Scientometrics (2021) 126:8375–8431
https://doi.org/10.1007/s11192-021-04128-1

Publication outperformance among global South researchers: An analysis of individual-level and publication-level predictors of positive deviance

Basma Albanna[1] · Julia Handl[2] · Richard Heeks[1]

Received: 19 December 2020 / Accepted: 5 August 2021 / Published online: 13 September 2021

Abstract

Research and development are central to economic growth, and a key challenge for countries of the global South is that their research performance lags behind that of the global North. Yet, among Southern researchers, a few significantly outperform their peers and can be styled research "positive deviants" (PDs). In this paper we ask: who are those PDs, what are their characteristics and how are they able to overcome some of the challenges facing researchers in the global South? We examined a sample of 203 information systems researchers in Egypt who were classified into PDs and non-PDs (NPDs) through an analysis of their publication and citation data. Based on six citation metrics, we were able to identify and group 26 PDs. We then analysed their attributes, attitudes, practices, and publications using a mixed-methods approach involving interviews, a survey and analysis of publication-related datasets. Two predictive models were developed using partial least squares regression; the first predicted if a researcher is a PD or not using individual-level predictors and the second predicted if a paper is a paper of a PD or not using publication-level predictors. PDs represented 13% of the researchers but produced about half of all publications, and had almost double the citations of the overall NPD group. At the individual level, there were significant differences between both groups with regard to research collaborations, capacity development, and research directions. At the publication level, there were differences relating to the topics pursued, publication outlets targeted, and paper features such as length of abstract and number of authors.

Keywords Positive deviance · High-performing researchers · Mixed-methods · Bibliometrics · Citation analysis · Topic modelling · Higher education policy · Global South

✉ Basma Albanna
basma.albanna@manchester.ac.uk

Extended author information available on the last page of the article

Introduction

A nation's scientific research capability, characterised by its direct engagement in the creation of knowledge, plays a vital role in its sustainable economic development, and the strong correlation between science and technology development and economic development is well documented (King, 2004; Man et al., 2004). Scientific research is required both to create the new technologies and techniques that increase local productivity and economic growth, and to adapt technologies imported from abroad (Goldemberg, 1998). A necessary part of this, in order to build a strong knowledge society with a thriving 'culture of science', is the publication and dissemination of research results (Salager-Meyer, 2008).[1]

A clear research divide is visible between the global South[2] and the global North. This can be seen in terms of research investment and capability. For example, the average national expenditure on research and development from 2005 to 2014 was 1.44% of GDP in Northern countries but only 0.38% of GDP in Southern countries (Blicharska et al., 2017) while the number of researchers per million population in 2017 was 4,351 in the global North and 713 in the global South (World Bank, 2020). The divide is also manifest in scientific outputs. In 2018, global North countries produced an average of more than 35,000 scientific and technical journal articles per country while global South countries produced an average of 9700, or 4000 if China and India are excluded[3] (World Bank, 2020). Despite some signs of progress, there also remains an important gap in terms of per-country and per-researcher citation rates between North and South (Confraria et al., 2017; Gonzalez-Brambila et al., 2016). The divide in terms of highly-cited outputs is even starker, with global South researchers (again excluding China and India) authoring less than 2% of the top 1% most-cited articles globally (National Science Board, 2018).

It is this latter issue—low citation rates for Southern research—that forms our particular focus in this paper, and for which a number of explanations have been put forward. Statistical evidence shows that the lower levels of investment and lower relative populations of researchers in the global South are key factors; the latter issue is exacerbated by the brain drain of Southern researchers who relocate to the global North (Man et al., 2004; Pasgaard & Strange, 2013; Salager-Meyer, 2008). Lower levels of English language proficiency are also a factor, given the skew of international journal publication towards English (Confraria et al., 2017; Gonzalez-Brambila et al., 2016; Man et al., 2004). Other recognised institutional exclusion factors and/or biases against Southern researchers include difficulty in securing research grants (Karlsson et al., 2007), and a greater likelihood that reviewers and editors of mainstream scientific journals will reject a paper from a global South institution than a paper of equivalent quality from a global North institution (Gibbs, 1995; Leimu & Koricheva, 2005).

Among the valuable research conducted on this issue to date, there have been three main approaches: country-level statistical analysis, paper-level statistical analysis, or

[1] Acknowledging that this takes a Western perspective on knowledge; a perspective that has been critiqued given the other forms of non-Western knowledge and knowledge production that exist (Thesee 2006).

[2] The terms "South" and "Southern" will be used to refer to countries classified as upper-middle income, lower-middle income, and low income. Accordingly, the terms "North" and "Northern" will be used to refer to countries that are members of the OECD (Organisation for Economic Co-operation and Development) or are classified as high-income economies by the World Bank based on estimates of gross national income per capita.

[3] Small island states and non-UN-member territories are excluded from this calculation.

individual-level analysis. While the latter includes author-related factors, Southern researchers as individuals are rarely investigated. In particular, there has been no previous research focusing on "exceptions to the rule": those few Southern researchers who are able to achieve much higher research performance than their peers. *Pre hoc*, it is reasonable to hypothesise that such researchers could provide valuable insights and lessons that might help to better understand and even mitigate the current North–South divide in research outputs and citation. It is therefore the purpose of this paper to specifically study these exceptions by investigating what characterises high-performing researchers and their publications in a global South context.

In order to do this, we make use of the "positive deviance" (PD) approach, given that this attempts to systematically identify and learn from "outliers"—individuals who are performing substantially better than expected and better than their peers, given the resources and socio-economic conditions they are exposed to (Sternin et al., 1997). First used at scale in order to learn from Vietnamese families with well-nourished children in contexts of widespread malnutrition (Sternin, 2002), the positive deviance approach has subsequently spread to other domains (Albanna & Heeks, 2019). However, the conventional PD approach relies heavily on primary data collection to develop a baseline from which positive deviants (PDs) are identified—a process that is both time- and labour-intensive with costs directly proportional to sample size. Recent developments in the availability of digital datasets have presented new possibilities for the identification and understanding of PDs (Albanna & Heeks, 2019). This new digital data-powered approach to positive deviance was seen as particularly relevant for investigation of scientific researchers given the existence of platforms that digitally index and/or analyse the scholarly work of researchers, enabling evaluation of their performance through multiple dimensions and metrics.

For this study, we chose a sample of 203 information systems (IS) researchers from Egypt to identify factors that enabled a few positive deviants to outperform their peers. Positive deviants are defined as researchers who outperform their peers in both productivity (articles published) and impact (article citations). They were identified based on six citation metrics that take into account those two dimensions of performance in different ways. We conducted an analysis of the researchers' attributes, attitudes, practices, and publications, based on a mixed-methods approach that employed interviews, surveys, and analysis of publication-related datasets. Two methodological innovations were developed in this study. The first was the use of multiple performance metrics in identifying PDs, which enabled us to profile PDs into groups based on those metrics. The second was the identification of extrinsic and intrinsic predictors of PDs' publications as a way of understanding and reflecting on some of their publication strategies. Hence, this paper has two main contributions. The first is *contextual,* which is the identification of predictors of high performers or PDs in a Southern country, who face challenges different from those facing researchers in Northern countries. And the second is *methodological*, where a combination of multiple performance indicators and a number of data sources were used to develop a holistic approach for identifying, profiling and characterising PDs.

In what follows, we first present a review of related work on high-performing researchers before explaining the data sources and methodology of this data-powered positive deviance approach. The methodology steps are then undertaken: defining the study focus, determining the positive deviants, and discovering the features of positive deviants and their published papers. We end with discussion and conclusions.

Related work

There is a substantial body of research on the predictors of individual-level high research performance over the last four decades. While the terminology of "positive deviants" has not been used; analogous concepts and synonymous terms have been. Relevant literature on *highly productive academics* includes work studying their attitudes, practices and perceptions in 11 European countries (Kwiek, 2016), and their attributes, perceptions and structural predictors in China, Japan and South Korea (Postiglione & Jung, 2013); a series of studies that investigated the characteristics and work habits of the top (three or four) educational psychology researchers in the US (Kiewra & Creswell, 2000; Patterson-Hazley & Kiewra, 2013) and in Germany (Flanigan et al., 2018); and a paper on the strategies and attributes of highly productive academics in school psychology, who were mainly Americans (Martínez et al., 2011). Other research that has looked into *top performers* includes Kwiek's study (2018), which investigated both individual and institutional variables to identify predictors of research success for the top 10% of Polish academics, and Kelchtermans and Veugelers' (2013) study on top performing Belgian researchers, which investigated the effects of co-authorship, gender and previous top performance. There are also studies on *research stars* such as the study by Yair et al. (2017) on Israeli Prize laureates in life and exact sciences; and the study by White et al. (2012) on American researchers in business schools, where individual and situational variables were explored. *High achievers* were identified in a study by Harris and Kaine (1994) that investigated the preferences and perceptions of high-performing Australian university economists and *highly cited scientists* were studied by Parker et al. (2010) and Parker et al. (2013) who sought to identify the social characteristics and opinions of the 0.1% most cited environmental scientists and ecologists worldwide. *Eminent scientists* were studied by Prpić (1996) to explore the most important predictors of productivity among Croatian scientists and *top producing* researchers were identified in a study by Mayrath (2008) which aimed at understanding the attributes of the authors having the most publications in educational psychology journals.

Beneath the factors identified in these studies have lain theoretical models proposed to explain the research performance and outperformance observed, of which three will be mentioned here. The *sacred spark theory* (Cole & Cole, 1973) states that highly productive researchers have an inner drive and motivation to do science that is fuelled by their love of the work. Other theories look more at the external environment. *Utility maximisation theory* (Kyvik, 1990) argues that the extent to which researchers research and publish—as opposed to other activities—is determined by the personal utility or benefit they perceive themselves getting; that utility often being significantly determined by external incentives or disincentives that attach to the different activities. *Cumulative advantage theory* (Merton, 1968) is somewhat similar in identifying external reward systems, and their reinforcement or otherwise of research and publication activity, as important (Cole & Cole, 1973). But it sees researchers who begin with some advantage (either innate or external) being increasingly more productive over time compared to others as they gain further advantage, such as greater likelihood of obtaining research grants, or participating in collaborations.

As summarised next, this work has been of significant value in providing insights into high-performing researchers. However, we can also identify three lacunae which the current paper seeks to address.

First, geographic. It is evident from the above that there is a geographic concentration of such studies on high-income countries of the global North: the one study including China is the sole exception. There has thus been practically no consideration of research performance in the resource-constrained countries of the global South. Addressing this unexplored topic is particularly pressing, given the imperative to improve the contribution of research to national development in these countries, and given that findings from such a study might lead to new context-aware predictors of high research performance that could mitigate some of the challenges reflected in the current North–South divide. Hence the justification for the current paper.

Second, methodological in relation to the dependent variable of performance measurement. The majority of studies identify and rank high-performing researchers based on (i) productivity, as measured by number of articles published (e.g. Harris & Kaine, 1994; Kwiek, 2016; Postiglione & Jung, 2013; Prpić, 1996) or, in case of some studies, (ii) impact as measured by number of citations (e.g. Parker et al., 2010, 2013). There are clearly benefits in incorporating both productivity and impact measures yet this was rarely found in the literature reviewed (Altanopoulou et al., 2012). Recent advances in citation metrics and availability of tools such as Harzing's Publish or Perish software (Harzing, 2007) provide an opportunity to measure performance along different dimensions[4] and using combined measures. The current study therefore combines a number of citation metrics to evaluate researchers; enabling a balanced consideration of both productivity and impact and allowing control for factors like article and author age.

Third, methodological in relation to the independent variables or predictors. Table 1 presents the significant predictors of high performers in research, identified from previous studies and forming a foundation for modelling and analysis for the current study. These can be grouped into seven main categories: *Personal or Demographic* characteristics such as age, gender and education; *Internationalisation and Research Collaboration* such as participation in domestic or international research teams; *Research Engagement* with publishing entities; *Research Approach* including focus*; Academic Roles* covering distribution of time between different academic activities; *Practices* associated with undertaking research; and *Institution* predictors related to work environment actuality or preferences.

These predictors were almost all identified using qualitative methods such as interviews (Flanigan et al., 2018; Kiewra & Creswell, 2000), quantitative methods such as surveys (Harris & Kaine, 1994; Kwiek, 2016, 2018; Postiglione & Jung, 2013; Prpić, 1996) or a mix of both (Martínez et al., 2011; Mayrath, 2008; Patterson-Hazley & Kiewra, 2013). What is consistent here is the focus on individual researcher-level data. Largely missing has been publication-level data.

