lenses from different disciplines to IRCM. Therefore, it is worthwhile revisiting whether many IRCM studies actually provide insights from just some narrow angle. The study also found that there are two core research topics that have been commonly discussed in the IRCM domain - namely scientific productivity measurement and scientific impact measurement - and that there are common terms used in the IRCM studies of these two research topics. Although only 65 relevant IRCM studies were chosen to analyse, the two core research topics were covered across disciplines in these IRCM studies.

There are limitations in the present study. Firstly, this study considered and analysed only literature found in Scopus, and there were only 65 publications found in the search step. This limited number of publications found within Scopus might affect the reliability of this study's findings. Secondly, this study applied a scoping review, which did not include a dedicated quality assessment of the included studies. Therefore, there might be a risk of bias assess-

ment when the studies were considered relevant and selected from the initial set of publications found. Thirdly, this study used keywords and terms from the IRCM publications found to analyse their co-occurrence. The co-occurrence of these publications' keywords and terms suggested that there were two research topics in the IRCM domain. Other techniques, such as co-citation analysis, could have been applied together with co-occurrence analysis to check if there was any inconsistency in the findings.

Therefore, these above limitations should be a focus of future studies. Future studies could fruitfully explore the state of IRCM research domain further by extending the scope of data and methods applied. The inclusion of Web of Science as another leading database for scientometric studies (Mongeon & Paul-Hus, 2016), for example, is necessary to validate the kinds of conclusions that can be drawn from this study. In addition, future studies should further develop the above initial findings by systematically reviewing the IRCM literature with the formal quality assessment of included studies. As also recommended above, future studies should include different bibliographic techniques in analysing the state of IRCM's research topics to confirm the consistency of their findings.

To conclude, this study filled the knowledge gap of what the common venues are for discussing IRCM, by pointing out that IRCM studies have been published in different journals for particular subject areas. We suggest relevant journals for bibliometric researchers or policy makers could be considered for future IRCM publications. Second, this study provides insights into the IRCM domain specifically. The study found that there are two primary research topics in the IRCM domain, and suggested key terms in each research topic so that future IRCM studies will be able to easily find the relevant publications. The study also found that there are not sufficient studies about the roles of indicator, measure and data in IRCM. These research gaps should be addressed in future studies. From these results, the present study derives practical advice for both finding and conducting research in IRCM.

References

Abramo, G., D´Angelo, C. A., & Solazzi, M. (2011). Are researchers that collaborate more at the international level top performers? An investigation on the Italian university system. *Journal of Informetrics, 5* (1), 204–213.

Arksey, H., & O´Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8* (1), 19–32.

Beaver, D., & Rosen, R. (1978). Studies in scientific collaboration: Part I. The professional origins of scientific co–authorship. *Scientometrics, 1* (1), 65–84.

Chen, K., Zhang, Y., & Fu, X. (2019). International research collaboration: An emerging domain of innovation studies?. *Research Policy, 48* (1), 149–168.

De Beaver, D., & Rosen, R. (1979). Studies in scientific collaboration: Part II. Scientific co–authorship, research productivity and visibility in the French scientific elite, 1799–1830. *Scientometrics, 1* (2), 133–149.

De Stefano, D., Fuccella, V., Vitale, M. P., & Zaccarin, S. (2013). The use of different data sources in the analysis of co–authorship networks and scientific performance. *Social Networks, 35* (3), 370–381.

Gómez–Núñez, A. J., Vargas–Quesada, B., Chinchilla–Rodríguez, Z., Batagelj, V., & Moya–Anegón, F. (2016). Visualization and analysis of SCImago Journal & Country Rank structure via journal clustering. *Aslib Journal of Information Management, 68* (5), 607–627. doi:10.1108/ajim–12–2015–0205

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

Jeangirard, E. (2021). Content–based subject classification at article level in biomedical context. *ArXiv preprint arXiv* : 2104.14800.

Katz, J. S., & Martin, B. R. (1997). What is research collaboration?. *Research Policy, 26* (1), 1–18.

Laudel, G. (2002). What do we measure by co–authorships?. *Research Evaluation, 11* (1), 3–15.

Luukkonen, T., Tijssen, R., Persson, O., & Sivertsen, G. (1993). The measurement of international scientific collaboration. *Scientometrics, 28* (1), 15–36.

Melin, G., & Persson, O. (1996). Studying research collaboration using co–authorships. *Scientometrics, 36* (3), 363–377.

Mongeon, P., & Paul–Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics, 106* (1), 213–228.

Sedighi, M.(2016). Application of word co–occurrence analysis method in mapping of the scientific fields (case study: the field of Informetrics). *Library Review.*

Sharma, S., & Thomas, V. (2008). Inter–country R&D efficiency analysis: An application of data envelopment analysis. *Scientometrics, 76* (3), 483–501.

Schmoch, U., & Schubert, T. (2008). Are international co–publications an indicator for quality of scientific research?. *Scientometrics, 74* (3), 361–377.

Xu, S., Hao, L., An, X., Yang, G., & Wang, F. (2019). Emerging research topics detection with multiple machine learning models. *Journal of Informetrics, 13* (4), 100983.

Yuan, L., Hao, Y., Li, M., Bao, C., Li, J., & Wu, D. (2018). Who are the international research collaboration partners for China? A novel data perspective based on NSFC grants. *Scientometrics, 116* (1), 401–422.