Yet there are a number of reasons for thinking that adding in publication-level data can provide valuable additional insights. Publication data can provide insights into some of the individual-level predictors: for example, the amount and type of research collaboration undertaken by high-performing researchers compared to other researchers. Past studies also identify three main groups of high citation predictors: author, journal and paper (Onodera & Yoshikane, 2015; Walters, 2006) and there is some evidence that within these predictors, the most important are those related to the paper (Stewart, 1983). Publication analysis can therefore determine whether the papers of high-performing researchers match characteristics of papers known to be associated with high citation rates, such as title length (Elgendi, 2019), paper length (Onodera & Yoshikane, 2015), number of references (Didegah & Thelwall, 2013) and figures (Haslam et al., 2008), coverage of certain topics (Mann et al., 2006) and keywords (Hu et al., 2020), number of authors (Peng & Zhu, 2012), quality of journal

[4] https://harzing.com/pophelp/metrics.htm.

Table 1 Significant predictors of high-performing researchers from previous studies

Significant predictors	References
Personal/Demographics	
Gender: being male	(Prpić, 1996) (Parker et al., 2010) (Patterson-Hazley & Kiewra, 2013) (Kwiek, 2016)
Being older in age or in years of active publication	(Prpić, 1996) (Parker et al., 2010) (Patterson-Hazley & Kiewra, 2013) (Kwiek, 2016)
Younger age of obtaining PhD	(Prpić, 1996)
Holding academic rank of professor	(Kelchtermans & Veugelers, 2013) (Kwiek, 2016)
Training in top programmes in top schools	(Kiewra & Creswell, 2000)
Having outside interests	(Kiewra & Creswell, 2000)
Having notable figures as advisors	(Mayrath, 2008) (Flanigan et al., 2018)
Ability in more than two foreign languages	(Prpić, 1996)
Good writing skills	(Kiewra & Creswell, 2000) (Mayrath, 2008)
Passion, curiosity and/or deep interest in research	(Mayrath, 2008)
Internationalisation and Research Collaborations	
Connectivity with top researchers	(Patterson-Hazley & Kiewra, 2013)
Domestic research collaborations	(Harris & Kaine, 1994) (Mayrath, 2008) (Kwiek, 2016)
International research collaborations	(Harris & Kaine, 1994) (Prpić, 1996) (Postiglione & Jung, 2013) (Kwiek, 2016) (Kwiek, 2018)
Publishing abroad	(Harris & Kaine, 1994) (Prpić, 1996) (Kwiek, 2016) (Kwiek, 2018)
Research is international in scope	(Kwiek, 2016) (Kwiek, 2018)
Co-authorship: more co-authors	(Prpić, 1996) (Kelchtermans & Veugelers, 2013)
Research Engagement	
Engaging in peer review	(Prpić, 1996) (Kiewra & Creswell, 2000) (Kwiek, 2016)
Being editor of journals/book series	(Harris & Kaine, 1994) (Prpić, 1996) (Kiewra & Creswell, 2000) (Kwiek, 2016) (Kwiek, 2018)
Research Approach	
Focus on basic/theoretical research	(Parker et al., 2010) (Postiglione & Jung, 2013) (Kwiek, 2016)
Focus on pioneering science i.e. exploring novel, under-researched issues	(Kiewra & Creswell, 2000)
Constant research focus i.e. finding a niche and carving it out	(Kiewra & Creswell, 2000) (Mayrath, 2008)
Research leading to acceptable results and not necessarily spectacular results	(Harris & Kaine, 1994)
Research which looks for immediate solutions	(Harris & Kaine, 1994)
Research that will enhance reputation and prospects for promotion	(Harris & Kaine, 1994)
Academic Roles	
Having an administrative role	(Patterson-Hazley & Kiewra, 2013) (Kelchtermans & Veugelers, 2013)
More time available for research: lower teaching load	(White et al., 2012) (Postiglione & Jung, 2013) (Kwiek, 2016) (Kwiek, 2018)
Mentoring/supervising students	(Prpić, 1996) (Kiewra & Creswell, 2000) (Mayrath, 2008) (White et al., 2012)

Table 1 (continued)

Significant predictors	References
Practices	
Time management practices/skills (e.g. stability in daily routines and working for long hours, fixed academic writing time)	(Kiewra & Creswell, 2000) (Mayrath, 2008) (Parker et al., 2010) (White et al., 2012) (Kwiek, 2016) (Flanigan et al., 2018)
Research management strategies (e.g. meeting weekly with collaborators)	(Mayrath, 2008) (Flanigan et al., 2018)
Publishing professional and/or scientific works during their undergraduate studies	(Prpić, 1996)
Receiving feedback on manuscripts from colleagues or mentors	(Mayrath, 2008)
Very good knowledge of the literature	(Mayrath, 2008)
Setting deadlines	(White et al., 2012)
Institution	
Research is considered in HR decisions such as promotion	(Kwiek, 2016)
Workplace has a strong performance orientation	(Postiglione & Jung, 2013) (Kwiek, 2016)
Workplace perceived as conducive to research	(Harris & Kaine, 1994)
Workplace perceived as a relaxed environment	(Harris & Kaine, 1994)
Workplace perceived as an environment that provides opportunity to work on challenging problems	(Harris & Kaine, 1994)

(Davis et al., 2008), etc. Additionally, predictors relevant to the global North–South divide have been identified in a number of paper-focused studies which found that the author's nationality (whether from the United States or not) (Walters, 2006), the regional focus of the articles (focusing on the United States or Europe) and the language of the journal (Van Dalen & Henkens, 2005) had a significant positive effect on the average citations.

Three particular theories are employed by these studies to examine the factors affecting publication citation. The *normative view* (Hagstorm 1965; Kaplan, 1965; Merton, 1973) assumes that science is a normative institution governed by internal rewards and sanctions (Baldi, 1998). According to this perspective the intrinsic characteristics of papers (i.e. content/quality) are the main driver of citations. The *social constructivists' view* (Gilbert, 1977; Knorr-Cetina, 1981; Latour, 1987) argues that scientific knowledge is socially constructed through the manipulation of political and financial resources and citations are used as persuasion tools (Peng & Zhu, 2012). The citing behaviour in this view is driven by a paper's extrinsic characteristics, such as the location of the cited paper author within the stratification structure of science, that would convince the reader with the validity of the arguments (Baldi, 1998). The *natural growth mechanism* (Glänzel & Schoepflin, 1995) states that citations are driven mainly by the interaction between publication-level time dependent factors and factors related to the publication outlets. It sees that the characteristics of academic journals (e.g. journal prestige, self-citation rate, maturation speed) will interact with time to affect citations of papers (Peng & Zhu, 2012).

All of this supports incorporation of publication-level predictors of outperformance when considering what predicts outperformance of individual researchers. In this study we therefore use a combination of researcher-level data gathering (via interview and survey) and publication-level analysis, in order to provide a fuller picture of research performance predictors.

Methodology and data

The positive deviance approach consists of five steps: "(1) *Define* the problem, current perceived causes, challenges and constraints, common practices, and desired outcomes. (2) *Determine* the presence of positive deviant individuals or groups in the community. (3) *Discover* uncommon but successful practices and strategies through inquiry and observation. (4) *Design* activities to allow community members to practice the discovered behaviours. (5) *Monitor* and evaluate the resulting project or initiative" (Positive Deviance Initiative, 2010). Time and resourcing constraints meant that only the first three steps could be essayed in this study. We also diverged from the traditional approach by using this study as a testbed for what we will call the "data-powered positive deviance" (DPPD) methodology. Where traditional PD relies on freshly- and specifically-gathered field data, the idea behind DPPD is that it uses pre-existing digital data sources instead of—or in conjunction with—traditional data sources. It uses digital datasets to identify positive deviants (those performing unexpectedly well in a specific outcome measure that is digitally recorded, mediated or observed) and potentially also to understand the characteristics and practices of those PDs if digitally recorded (Albanna & Heeks, 2019). The potential of DPPD is that it can mitigate some of the challenges of traditional PD approaches by reducing time, cost and effort, and can add to the positive deviance approach by identifying positive deviants in new ways and domains (Albanna & Heeks, 2019).

Scientometrics—the field of study that focuses on measuring and analysing scientific literature ('Scientometrics' 2020)—is well suited for a positive deviance approach because, as reflected in the discussion of literature above, research performance does not follow a normal distribution (O'Boyle & Aguinis, 2012). Instead, it follows a Pareto or power distribution characterised by strong skewness with a long tail to the right that includes a number of high-performing outliers; sufficient to provide a sample of positive deviants. Scientometrics is well suited to DPPD specifically for two reasons. First, because the proliferation of electronic research databases has made it possible to develop scientific evaluation indicators that can be used to digitally measure the performance of researchers (e.g. h-index) and journals (e.g. impact factor). Second, because the emergence of advanced data analytics tools alongside the emergence of a variety of large scale datasets (such as citations, references, publication outlets, usage data, paper content, etc.) has made it possible to not only measure performance, but to also analyse the practices of the researchers, characterise their scientific outputs, and predict their future performance. Specifically, this can be rendered possible through techniques such as network analysis, topic modelling, predictive analytics and co-citation analysis.

As discussed above, in this study we used a mixed-methods approach to identify PDs and to analyse their practices. In the *Define* step, we used secondary data from Egyptian university websites and from Google Scholar to set the frame of Egyptian researchers for analysis. In the *Determine* step, we extracted for each researcher the bibliometric data of all his/her academic outputs that were produced while being affiliated to an Egyptian university, without the exclusion of publication or research type. This data was analysed with statistical software R v3.4.1 to identify the positive deviants and non-positive deviants (NPDs) within the overall population. During Stage 1 of the *Discover* step, primary

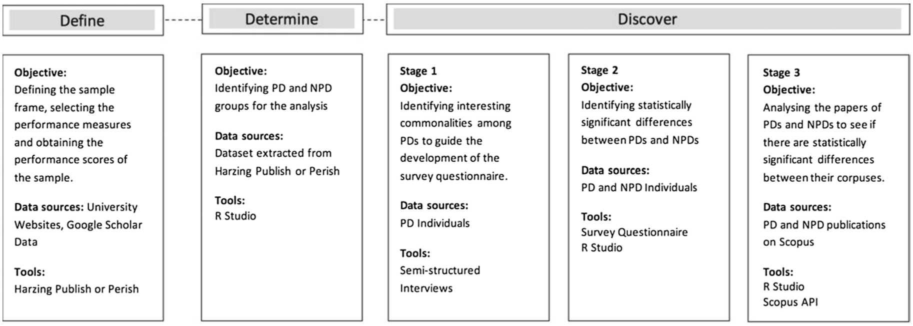

Define
Determine
Discover
Objective:
Defining the sample frame, selecting the performance measures and obtaining the performance scores of the sample.
Data sources: University Websites, Google Scholar Data
Tools:
Harzing Publish or Perish
Objective:
Identifying PD and NPD groups for the analysis
Data sources:
Dataset extracted from Harzing Publish or Perish
Tools:
R Studio
Stage 1
Objective:
Identifying interesting commonalities among PDs to guide the development of the survey questionnaire.
Data sources:
PD Individuals
Tools:
Semi-structured Interviews
Stage 2
Objective:
Identifying statistically significant differences between PDs and NPDs
Data sources:
PD and NPD Individuals
Tools:
Survey Questionnaire
R Studio
Stage 3
Objective:
Analysing the papers of PDs and NPDs to see if there are statistically significant differences between their corpuses.
Data sources:
PD and NPD publications on Scopus
Tools:
R Studio
Scopus API

Fig. 1 Summary of the applied data-powered positive deviance process

data was collected through in-depth interviews from a sample of PDs to explore practices, attitudes and attributes that might distinguish them from NPDs. During Stage 2 of the *Discover* step, the key findings from Stage 1 plus other predictors of research performance drawn from the literature (see Table 1) were used to design a survey tool. That survey then targeted the whole population and tested if the proposed differentiators were significantly different between the two groups (PDs and NPDs). Finally, in Stage 3 of the *Discover* step, the Scopus secondary dataset was used as the basis for analysis of researcher publications; extending and validating some of the findings identified in the previous steps. Figure 1 summarises the process used to identify PDs and to discover predictors of their performance, and outlines the structure of findings, presented next.

Findings

Define

The study population comprises IS researchers in Egyptian public universities. A single discipline was chosen to avoid variations in, for example, typical publication and citation rates that arise between different disciplines. Information systems was chosen because of the growing importance of research on digital technologies including technological development and implementation research to economic development, and because a pre-check showed ready presence of a substantial number of Egyptian IS researchers and publications in the main secondary datasets. Egypt was chosen because it was the first author's home country, with social contacts affording ready access to university departments and staff. Public universities were chosen to ensure that all researchers in the sample worked in a context of similar resource constraints, albeit with slight variations between universities within or outside the Greater Cairo area.

In Egypt there are 29 public universities, 11 of which do not have computer science faculties or IS departments and seven of which do not have an online directory of the IS department staff. So the final sample included 11 universities: Cairo, Ain Shams, Benha, Helwan, Mansoura, Fayoum, Menofeya, Assiut, Zagazig, Kafrelsheikh and Port Said. The total number of faculty members in those universities was 304 but for this study we only included those researchers who hold at least a Masters' degree[5] and have published at least one article. This guarantees that they have some publishing experience. Consequently, the final sample that we targeted for this study included 203 researchers who were assistant lecturers, lecturers, assistant professors and professors. (In the Egyptian higher education system, the first academic rank is assistant lecturer, which you receive once you obtain your Masters' degree and then you become a lecturer when you obtain your PhD. The following rank is associate professor and then professor, which are obtained based on both years of experience and publications.)

Using the university websites, the names, degrees and email addresses (if provided) of these researchers were identified and this information was used to extract their citation data from the Publish or Perish (PoP) software. PoP is a freely accessible software program that extracts data for researchers from a number of sources (e.g. Web of Science, Scopus, Google Scholar) to provide a variety of research citation metrics (Harzing, 2007). The citation data was extracted for the study sample in July 2018. We only included the papers that had the Egyptian university affiliation i.e. publications produced while doing a PhD abroad were excluded to ensure a fair comparison and to reduce the effects of confounding variables associated with universities abroad.

For this study, Google Scholar was chosen as the source for bibliometrics. The choice of Google Scholar was driven by the fact that ISI citation databases, such as Web of Science, limit citations to journals in the ISI databases. They do not count citations from books and conference proceedings, cover mainly English language articles and provide different coverage in different fields. Such databases significantly underestimate researchers' publications and citations (Ortega, 2015). Prior literature also supports the fact that Google Scholar outperforms Web of Science in coverage (Kousha & Thelwall, 2007), especially for articles that were published from 1990 onwards (Belew, 2005) and for computer science-related research in which conference papers form a key means of publication (Franceschet, 2010). Additionally, Google Scholar is freely accessible, which makes the DPPD method used in this case study easily replicable in different scientific fields and countries. The main drawback of Google Scholar is that its consistency and the accuracy of data is lower than commercial citation enhanced databases such as Web of Science (Jacso, 2005). Hence, extra time was needed to check the accuracy of obtained results.

For every researcher six main citation metrics—extracted and derived from PoP query results—were used to measure research performance as shown in Table 2.

In this study positive deviants are researchers who outperformed their peers in at least one of the six citation metrics presented in Table 2. The need to use multiple measures was motivated by the drawbacks of relying only on the h-index. These drawbacks include the influence of length of researcher's scientific career, with the h-index reflecting longevity as much as it reflects quality (Alonso et al., 2009; Van Noorden, 2010), in addition to its insensitivity to highly cited papers (Egghe, 2006). Using multiple measures enabled us to avoid putting certain groups at a disadvantage due to factors such as the length of their research career, the size of their research departments, or their publication strategies. Measures like the m-quotient, the aw-index and the hc-index ensured that outperformance is detected regardless of the publication age of the author or the age of the paper. Similarly, some IS departments are larger than others, enabling them to have more research collaborations, and to reduce the potential bias due to the larger pool of research collaborators, the hi-index was employed.

[5] It is a prerequisite for confirmed appointment to publish at least two papers towards your Masters' degree in the majority of computer science faculties across the country (those faculties being the typical home of information systems departments and/or researchers).

Table 2 Citation metrics extracted for each researcher to measure performance

Citation metric	Description
h-index	Hirsch's h-index (Hirsch, 2005) is the most widely used single-number measure for assessing the research performance of a researcher. It provides a metric that balances impact and productivity. For example, a researcher has a h-index equal to 10 if 10 of his/her papers have received at least 10 citations and the remaining papers received no more than 10 citations
g-index	Egghe's g-index (Egghe, 2006) aims at improving the h-index by giving more weight to highly cited papers. It is calculated based on the distribution of citations received by a given researcher's publications. A g-index of 20 means that an academic has published at least 20 articles that, combined, have received at least 400 citations (i.e. g^2). Unlike the h-index, which requires that each one of the 20 publications should have at least 20 citations, the g-index takes the cumulative number of citations, allowing high numbers to be driven by a small number of articles
hc-index	The contemporary h-index (Sidiropoulos et al., 2007) rewards academics who maintain a steady level of research activity by giving more weight to recently published articles. The weighting in both the original and the PoP implementations mean, for example, that citations for an article published in the current year count four times whereas citations for papers published four years previously count only once
hi-index	The individual h-index (Batista et al., 2006) reduces the effects of co-authorship by dividing the standard h-index by the average number of authors in the articles that contribute to the h-index
aw-index	The aw-index is derived from the age weighted citation rate (AWCR) which measures the number of citations for the articles contributing to the h-index, adjusted for the age of each article, where the count of citations for a specific article is divided by how old it is and then summed (Sidiropoulos et al., 2007). The aw-index is defined as the square root of the AWCR to make it more comparable with the h-index. In PoP, the adjusted citation counts are summed across all papers, not just those contributing to the h-index, as this captures the impact of the total body of work more accurately. It also allows more recent and less-cited papers to contribute to the AWCR, even though they might not yet contribute to the h-index
m-quotient	The m-quotient was proposed by Hirsch to avoid putting early career researchers at a disadvantage (Hirsch, 2005) and enabled the inclusion of young researchers in the study sample. It is calculated by dividing the h-index by the publication span (i.e. the number of years since the first publication)

We were also interested in researchers with selective publication strategies: those who do not necessarily publish a very high number of papers but who do attain a high impact. This group of researchers can be unfairly assessed using the h-index, while led us to use the g-index. In summary, we can see that these established citation metrics are complementary, as they make different assumptions and have different biases, and that combining the different measures provides a more comprehensive picture of performance.

Determine

Positive deviants are typically identified as specifically-calculated outliers from some measure of central tendency. As can be seen from Fig. 2—violin plots[6] of the distribution of each of the six measures across the entire sample—the data here is not normally

[6] Violin plots are similar to box plots except that they also show the probability density of the data at different values, usually smoothed by a kernel density estimator.

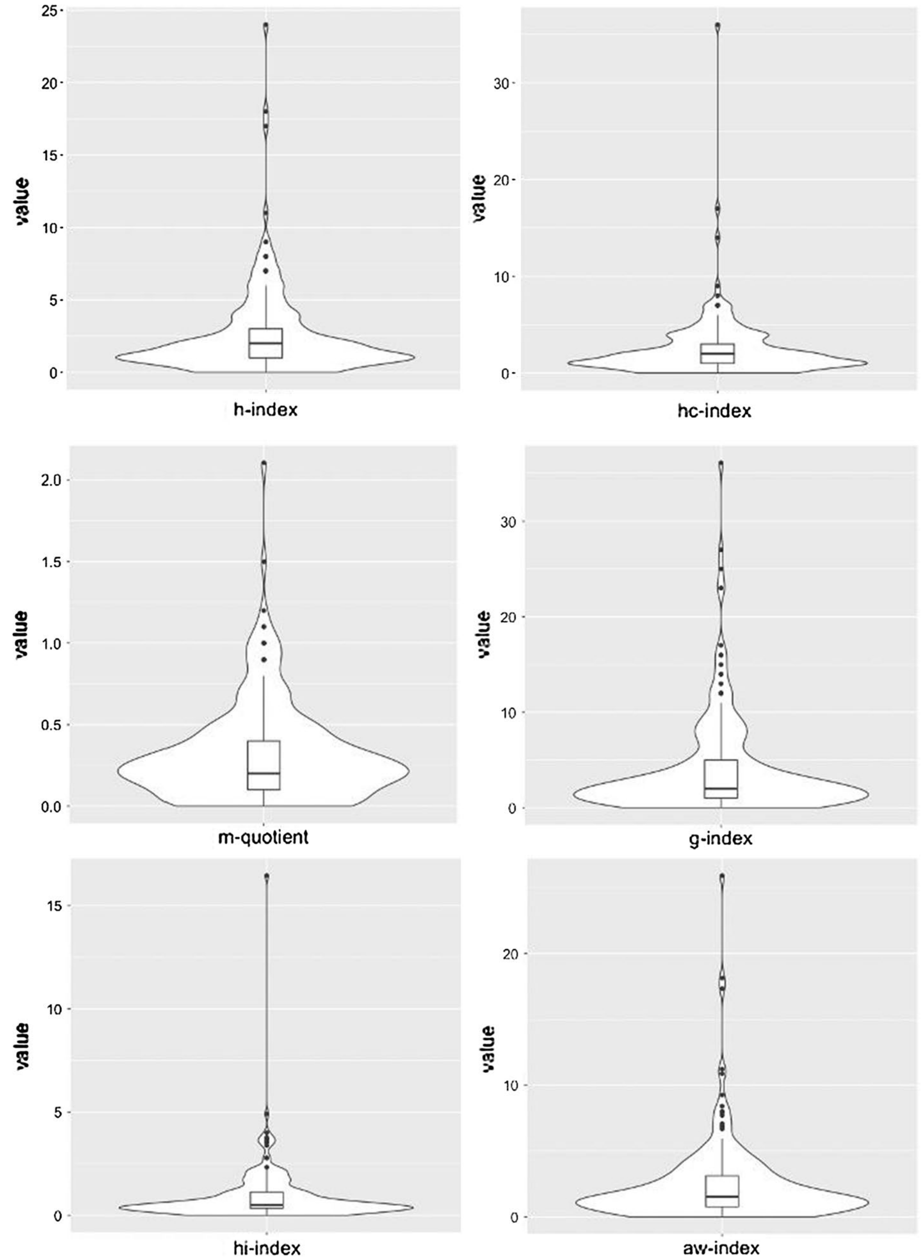


Fig. 2 Violin plots of the sample's scores across the six measures showing the outliers

distributed. Instead, and consistent with the past findings on researcher performance reported above (O'Boyle & Aguinis, 2012) it shows a skewed, Pareto distribution with a long tail above the mean. This makes the *mean* a skewed indicator of central tendency and invalidates the method of identifying positive deviants or outliers in a normally-distributed

Table 3 Summary statistics of the study population

	PDs ($n=26$)	NPDs ($n=177$)	Population ($n=203$)
Average h-index	7.7	1.7	2.5
Average hc-index	7.3	1.7	2.4
Average m-quotient	0.82	0.24	0.3
Average g-index	13.5	2.6	4.0
Average hi-index	2.8	0.58	0.8
Average aw-index	7.8	1.6	2.4
Percentage of assistant lecturers	7.6%	49.2%	43.8%
Percentage of lecturers	26.9%	21.3%	22.2%
Percentage of associate professors	19.2%	18.8%	18.7%
Percentage of professors	46.3%	10.7%	15.3%
Average number of publication years	10.7	6.9	8.3
Average number of papers	43.7	5.9	10.7
Average number of citations	387.1	19.7	66.8

As a reminder, these figures exclude papers published while researchers were overseas

population, which would define them as those observations lying beyond two or three standard deviations above the mean.

Instead, we used the *median* as an indicator of central tendency and employed the interquartile range (IQR) method (Hampel, 1974) to identify positive deviants based on their deviation from the median. In the IQR method, the dataset is divided into four parts, the values that separate the parts are called the first, second, and third quartiles; and they are denoted by Q1, Q2, and Q3, respectively. Q2 is the median of ordered observations, Q1 is the median of observations ordered before Q2 and Q3 is the median of observations ordered after Q2. IQR is Q3–Q1 and outliers are defined as observations that lie beyond 1.5*IQR (Walfish, 2006). In this case study, PDs were defined as individuals lying beyond the 1.5*IQR added to the third quartile in at least one of the six citation metrics that we used as measures of performance. In total, 26 unique PDs were identified and their average performance metrics in comparison to the NPDs are summarised in Table 3.

Cluster analysis

Hierarchical clustering was used to identify groups of PDs based on the citation metrics in which they were found similar, i.e. all members of a cluster are outliers in similar citation metrics. To support this analysis, a binary vector composed of six dummy variables (representing the six citation metrics) was constructed for each of the positive deviants identified,. A value of “1” indicates that this PD is an outlier in the corresponding metric and a value of “0” indicates that this PD is not an outlier in this metric. We then used the *hclust* function of the R *cluster* package[7] to implement complete linkage agglomerative hierarchical clustering using the Gower distance. This method usually yields clusters that are compact and well separated, and the complete linkage criterion ensures direct control of

[7] https://cran.r-project.org/web/packages/cluster/cluster.pdf

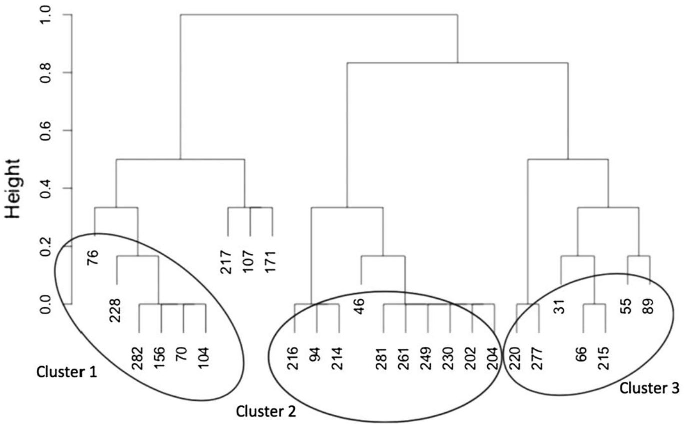


Fig. 3 Hierarchical clustering of PD researchers based on their outlier scores

the maximum dissimilarity in each cluster. A graphical representation of the resulting hierarchical tree (i.e. dendrogram) is presented in Fig. 3.[8]

As shown in Fig. 3, we were able to cluster the 26 researchers into three main clusters as follows.[9]

Cluster 1: Rising stars This cluster includes six researchers who were outliers either in the m-quotient, which discounts longevity and citation skews against junior researchers, and/or the aw-index, which gives weight to more recent and as yet less cited papers by calculating age weighted citation rates for the researcher's papers. Researchers belonging to this group were mainly assistant lecturers and lecturers (with the exception of one associate professor) and they were characterised by a short publication span and a small publication volume (as shown in Table 5).

Cluster 2: Exceptional performers This cluster includes ten researchers who were outliers in all the six metrics collectively, each being an outlier in at least five metrics. Researcher 46 was an outlier in all metrics except for the hi-index, which might indicate that they have very few single authored papers or that they usually publish with a large number of authors. Researchers 216, 94 and 214 were outliers in all measures except for the m-quotient. The remaining six researchers were outliers in all six metrics. Researchers belonging to this group are characterised by balancing all performance measures i.e. productivity, impact and

[8] The numbers in Fig. 3 are the unique ID numbers allocated to each individual researcher; allocated from the broad initial population of 304 researchers.

[9] We could not directly group the remaining three researchers (107, 171 and 217) into any of the previous clusters, since each was an outlier in a unique metric, which was not the characterising metric(s) of the identified cluster. Researcher 107 was an outlier in the hc-index, which gives more weight to citations from recent papers. Therefore, this researcher might be considered as one of the "rising stars" despite not being identified as a potential candidate in the dendrogram and having a high publication age (11) in comparison to the average publication age of the group, which is 3.5 as shown in Table 5. Researcher 217 was an outlier in the h-index and researcher 171 was an outlier in the hi-index. Being an outlier in the hi-index indicates independence: researcher 171 has a lot of single-authored papers or papers with a small number of researchers, and his/her average number of authors per paper was only two authors.

Table 4 Average group scores in each of the six citation metrics with grouping measures highlighted by colour shading

Cluster	m-quotient	h-index	g-index	hc-index	hi-index	aw-index
1	0.95	3	4.5	3.3	0.84	5.13
2	0.98	11.7	20.6	11.9	4.7	11.6
3	0.51	6.8	13.2	5	2.3	6.09

Table 5 Group scores in relevant performance indicators

Cluster	Average publication age	Average no. of papers	Average no. of citations	Average no. of authors/paper	Average no. of cites/paper
1	3.5	6.3	56.1	3.4	8
2	13.3	87.2	726.6	3.0	9
3	14	25	313.85	2.9	16

consistency, and having an old average publication age. They also have the highest average aw-index indicating sustained production of highly cited articles. They were mainly professors with the exception of one lecturer and one associate professor.

Cluster 3: Highly cited researchers This cluster includes seven researchers who are all outliers in the g-index, which gives more weight to highly cited papers. In addition to the g-index, researchers 220 & 227 were outliers in the hi-index which means that they publish mainly individually or with small groups of co-authors; researchers 31 and 89 were outliers in the aw-index, meaning that their highly cited papers are recent, resulting in a high age weighted citation rate. As shown in Table 5, they are characterised by having the longest publication span and the highest number of citations per paper across all clusters.

Table 4 shows the average scores of the three clusters across the six performance measures and Table 5 shows the average scores of those clusters across other relevant performance indicators.

Discover

This study used three separate methodologies—in-depth interviews, surveys and publication analysis—to triangulate data on underlying attributes, practices and attitudes of PDs, thus helping to validate findings. The three methodologies are interrelated and were undertaken sequentially in three stages, with the findings from one stage guiding design of the following stage.

Stage 1: Interviews

The objective of this stage was to identify uncommon strategies and practices among positive deviant researchers, which could be used to guide the design of the Stage 2 survey questionnaire. This stage was incorporated into the methodology because predictors of high research performance used in prior studies did not take into account the particular

Table 6 PD interviewees across gender and rank measures

Gender	
Male	10
Female	2
Rank	
Assistant Lecturer	1
Lecturer	3
Associate Professor	3
Professor	5
Total Number of Interviews	12

challenges of global South researchers. Hence there was a need to check the relevance of predictors from past literature and also to explore any additional context-specific predictors.

In order to do this, a semi-structured interview guide (English language version in Appendix A) was developed based on a combination of past literature on high-performing researchers and on the context of global South research. To reduce the need for extensive travel, interviews were restricted to the four universities in Greater Cairo: Cairo, Ain Shams, Helwan and Benha. Those universities were home to 12 of the 26 PDs identified in the *Determine* phase, all of whom were interviewed along with the heads of the IS departments in each university.[10] Table 6 shows the distribution of the interviewed PDs across gender and rank.

Permission was obtained for the interviews to be recorded so that the transcript could subsequently be analysed to identify common themes, patterns and explanations. In all, interview data was coded into nine main categories of potential differentiators of PDs:

- ***Previous education****:* A number of PDs mentioned that they obtained their PhD degrees from global North universities, explaining how it had a fundamental role in changing how they viewed and practised scientific research.
- ***Research motives:*** PDs were seen as having different motives and drivers for conducting and publishing research. Getting a promotion was definitely one of those drivers, especially for early-career researchers. However most of the PDs mentioned motives related to international recognition, staying competitive and how they enjoy the process of publishing research. A number also mentioned that their research satisfies a personal interest they have and publishing in it adds to their satisfaction.
- ***Research type:*** Almost all of the interviewed PDs worked on applied and experimental research, while only two focused mainly on theoretical research. In terms of topics, all that stood out were areas avoided by most PDs: only one did research in the management of information systems, and only two showed interest in research that had broader social and developmental impact.
- ***Research strategies:*** A number of PDs said that they were more inclined to do incremental research i.e. building upon previous work; one of them saying *"I do not innovate by finding new problems; I innovate by finding new ways and methods to solve a well-established problem"*. Applying for research funding from schemes like the Euro-

[10] An initial observation was that a number of PDs ($n=5$ of 26) were also department heads; a predictor that was added to the survey questionnaire.

pean Region Action Scheme for the Mobility of University Students (ERASMUS+) and the German Academic Exchange Service (DAAD) was also mentioned by a number of them. Another strategy that was mentioned by most of the PDs is reaching out to foreign authors to conduct collaborative research with them. When they were asked why they do that, their answers varied. Some said that it ensures better access to resources; with sample statements including *"When a paper is accepted in a conference or a journal, their universities can fund their travel expenses or pay for journal submission fees"*, and *"my research requires computing facilities that are hard to provide here and my research partner in Canada has access to those facilities"*. Such resources could include complementary skill sets: *"he [*foreign collaborator*] is good at scientific writing and in the statistical analysis of results and I'm good at coming up with ideas and in the interpretation of results, we were a great team"*. Another view was that foreign authorship assisted publication: *"foreign authors increase the chances of paper acceptance and reduce the time of acceptance drastically"*.

– ***Publication strategies:*** PDs were aware of the importance of publishing in indexed journals and conferences (such as those indexed in Scopus or ISI), stating it as a major criterion in selecting where to publish. Within this overall focus, the interviewees' strategies could be grouped into three main categories: (a) Publishing in international indexed journals and in international indexed conferences. These interviewees saw top-tier international conferences (e.g. the Very Large Databases series) being as prestigious as journals rated Q1 and Q2 in the SCImago Journal Ranking (SJR)[11] in addition to providing very high paper visibility. (b) Publishing in international indexed journals and in local indexed conferences; the majority of the interviewed PDs fell into this category due to financial constraints that limited their ability to attend international conferences. They used conferences as a medium to retain ownership, promote, refine and develop their research ideas before submitting extended versions of papers to journals. *"The journal paper should have at least 30% expansion to the work in the conference paper"* said one of the PD researchers. (c) Publishing in international indexed journals only: this group could not afford travel to international conferences and could not find any value in publishing in local conferences stating, for example, that *"Journal papers are more respected in Egypt"*. A number of PDs also stressed the importance of the publisher, indicating that they noticed that there are certain publishers which provided very high visibility to their papers which lead to higher citation. One of them said *"I started to focus on publishers instead of journals, because a strong publisher will make a journal powerful but a strong journal without a strong publisher, will die … any paper published by Elsevier will have great visibility, even if it is a new journal … I would rather publish in a Q4 journal published in Elsevier than publish in a Q3 journal published somewhere else"*.
– ***Research direction:*** A few PDs mentioned that they usually trace their own citations to see what other authors are saying about their work, and to see how their research is evolving, to get ideas for future work. One of the PDs also mentioned that he follows publishers like ACM, IEEE, Springer and Elsevier to keep informed of new conferences, and thereby indicating hot topics, *"if ACM decided to do a conference on recom-*

[11] "The SCImago Journal Rank (SJR) indicator is a measure of the scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where the citations come from." ('SCImago Journal Rank' 2020).

mender systems, this implies that recommender systems are picking up or will become a hot topic".

- ***Writing their papers:*** Interviewees were asked about factors that increase the chances of paper acceptance, and almost all of them agreed on the importance of the paper structure and presentation. One of them even stated that *"a well-written average idea is more likely to be accepted than a poorly-written great idea."* Interviewees also mentioned the importance of issues including: showing the contribution clearly and frequently in the paper, mathematical and theoretical validity, recency of references, use of formal and scientific writing, mastery of the English language, and a self-contained abstract showing clearly the contribution, the method used and the study findings or results. They were also asked about the process of writing a paper but nothing seemed unusual in that regard. Finally, they were asked about factors that could increase paper citations. The answers varied but, again, publishing with a reputable foreign author was mentioned as a key factor in attracting citations. One of the PDs said *"When you are publishing with a trusted author in the field, people feel comfortable to cite his work".* Publishing in top journals and conferences was also mentioned several times although a few PDs stated that some of their most highly cited work was published in local journals and conferences. PDs also mentioned the importance of publicising research work either through sending emails to researchers they thought would benefit from a paper or through making it available on academic networking sites like ResearchGate and Academia. A number of PDs mentioned the role of the title in attracting citations and one stated that *"I always try to borrow the same keywords used in the titles of the highly cited related papers, because when they search for them, mine will appear."* Some of them also mentioned that survey papers in new fields guarantee high citation and the same for publishing in hot topics at the beginning of their hype cycle.
- ***Research challenges:*** PDs were asked about their research challenges and how they were able to overcome them. A number of PDs mentioned that they encounter difficulties in choosing the right journal for their publications. Only one PD suggested a way to overcome this, which was through use of online journal finder tools. PDs mentioned the language barrier especially with the students they supervise; one PD stated that he uses the paid-for language editing services provided by Elsevier. Another PD said *"I asked one of my students to stop his PhD for 3 months just to enhance his English writing by taking courses".* Some of them mentioned having overseas contacts that proofread their work. The prolonged time from submission to acceptance was repeatedly mentioned by PDs as a major challenge, especially when the topics they want to publish are time sensitive. In such cases, they would resort to conferences for early communication of those ideas. Finally, all of them mentioned that the limited financial support they receive from the university—to attend conferences and to publish in open access journals—is a major challenge. The alternative was to self-finance their travel and publishing activities or to seek support from funding agencies. Some PDs also mentioned that they overcame this challenge by having as co-authors their former supervisors from their foreign PhD-granting universities, which would sometimes cover conference travel expenses and journal submission fees.
- ***Research skills development:*** A number of PDs mentioned taking scientific and technical writing courses. One of them also mentioned the importance of formal writing saying that *"I paid a lot of attention to learn the formal way of writing; you learn it from observation, trial and error".* Indeed, a number mentioned observing highly cited papers written by top authors to see how it is written and structured as a means to enhancing their writing skills. A lot of PDs mentioned using tools like Grammarly *"A*

Table 7 Distribution of the survey responses from PDs and NPDs across gender and rank measures

	PDs	NPDs
Gender		
Male	17	28
Female	3	42
Rank		
Assistant Lecturer	2	28
Lecturer	5	18
Associate Professor	4	13
Professor	9	11
Total number of Responses	20	70

> *number of powerful researchers I know recommended this tool"* said a PD, and Latex, *"I could spend a whole day rearranging figures on Word while it takes me a few minutes using Latex".*

In summary, the interviews led to the identification of potential patterns in attributes, attitudes and practices amongst PDs: some similar to those from earlier studies but a number that had not previously been identified. Some of these—such as use of keywords from titles of highly cited papers—were practices identified by only one interviewee that, while interesting, were not seen to warrant inclusion in the survey questionnaire. But those appearing repeatedly—studying for a PhD abroad, taking scientific and formal writing courses, publishing with foreign reputable authors, etc.—were incorporated into the Stage 2 questionnaire.

Stage 2: Survey

The primary objective of this stage was to validate the findings from the earlier parts of the methodology and to identify predictors of PDs that are significantly different from those of NPDs. An online survey questionnaire (English language version in Appendix B) was developed based on the review of related work (see e.g. Table 1), amended in light of the findings from Stage 1. A message and link to the survey was sent to the whole sample of PDs and NPDs ($n=203$) in the 11 universities, including PDs who were interviewed in the previous stage. In total, 90 survey responses were collected: 20 from PDs and 70 from NPDs yielding an overall response rate of 44%.

Survey responses ($n=90$) were entered and analysed with the use of the statistical software R Studio. Table 7 shows the distribution of the sample of PDs and NPDs across gender and rank. 70% of the respondents held PhD degrees (i.e. were lecturer rank or above) and 30% had MSc degrees (i.e. were assistant lecturers) and the responses came evenly from males and females. It also shows pronounced gender imbalance within the PDs and how seniority still plays a role in being a PD, despite incorporation of measures of performance that would control for that (e.g. hi-index).

Feature selection The survey tool had 38 questions covering researcher attributes such as gender and rank; attitudes such as what motivates them to publish research; and practices such as the type of research collaborations they engage in. After transforming categorical variables into dummy variables, the final sample had 90 observations and 185 variables. The next step was to build a predictive model to identify significant predictors of PDs among those 185 variables. But before building such a model, it is important to undertake two necessary steps. The first is to reduce complexity through feature selection i.e. selecting the predictor/independent variables that will be used to predict the dependent variable which in our case was a binary variable with the value of 1 for PD researchers and 0 for NPD researchers. And the second is to identify and address potential issues of multicollinearity.

Feature selection was done by running a simple univariate logistic regression (i.e. relation of the dependent variable with each predictor, one at a time) and then including only predictors that met a certain pre-set cut-off for significance to run in the multiple regression. For the simple regression a cut-off of $p < 0.1$ was used since its purpose was to identify potential predictor variables rather than test a certain hypothesis (Ranganathan et al., 2017). A stricter cut-off point ($p < 0.05$) was then used in the multiple logistic regression to identify significant predictors of PD. Out of all the explored predictors, 23 were identified as potentially significant predictors as shown in Table 8. Predictors derived from the interviews in Stage 1 are denoted by "(i)".

Following the construction of the simple univariate regression models, we proceeded to check multicollinearity. Specifically, the strength of the association between all possible pairs of the 23 predictors was determined using the Spearman rank correlation (for numeric variables), Chi square (for categorical variables) and Anova (for pairs involving one categorical and one numerical variable) implementations in the *cor* function of the caret package.[12]A lot of the predictors identified were significantly correlated with each other, which would be problematic when jointly used in a multiple logistic regression, creating what is referred to as the separation problem (Mansournia et al., 2018). In practice, stepwise regression can be used to overcome this issue but the problem with this approach is that it might eliminate important predictors that are correlated with the response variable and are important for the user. Partial least squares (PLS) regression allows us to retain in the model all the predictors that have a strong explanatory power. For that reason, it was our preferred method for multiple regression. This is further explained in the following section. PLS regression is a technique that reduces the predictors to a smaller set of uncorrelated components or latent variables and performs least squares regression on these components, instead of performing it on the original predictors. PLS regression is particularly useful when there are more predictors than observations and when the predictors are highly collinear (Abdi, 2003).

Multiple regression PLS regression is a technique that reduces the predictors to a small set of uncorrelated components and performs regression on those components instead of performing it on the predictors (Tobias, 1995). The plsRglm package[13] implements the PLS regression for generalised linear models which is an extension of the classical PLS regression introduced by Bastien et al. (2005). We also used the *cv.plsRglm* function to identify the ideal number of components to retain in a ten-fold cross-validation ($k = 10$), using six components ($n = 6$) as the maximum number of components to try with each group or fold. After plotting the results of the cross-validation, we decided to retain only two components based on the mis-classed criterion (i.e. components achieving the least number of misclassified observations) and the non-significant predictor criterion (i.e. components that had

[12] https://cran.r-project.org/web/packages/caret/caret.pdf.

[13] https://cran.r-project.org/web/packages/plsRglm/plsRglm.pdf.

Table 8 Estimated coefficients of significant predictors resulting from the simple logistic regression $***P<0.001$; $**P<=0.01$; $*\ P<0.05$; '.' $P<0.1$. Only those predictors with $P<0.1$ have their estimates presented in the table

Predictor	Estimates
Gender	
Male	2.012**
Female	
Marital status	
Married	
Single	
Number of children	
Last Degree	
PhD	
MSc	
Number of years to complete PhD degree	
1–2 years	
2–3 years	
4–5 years	
Foreign PhD degree (i)	0.0938
Faculty rank	
Ass. Lecturer	
Lecturer	
Ass. Professor	
Professor	1.6946**
Department chair (i)	1.1632*
Supervision	
Undergraduate groups	
MSc students	0.11991*
PhD students	0.18566*
Admin & teaching load	
Publication financial support	
Research grants	1.0468
College scholarship	
Travel funds	1.2417*
Conference attendance	
Department climate	-0.5078
Language school	
Work outside university	
Hours of work outside	
Research motivation	
I publish research to get a promotion	
I publish research for international recognition	
I publish research to stay competitive	
I publish research because I enjoy it (i)	
Type of research	
Studies suggesting new ways of viewing/implementing information processing systems e.g. theories, new architectures, new frameworks, ontologies, network protocols	

Table 8 (continued)

Predictor	Estimates
Research involving the creation of new information-processing systems	
Research involving the creation and evaluation of tools, formalisms, techniques/methods to support existing information processing systems	
Research on social and economic issues related to information processing systems (Including studies of the social and economic impact of information systems, ethical issues, changing views of humanity, etc.)	
From where do they get their research ideas	
Publications of researchers I follow on academic platforms (e.g. Google Scholar)	−0.9445
Live or recorded webinars (e.g. IEEE webinars)	
Papers citing my work (i)	
Predictor	Estimates
Conference attendance	
Future work section of papers	
From where do they get their research ideas	
Research strategy	
I prefer to do radical research	−0.9673
I prefer to do incremental research (i)	
I prefer to map out broad features of important new areas (i)	
I prefer to probe deeply and thoroughly in narrow areas	
I prefer research which looks for immediate solutions to real life problems (e.g. social problem or industry need)	
I prefer purely theoretical research (i)	
I prefer to carry out research work pretty much on my own	
I prefer to carry out research within a research team	
I prefer long-term projects to short-term ones	
I prefer short-term projects to long-term ones	
Research collaborations	
Doing research with academics in other universities in Egypt	0.4220*
Doing research with academics in other departments in my university	0.4482*
Doing research with academics overseas (i)	0.5985**
Where do you publish research	
Journals indexed in Scopus	
Journals indexed in Thomson ISI (Clarivate Analytics)	
International Conferences with Proceedings indexed in Scopus	
International Conferences with Proceedings indexed in Thomson ISI (Clarivate Analytics)	
Local Indexed Conferences	
Non-indexed Journals	
Non-indexed Conferences	
Factors affecting journal selection	
The publisher of the journal (i)	
Number of issues per year	−0.413*
Editorial board	
Journal fees	
Journal impact factor	
SCImago Journal Rank	

Table 8 (continued)

Predictor	Estimates
Factors affecting acceptance at top conferences and journals	
Presentation/Structure of the paper (i)	
Reputable co-authors (i)	
Strength of the authors' affiliated universities (i)	
Recency of references (i)	
Including references from the targeted journal/conference proceedings	
Technical depth (i)	
Significance of the contribution (i)	−0.9185*
Theoretical foundation (i)	
Previous publications in the targeted journal/conference	
Primary reason for presenting in conferences	
Interaction with peers and getting feedback	
To be known among my research community	
To publicise my research and attract paper citation	
To gain knowledge about new research areas and trends	
To search for academic posts, possible grants and project collaborations	
Research platforms they use	
Academia	
Semantic Scholar	1.5976
ResearchGate	
Google Scholar Profile	
Arxiv	
DBLP	
ORCID	
Researcher ID	
ACM	
Publication strategies	
When I start in a new area of research, I prefer publishing the first paper by myself and then including other authors in the following papers (i)	
I publish part of my research work in a conference before publishing it in a journal (i)	
I submit my paper in top conferences (knowing it might get rejected) before submission in journals to get useful feedback/review (i)	
I submit papers in workshops of top conferences (i)	
I publish papers extending/based on the graduation projects of my last year (undergraduate) students (i)	
I publish papers with foreign reputable co-authors (i)	0.4183*
I publish papers in highly ranked journals/conferences (i)	
I publish papers with top publishers (e.g. Elsevier) (i)	
I add my papers in academic networking platforms (e.g. ResearchGate) (i)	
I send hard or soft copies of my paper to researchers in the same field once its published (i)	
I publish papers in specialised journals	
I publish papers in multidisciplinary journals	
I publish papers with new ideas, models or frameworks without experimentation	
I publish papers with new ideas, models or frameworks with experimentation and results	
I publish papers with tools or datasets	

Table 8 (continued)

Predictor	Estimates
I publish papers in open access journals	
Research challenges	
Motivation to carry out research is a challenge	
Finding the right journal/conference for my paper is a challenge (i)	
Lack of financial support needed for attending conferences is a challenge (i)	
Proficiency of written English is a challenge (i)	
Formal/Scientific writing is a challenge (i)	
Time from submission to acceptance in a journal is a challenge	
Insufficient time because of teaching/admin commitments is a challenge	
Overcoming challenges	
I use journal finder online tools (i)	
I pay for proofreading and editing services for my paper (i)	
I seek external funding agencies (e.g. ITIDA, ASRT, TIEC) to cover the costs of travelling to attend conferences (i)	
I use the financial support provided by the university to cover my travel and publication fees	
I apply for research grants (e.g. Erasmus) (i)	1.2135*
I establish research teams overseas (i)	1.9588**
Research tools	
Grammarly (i)	
Reference managers (e.g. Mendeley)	
Latex (e.g. Sharelatex) (i)	
Enhancing publication quality	
Observing highly cited papers to see how they are written and structured (i)	
English writing courses (i)	1.1386
Scientific writing/Formal writing courses (i)	1.3458*
Technical courses related to the field (i)	
Using a graphic designer to represent results in an attractive manner (i)	
Sending papers to friends/relatives for proof editing (i)	
Checking paper citations	−1.0756
Why do they check paper citations	
To check the geographical distribution of the citing papers	
See the impact of the paper after removing self-citation	
Get ideas on future research areas / improvement areas (i)	

significant predictors). Cross-validation with a 70–30 split was used in each of ten training datasets with their test data pairs to calculate the model's prediction accuracy and its AUC i.e. area under the receiver operating characteristics (ROC) curve[14]; which is considered a good metric for evaluating the performance of binary classifiers. Across the ten folds, the model resulted in an average accuracy of 0.78 and an average AUC of 0.70. Significant pre-

[14] ROC curve is plotted with true positive rate (TPR) against the false positive rate (FPR) where TPR is on the *y*-axis and FPR is on the *x*-axis.

Table 9 Component estimates along with the loadings of their significant predictors and their predictive power in a ten-fold cross-validated PLS model

Significant variables	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	Avg
Component 1	2.2	1.08	2.49	1.13	1.51	1.26	1.45	1.39	1.28	1.17	1.50
Male	0.28	0.38	0.25	0.24	0.40	0.31	0.28	0.30	0.32	0.31	0.31
Professor	0.32	0.32	0.31	0.29			0.30	0.30		0.33	0.31
Foreign PhD degree		0.26								0.26	0.26
Department chair	0.28		0.31	0.29							0.29
Number of supervised MSc students	0.37	0.27	0.32	0.33			0.30	0.35			0.32
Number of supervised PhD students	0.34		0.34	0.33			0.31				0.33
Received research publication grant	0.23										0.23
Received travel funds			0.24				0.25				0.25
Rating of department climate		−0.06					−0.07	−0.05		−0.12	−0.08
I prefer to do radical research that suggests new models frameworks methods and architecture that were not implemented before	−0.01						−0.02	−0.06			−0.03
The get ideas from publications of researchers they follow on academic platforms e.g. Google Scholar			−0.22								−0.22
Doing research with other academics in other universities in Egypt	0.28		0.23								0.26
Doing research with academics overseas	0.34		0.25	0.34	0.36	0.35	0.34	0.34	0.33	0.34	0.33
Doing research with academics in other departments in my university	0.21		0.21			0.22					0.21
Number of issues per year		−0.29			−0.2		−0.21		−0.21		−0.23
They believe that the significance of the contribution increases the chances of acceptance of a paper in a journal			−0.1			−0.27			−0.24		−0.20
I publish papers with foreign reputable authors	0.33		0.26			0.37	0.33		0.34	0.34	0.33
I apply for research grants				0.25							0.25
I establish research teams overseas	0.24	0.23	0.22	0.24	0.25	0.20	0.24		0.24		0.23
I took English writing courses			0.13					0.18		0.178	0.16
I took Scientific or Formal Writing courses				0.22		0.30			0.25		0.26
They have a profile on Semantic Scholar				0.09							0.09
Component 2	2.4	1.13	2.68	1.46	1.33	1.35	1.29	1.29	2.02	1.48	1.64

Table 9 (continued)

Significant variables	*K* = 1	*K* = 2	*K* = 3	*K* = 4	*K* = 5	*K* = 6	*K* = 7	*K* = 8	*K* = 9	*K* = 10	Avg
I prefer to do radical research that suggests new models frameworks, methods and architecture that were not implemented before	−0.36	−0.49			−0.42		−0.02				−0.32
Rating of department climate				−0.52							−0.52
Prediction Accuracy	0.78	0.73	0.86	0.69	0.91	0.86	0.69	0.69	0.82	0.73	0.78
AUC	0.79	0.60	0.80	0.52	0.88	0.80	0.57	0.63	0.72	0.66	0.70

dictors (p values < 0.05) of the two components we retained are presented in Table 9. The table also shows the estimates of the two retained components across the ten folds with an average coefficient of 1.5 for component one and 1.64 for component two.

In the analysis shown in Table 9, significant differences between PDs and NPDs emerged, covering attributes such as gender (PDs were mainly males) and rank (a large number of PDs were professors who are also department chairs). However, it is hard to tell if the latter is a cause or an effect. This is because becoming a department chair in the higher education system in Egypt is mainly based on years of experience rather than academic merit. Additionally, department chairs get the biggest share of MSc and PhD student supervisions, which are also significant predictors of PDs. Having a larger number of students implies a larger number of publications and citations, hence better citation metrics. Differences related to practices included the ways PDs developed their skills, such as taking scientific writing courses and English writing courses and travelling abroad for their PhD degrees. It was also strongly evident that PDs publish more papers with foreign authors. This links to a key difference that persistently appeared, with a relatively high loading, which was doing research with academics overseas. Other collaborations such as doing research with academics in other universities in Egypt and in other departments in the same university, were also significantly higher among PDs but they were not as strong as collaborations overseas. Practices that were found to be less common among PDs included getting research ideas from publications of researchers online, and surprisingly, doing radical research that suggests new models, frameworks, methods and architectures that were not implemented before; which is somewhat counterintuitive. Finally, differences relating to attitudes included how the researchers rated the climate of their department: PDs perceived their departments as more hostile and competitive while NPDs viewed departments as more friendly. They also viewed the number of issues as a less important factor when selecting the journals to publish in.

Table 9 also shows that component one included key predictors that are positively correlated with high research performance while component two was able to better capture the direction of variation of two predictors that are negatively correlated with high research performance ("Rating of department climate" and "I prefer to do radical research that suggests new models frameworks, methods and architecture that were not implemented before") and had very small loadings in component one.

We were also interested in developing a model that would exclude non−controllable factors in order to identify transferable practices that could be adopted by other researchers. It excluded gender, rank and being a department chair. Table 10 presents the findings of this model which resulted in an average accuracy of 0.77 and an average AUC of 0.72. The table also shows the estimates of the two retained components across the ten folds with an average coefficient of 1.55 for component one and 1.58 for component two. Reassuringly, this model's predictive power was very close to the average predictive power of the previous model (mean 0.78 and AUC 0.70) despite the exclusion of those significant predictors that had a relatively high loading. This model reinforces the results from the previous model on the importance of international research collaboration since "Doing research with academic overseas", "Establishing research teams overseas" and "Publishing with foreign reputable authors" appeared repeatedly as significant predictors with high loadings across the ten folds. The significance was also evident of supervising more students (MSc and PhD), having a foreign PhD degree, receiving travel funds, and taking scientific or formal writing courses.

Table 10 Component estimates along with the loadings of their significant predictors after excluding gender, rank and role

Significant Variables	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	Avg
Component 1	2.44	1.38	1.69	1.22	1.64	1.41	1.65	1.42	1.31	1.33	1.55
Foreign PhD degree		0.28								0.30	0.29
Number of supervised MSc students	0.42	0.31	0.37	0.35			0.32	0.38			0.36
Number of supervised PhD students	0.38		0.39	0.35			0.34				0.37
Received research publication grant	0.23			0.27							0.25
Received travel funds			0.27				0.27				0.27
Rating of department climate		−0.10					−0.11	−0.09		−0.18	−0.12
I prefer to do radical research that suggests new models frameworks methods and architecture that were not implemented before	−0.02						−0.05	−0.1			−0.06
The get ideas from publications of researchers they follow on academic platforms e.g. Google Scholar			−0.24								−0.24
Doing research with other academics in other universities in Egypt	0.34		0.27								0.31
Doing research with academics overseas	0.39		0.31	0.39	0.44	0.39	0.40	0.42	0.38	0.41	0.39
Doing research with academics in other departments in my university	0.25		0.23			0.30					0.26
Number of issues per year		−0.34			−0.26		−0.24		−0.22		−0.27
They believe that the significance of the contribution increases the chances of acceptance of a paper in a journal			−0.14			−0.31			−0.26		−0.24
I publish papers with foreign reputable authors	0.38		0.32			0.40	0.37		0.38	0.41	0.38
I establish research teams overseas	0.28	0.27	0.26	0.29	0.32	0.23	0.27		0.28		0.28
I took English writing courses			0.15					0.21		0.21	0.19
I took scientific or formal writing courses				0.25		0.33			0.28		0.29
They have a profile on Semantic Scholar				0.13							0.13
Component 2	2.20	1.31	1.92	1.47	1.31	1.59	1.40	1.15	1.88	1.57	1.58
I prefer to do radical research that suggests new models frameworks, methods and architecture that were not implemented before	−0.30	−0.52						−0.48			−0.43
Rating of department climate				−0.51							−0.51
Prediction Accuracy	0.74	0.78	0.82	0.74	0.86	0.78	0.73	0.65	0.82	0.78	0.77

Table 10 (continued)

Significant Variables	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	Avg
AUC	0.82	0.74	0.77	0.55	0.85	0.69	0.66	0.65	0.72	0.74	0.72

Stage 3: Publication analysis

While Stages 1 and 2 were focused on the identification of individual-level predictors of PDs, Stage 3 is focused on the identification of publication-level predictors. In other words, in this stage, the unit of analysis is the paper instead of the researcher. The general motivation for publication-level analysis was noted above but, in addition, some of the significant predictors identified in the previous stages required validation that could only be done through publication analysis. For instance, while PDs mentioned publishing with foreign authors and this was established as a significant predictor, it was not possible to validate this practice and quantify its prevalence within PD publications, relative to NPD publications, without analysing the actual papers. The same was true for the number of authors, the choice of publication outlet, the frequency of research collaborations, etc. In summary, the objective of Stage 3 is twofold: the first is to quantify and validate some of the findings of Stages 1 & 2 through the analysis of the researchers' publications. The second is to identify additional predictors of PDs that can be derived directly from their publications.

In this stage we analysed the publication corpus of PDs versus the publication corpus of NPDs. We defined a PD publication as a paper that has at least one PD author from the 26 high-performing researchers identified in the *Determine* Phase. In contrast, an NPD publication is defined as a paper with at least one NPD author but where none of the authors is a PD. By doing so, we were able to create two mutually exclusive corpora to capture distinguishing characteristics of each. The papers were collected from the Scopus database using the Rscopus[15] data package which links R Studio to the Scopus database API interface. For every researcher in the study population ($n=203$) a Scopus ID was identified manually through the Scopus advanced search tool form on the website. This ID was then used to retrieve all the possible information associated with their publications including co-authors, co-author affiliations, abstracts, keywords and titles. For consistency purposes, we excluded publications not having the Egyptian university affiliation and/or produced while researchers were abroad (e.g. during overseas PhD study) or produced after 2018 (since the citation metrics upon which we selected the PDs were calculated in 2018). In total, 991 publication records were extracted for PDs and 677 publications were extracted for NPDs. Those publications were further reduced to 876 unique publications (in total), after excluding duplicate publications and publications that did not have abstracts on Scopus. The final corpus of papers included 392 PD papers and 484 NPD papers. Skews consistent with the early-discussed Pareto distribution of performance (O'Boyle & Aguinis, 2012) were immediately reflected: PDs make up 13% of the study population but contributed to the creation of 48% of the publications. Those 392 papers were cited 3210 times while NPD papers were cited 1810 times.

We proceeded to examine the three types of paper-level predictors of citation rates used in previous studies: (1) "extrinsic" features of the paper that are not directly related to its content (e.g. paper length, number of authors, etc.); (2) "intrinsic" or content-related features such as the topics covered in the paper; and (3) the publication "outlets" of the paper (e.g. conference or journal paper, journal SJR, etc.). The papers' "extrinsic" features were extracted for each paper using the Scopus API functions. Paper "intrinsic" features were extracted from the paper title, abstract and keywords using a topic modelling technique that is further explained in a subsequent section. The publication "outlet" features were

[15] https://cran.r-project.org/web/packages/rscopus/rscopus.pdf.

Table 11 Paper and publication outlet features used as predictors

Predictor	Feature type	Source
Paper length (number of pages)	Paper-Extrinsic	(Stewart, 1983) (Peters & van Raan, 1994) (Van Dalen & Henkens, 2001) (Walters, 2006) (Davis et al., 2008) (Haslam et al., 2008) (Lokker et al., 2008) (Peng & Zhu, 2012) (Onodera & Yoshikane, 2015) (Elgendi, 2019)
Number of authors	Paper-Extrinsic	(Peters & van Raan, 1994) (Van Dalen & Henkens, 2001) (Walters, 2006) (Davis et al., 2008) (Haslam et al., 2008) (Lokker et al., 2008) (Fu & Aliferis, 2010) (Peng & Zhu, 2012) (Didegah & Thelwall, 2013) (Onodera & Yoshikane, 2015) (Elgendi, 2019)
Intranational affiliations (in Egypt)	Paper-Extrinsic	(Lokker et al., 2008) (Fu & Aliferis, 2010) (Didegah & Thelwall, 2013) (Onodera & Yoshikane, 2015)
International affiliations	Paper-Extrinsic	(Didegah & Thelwall, 2013) (Onodera & Yoshikane, 2015)
US affiliations	Paper-Extrinsic	(Van Dalen & Henkens, 2001)
Type of article	Paper-Extrinsic	(Peters & van Raan, 1994) (Van Dalen & Henkens, 2001) (Walters, 2006) (Davis et al., 2008) (Lokker et al., 2008)
Number of figures	Paper-Extrinsic	(Haslam et al., 2008) (Onodera & Yoshikane, 2015) (Elgendi, 2019)
Number of references	Paper-Extrinsic	(Stewart, 1983) (Peters & van Raan, 1994) (Walters, 2006) (Davis et al., 2008) (Haslam et al., 2008) (Lokker et al., 2008) (He, 2009) (Didegah & Thelwall, 2013) (Onodera & Yoshikane, 2015)
Title length	Paper-Extrinsic	(Haslam et al., 2008) (Elgendi, 2019)
Colons in title	Paper-Extrinsic	(Haslam et al., 2008) (Elgendi, 2019)
Abstract length	Paper-Extrinsic	(Lokker et al., 2008)
Paper topics (manual content analysis of popularity and focus)	Paper-Intrinsic	(Peters & van Raan, 1994) (Van Dalen & Henkens, 2001) (Van Dalen & Henkens, 2005) (Lokker et al., 2008) (Peng & Zhu, 2012) (Hu et al., 2020)
Open access	Publication Outlet	(Davis et al., 2008)
Publisher	Publication Outlet	This was examined based on insights from the interviews
Journal SJR	Publication Outlet	(Van Dalen & Henkens, 2001) (Walters, 2006) (Davis et al., 2008) (Haslam et al., 2008) (Fu & Aliferis, 2010) (Peng & Zhu, 2012) (Didegah & Thelwall, 2013)
Publication type (conference paper vs journal article)	Publication Outlet	(Fu & Aliferis, 2010)

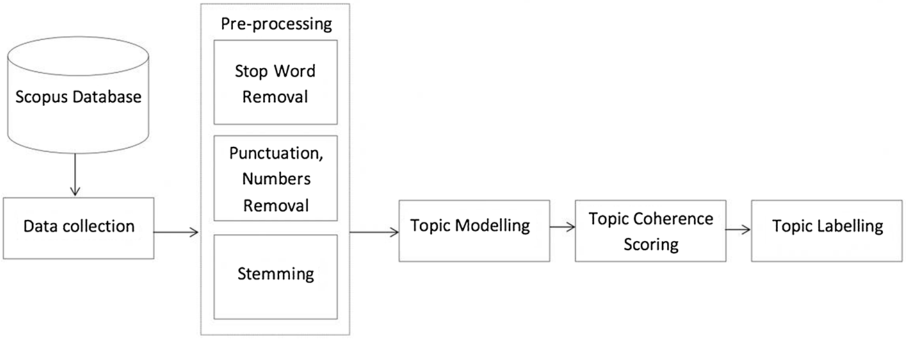


Fig. 4 Topic extraction process, developed from Mahanty et al. (2019)

obtained using the sjrdata[16] package which contains data extracted from the SCImago Journal & Country[17] open data portal.

Publication predictors

There is a substantial body of research on publication-level predictors of citation rates. In this study we selected several of those predictors based on three main conditions. The first is their relevance to measuring or validating findings from the previous two stages. The second is their relevance to the issues previously raised in the literature relating to Southern researchers. Finally, we excluded features that are difficult to ascertain or would require manual validation (e.g. gender of authors), subjective features (e.g. title attractiveness) or features that would require extensive additional computation or additional measure development e.g. internationalisation of journals. Table 11 presents the different publication features that were used as predictors, and references studies that used them as potential predictors. How paper topics (i.e. paper intrinsic features) were identified and converted into a feature space is explained in the following section.

Topic extraction

The objective of this analysis is (i) to identify the various research topics of the study population, and (ii) to develop for every paper a vector representing the distribution of the content across the topics identified. The author topics resulting from this analysis were used as the paper "intrinsic" features in the regression analysis (presented below) in order to explore if there are certain topics that are associated with PD performance and vice versa. Figure 4 explains the steps involved in the topic extraction process. Abstracts are considered a compact representation of the whole article, so we used them as a proxy of the paper content to identify the topics of research. We started by extracting publication data for each author from the paper corpus (876 unique abstracts). Standard text mining pre-processing steps were applied on the entire corpus of abstracts such as lowercasing the corpus, removal of standard stopwords (e.g. a, an, and, the), stemming of terms to remove pluralisation or

[16] https://github.com/ikashnitsky/sjrdata.

[17] https://www.scimagojr.com/.

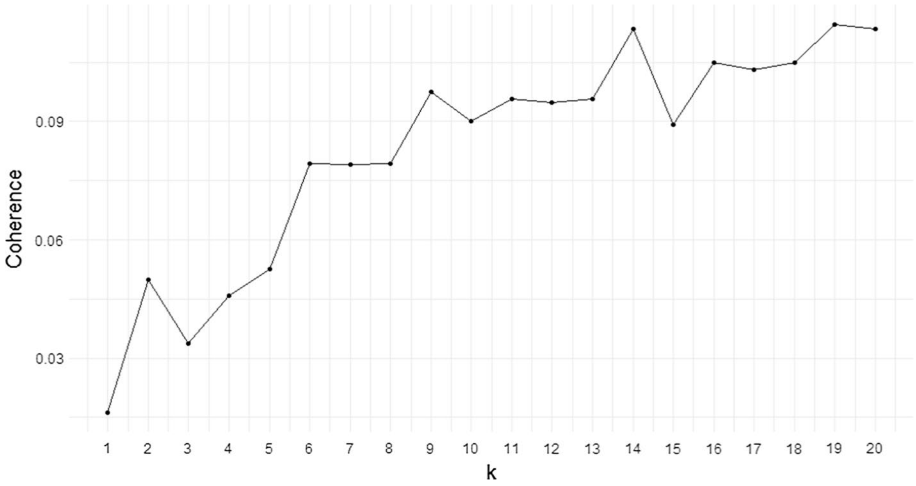


Fig. 5 Topic coherence scores

other suffixes and to normalise tenses. Additional pre-processing steps involved removing numbers, special characters and white spaces (Kao & Poteet, 2007; Mahanty et al., 2019). An unsupervised topic modelling technique called the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) was used to identify the topics within the abstracts corpus. The basic idea of LDA is that articles will be represented as a mixture of topics, and each topic is characterised by a distribution over words. LDA was applied over the entire corpus to identify topics and calculate the probability distribution across topics for each document.

Automatic topic coherence scoring: To develop the LDA model, we need to have a predetermined value for the number of topics (k). A small number of topics can lead to very generic topics and a large number can result in the generation of overlapping and non-comprehensive topics. Hence, we decided to calculate automatically the topic coherence (the degree of semantic similarity between high scoring words in the topic) at every k from 1 to 20, and established that $k = 19$ achieved the highest topic coherence score as shown in Fig. 5.

Topic labelling: Since the labelling of the topics is not done automatically by LDA, we assigned for every topic a relevant label based on the abstracts and keywords of articles with a probability > 90% of falling into that topic. We then validated those manually-generated labels by checking if their terms were automatically generated in the most frequent words within that topic. Table 12 summarises the topics identified, the topic labels assigned and the most frequent keywords.

Time series analysis: We were also interested in visualising topic prevalence over time for the PDs and NPDs separately, in a similar way to the analysis presented in the study by Mahanty et al. (2019). This was done by calculating the mean topic proportion per year for the PD corpus and in the NPD corpus as shown in Fig. 6. The first finding was that NPDs had a longer publication span starting in 1988 while PDs had a shorter publication span starting in 1993. However, for better visualisation we used the same chronological scale for both groups (2002–2018). There were topics, such as Classification Models, where PDs were early movers and then they were followed by NPDs. We can also clearly see the prevalence of Expert Systems and GIS-related topics in the PD corpus in comparison to the NPD corpus, where there is more prevalence of Neural Networks and Business Process

Table 12 LDA generated topics with their corresponding coherence scores and most frequent terms

Topics	Labels	Coherence	Most frequent terms
t_1	Neural Networks	0.361	Neural, network, neural_network, detect, fast, input, time, result, imag, domain, weight, frequenc, comput, paper, networks, normal, neural_networks, frequenc_domain, number, present
t_2	Distributed Database Management Systems	0.068	Data, databas, queri, time, system, propos, process, cloud, distribut, big, result, paper, perform, stream, improv, big_data, storag, update, increas, effici
t_3	Multilevel Programming	0.092	Problem, algorithm, optim, propos, solv, object, solut, model, paper, program, level, perform, multi, approach, result, function, genet, genet_algorithm, fuzzi, linear
t_4	Data Mining	0.057	Algorithm, propos, graph, cluster, method, similar, protein, paper, object, show, base, data, index, approach, effici, structur, comput, present, mine, mani
t_5	Information Retrieval	0.191	Semant, web, arab, search, languag, user, text, retriev, algorithm, inform, propos, document, content, paper, rank, result, extract, generat, approach, model
t_6	Cloud Computing	0.089	Eecur, data, propos, encrypt, scheme, privaci, key, system, attack, user, imag, access, share, protect, secret, cloud, inform, watermark, digit, util
t_7	Networks	0.222	Rout, network, system, protocol, node, propos, base, mobil, blood, biometr, time, traffic, ad_hoc, hoc, keystrok, ad, method, vessel, packet, user
t_8	Web Services	0.103	Servic, cloud, mobil, locat, comput, provid, base, communic, propos, web, services, web_servic, user, applic, cloud_comput, approach, paper, system, integr, cost
t_9	Wireless Sensor Networks	0.157	Network, algorithm, sensor, node, optim, energi, cluster, propos, model, wsn, protocol, wireless, data, power, sensor_network, effici, differ, wireless_sensor, time, springer
t_10	Face Detection	0.069	Data, iot, face, neural, process, detect, test, result, comput, time, paper, phase, propos, approach, human, neural_net, net, neural_network, thing, cooper
t_11	Expert Systems & GIS	0.147	Decis, gis, make, govern, system, evalu, select, factor, inform, studi, develop, process, problem, criteria, research, success, spatial, decis_make, project, solv
t_12	Product Service Systems & Mobile Based Applications	0.046	Learn, technolog, system, agent, busi, framework, paper, complianc, student, differ, present, smart, mobil, process, manag, monitor, bas, support, communic, environ
t_13	Clinical Decision Support Systems	0.086	Case, ontolog, base, system, medic, knowledg, bas, propos, cbr, data, health, fuzzi, domain, rough, result, case_bas, model, standard, set, rule
t_14	Hardware Systems	0.084	Imag, result, springer, part, optim, nois, signal, state, structur, differ, model, paper, quantum, low, high, method, part_springer, rate, depth, error
t_15	Information System Development Methodologies	0.082	Test, softwar, model, system, develop, propos, case, qualiti, approach, requir, paper, test_case, generat, process, design, perform, effort, autom, engin, provid

Table 12 (continued)

Topics	Labels	Coherence	Most frequent terms
t_16	Business Process Management & Process Mining	0.003	Data, process, approach, model, differ, integr, techniqu, mine, busi, event, exist, paper, work, propos, organ, rule, propos_approach, approaches, analysi, sourc
t_17	Classification Models	0.198	Featur, classif, propos, classifi, accuraci, techniqu, differ, appli, result, dataset, select, machin, extract, learn, data, set, approach, algorithm, cancer, support
t_18	Social Network Mining	0.083	User, social, network, recommend, propos, predict, data, base, social_network, detect, opinion, communiti, sentiment, spatial, mine, analysi, system, users, algorithm, show
t_19	Computational Grids	0.037	Predict, schedul, system, level, grid, task, result, hcv, signific, wind, studi, patient, risk, time, propos, cell, comput, introduce, respons, resourc

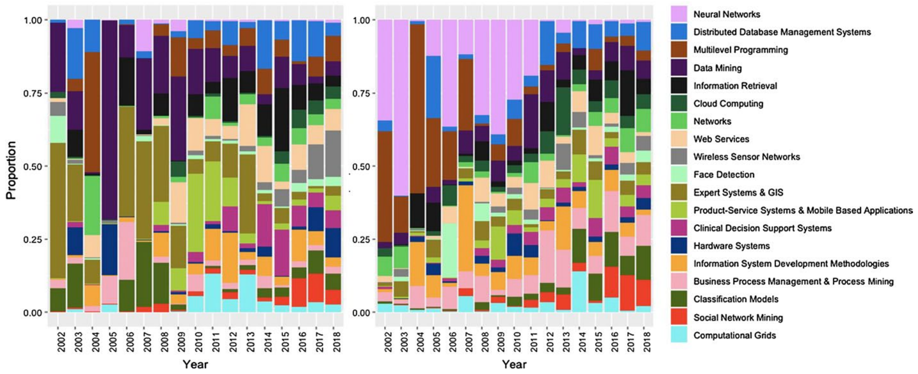


Fig. 6 Topic proportions of PD corpus (left) and NPD corpus (right) over time

Management & Process Mining. There are also topics that had very similar proportions over time for both groups, such as Social Network Mining.

Feature selection

Table 13 presents paper features that were used as predictors of PD-authored papers in the multiple logistic regression model. The features with P values less than 0.1 were the ones identified as potential predictors using the same approach adopted in Stage 2. Those 23 features are the ones we selected for the multiple logistic regression model. Subsequently, we calculated all pairwise correlations, using the *cor* function, and we found that a large number of them were correlated. Hence, consistent with Stage 2, we used the PLS regression for generalised linear models as it allowed us to retain all the potential predictors that could have a strong explanatory power. This is further explained in the following section.

Multiple regression We started by using *cv.plsRglm* function to identify the ideal number of components to retain in a ten-fold cross-validation ($k = 10$), using six components ($nt = 6$) as the maximum number of components to try with each group or fold. After plotting the results of the cross-validation, we decided to retain only three components based on the misclassed criterion and the non-significant predictor criterion. Cross-validation with a 70–30 split was used in ten different samples of test and training datasets to calculate the model's prediction accuracy and its AUC. Across the ten folds, the model resulted in an average accuracy of 0.74 and an average AUC of 0.73. Significant predictors (P values < 0.05) within each of the three components were retained and their loadings are presented in Table 14. The table also shows the estimates of the three retained components across the ten folds with an average coefficient of 1.09 for component one and 0.58 for component two and 0.41 for component three.

In the analysis (results shown in Table 14), significant differences emerged between the PD and NPD corpuses. Regarding the paper extrinsic features, it was clear that papers of PDs had longer titles and abstracts and more pages and references. Their papers had a larger number of authors and affiliations which supports the findings of Stage 2 around research collaborations. While Stage 2 showed that PDs are more likely to publish their papers with foreign authors and establish research teams overseas, this analysis enables us to better understand the type of collaborations by showing us that they were mainly with

Table 13 Estimated coefficients of significant predictors resulting from the simple logistic regression $***P < 0.001$; $**P < = 0.01$; $*\ P < 0.05$; '.' $P < 0.1$. Only those predictors with $P < 0.1$ have their estimates presented in the table

Predictor	Estimates
Paper Extrinsic Features	
Paper Length (number of pages)	0.079***
Number of Authors	0.166***
Number of Affiliations (total number of affiliations)	0.119*
Intranational Affiliations (affiliations within Egypt)	
International Affiliations (affiliations overseas)	
US Affiliations (US university affiliations)	0.691*
Type of Paper	
Conference Paper	−0.90***
Journal Article	0.478***
Review Paper	
Number of References	0.019
Title Length	0.072***
Colons in Title	0.467*
Abstract Length	0.012***
Paper Topics or Intrinsic Features	
Neural Networks	−4.77***
Distributed Database Management Systems	
Multilevel Programming	
Data Mining	
Information Retrieval	
Cloud Computing	
Networks	
Web Services	1.25**
Wireless Sensor Networks	2.49***
Face Detection	−1.26*
Expert Systems & GIS	1.31**
Product Service Systems & Mobile Based Applications	
Clinical Decision Support Systems	1.1582*
Hardware Systems	1.64**
Information System Development Methodologies	
Business Process Management & Process Mining	−4.38***
Classification Models	
Social Network Mining	
Computational Grids	
Publication Outlets	
Open Access	
Journal SJR	0.515**
Publisher	
Wiley	1.63**
Springer	0.63***
ACM	−0.68
Elsevier	0.90***
Inderscience	
IGI Global	
Taylor and Francis	

Table 13 (continued)

Predictor	Estimates
IEEE	

authors from US universities. Additional findings include PDs having more references in their papers and more titles with colons.

Paper intrinsic features, represented by the topics covered in a paper, turned out to be an important distinguishing predictor. It seems that PDs publish fewer papers covering business process management and neural networks in comparison to NPDs. The latter can be linked to an earlier finding from Stage 2 that PDs "do not prefer doing radical research that suggests new models, frameworks, methods and architecture that were not implemented before". One possible explanation could be that neural networks, despite being a cyclical phenomenon, requires radical research whenever there is a recurrence. There were also topics that had much larger coverage in PD papers in comparison to NPD papers, e.g. wireless sensor networks and hardware systems.[18]

As for the publication outlet, we can see that PDs published more journal articles and fewer conference papers; an important predictor that persistently appeared with a high loading. Their preferred publishers were Springer, Elsevier and Wiley, with Elsevier being the one with the highest average loading, supporting the comments of one of the PDs we interviewed who believed that Elsevier journals enable better visibility and impact. PDs were also less likely to publish their papers in ACM. SJR of PD papers was also significantly higher than the SJR of NPD papers, which implies that PD researchers targeted journals with higher quality and impact.

Table 14 also shows that the first component captures variation associated with PD journal articles while the second component appears to relate to PD conference papers, making it possible to infer some of the characteristics of PD publications in either type of outlet. For instance, component one shows that PD journal articles were correlated with a larger number of authors and affiliations, longer abstracts and higher SJR scores. On the other hand, PD conference papers had fewer affiliations and lower SJR values, while still having long abstracts.

Discussion and conclusions

The main motivation of this study was to understand more about research in the global South through a first application of the data-powered positive deviance methodology; a methodology that helped identify and understand those researchers who were able to achieve better research outcomes than their peers. We used a combination of data sources (interviews, surveys and publications) and analytical techniques (PLS regression and topic modelling) to identify predictors of positively-deviant information systems researchers

[18] The significant topics identified in this analysis might not be particularly aligned with the time series analysis conducted earlier (Fig. 6) due to the difference in the unit of analysis. The time series compares topic proportions for papers per year, while the regression analysis looks into topic proportions per paper. The former is cumulative, so some topics might look significant cumulatively, such as data mining, but they happen to be insignificant in the regression analysis when topic proportions are analysed individually for each paper.

Table 14 Component estimates along with the loadings of significant predictors and their predictive power in a ten-fold cross-validated PLS model

Significant variables	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	Avg
Component 1	1.14	1.07	1.01	1.07	1.05	1.12	1.15	1.01	1.15	1.14	1.09
Journal Articles	0.34	0.33	0.31	0.35	0.31	0.32	0.31	0.32	0.31	0.33	0.32
Conference Papers	−0.42	−0.41	−0.38	−0.43	−0.39	−0.40	−0.39	−0.40	−0.37	−0.41	−0.40
US Affiliations	0.16	0.16	0.19		0.18	0.22	0.16	0.20		0.18	0.18
Title with Colon					0.08						0.08
ACM_Papers	−0.13				−0.17		−0.17	−0.17	−0.13	−0.12	−0.15
Springer Papers	0.11	0.11	0.07	0.11	0.12	0.07	0.12	0.07	0.11	0.11	0.10
Elsevier Papers		0.18	0.16	0.19	0.18	0.18	0.17		0.11	0.15	0.17
Wiley Papers	0.18	0.16	0.16	0.15	0.18	0.16	0.10	0.21	0.17	0.14	0.16
t_1 Neural Networks	−0.16	−0.18	−0.2	−0.28	−0.16	−0.13	−0.20	−0.14	−0.25	−0.21	−0.19
t_8 Web Services	0.06	0.07	0.09	0.04	0.06		0.08	0.05	0.08	0.06	0.07
t_9 Wireless Sensor Networks	0.23	0.21	0.19	0.24	0.25		0.25	0.20	0.20	0.23	0.22
t_10 Face Detection	−0.03	−0.08					−0.07	−0.07	−0.06		−0.06
t_11 Expert systems & GIS	0.08	0.06	0.07		0.06		0.09	0.08	0.08	0.08	0.08
t_13 Mobile Based Applications		0.09	0.07	0.07	0.08	0.05	0.11	0.06	0.08		0.08
t_14 Hardware systems	0.12	0.15		0.17	0.12	0.14	0.13	0.10	0.13	0.12	0.13
t_16 Business Process Management	−0.17	−0.19	−0.19	−0.18	−0.21	−0.22	−0.25	−0.21	−0.18	−0.17	−0.20
Number of Authors	0.24	0.27	0.26	0.23	0.23	0.26	0.27	0.24	0.24	0.24	0.25
Number of Affiliations	0.27	0.28	0.25	0.25	0.29	0.29	0.28	0.28	0.26	0.27	0.27
Paper Length	0.39	0.40	0.36	0.39	0.38	0.37	0.38	0.39	0.39	0.39	0.38
Number of References	0.26		0.29						0.38		0.31
Title Length	0.21	0.25	0.20	0.20	0.23	0.22	0.21	0.22	0.16	0.21	0.21
Journal SJR	0.29	0.32	0.29	0.30	0.29	0.29				0.29	0.30
Abstract Length	0.35	0.36	0.36	0.38	0.37	0.38	0.40	0.35	0.36	0.35	0.37
Component 2	0.68	0.59	0.5	0.54	0.54	0.68	0.58	0.50	0.52	0.68	0.58
Journal Article	−0.50	−0.42	−0.45	−0.41	−0.44	−0.44	−0.47	−0.37	−0.38	−0.50	−0.44
Conference Paper	0.47	0.41	0.43	0.41	0.43	0.41	0.46	0.34	0.38	0.47	0.42

Table 14 (continued)

Significant variables	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$	$K=8$	$K=9$	$K=10$	Avg
Elsevier Papers						−0.25					−0.25
t_1 Neural Networks		−0.18									−0.18
t_16 Business Process Management		−0.19		−0.25	−0.27						−0.24
Number of Affiliations	−0.09	−0.13	−0.19	−0.21	−0.14	−0.14	−0.17	−0.10	−0.13	−0.09	−0.14
Paper Length									−0.15		−0.15
Number of References								−0.08			−0.08
Title Length								−0.14			−0.14
Journal SJR	−0.42	−0.36	−0.36	−0.40	−0.39	−0.46	−0.41	−0.36	−0.33	−0.42	−0.39
Abstract Length	0.19	0.20	0.19	0.15	0.17	0.27	0.15	0.17	012	0.35	0.20
Component 3	0.53	0.46	0.41	0.40	0.35	0.35	0.41	0.21	0.43	0.53	0.41
Conference Paper	−0.45	0.43	−0.45	−0.44	−0.47	−0.5	−0.46	−0.36	−0.48	−0.49	−0.37
Prediction Accuracy	0.71	0.76	0.75	0.74	0.75	0.72	0.77	0.71	0.73	0.74	0.74
AUC	0.70	0.76	0.74	0.73	0.75	0.71	0.77	0.70	0.73	0.73	0.73

in 11 Egyptian public universities. We found that PDs, despite representing roughly one-eighth (13%) of the study population, contributed to the creation of roughly half (48%) of the publications and achieved nearly double (1.7x) the total number of citations of NPDs.

Significant Predictors of PDs and their Publications

Starting with the practices of PDs, a reasonably clear picture emerged from the analysis showing that significantly more PDs had travelled to get their PhD degrees from global North universities in comparison to NPDs. They had been part of multi-country research teams and published papers with foreign reputable authors. It seems that studying abroad did not just equip them with the technical know-how and the degree needed to pursue their academic careers, but also helped them establish channels of collaboration with their supervisors and their PhD granting universities, long after they returned to their home countries. This confirms findings from previous studies regarding the importance of international research collaborations (Harris & Kaine, 1994; Kwiek, 2016, 2018; Postiglione & Jung, 2013; Prpić, 1996). Another significant predictor of PDs was their receipt of research grants and travel funds. The findings also show that PDs took scientific/formal writing and English language courses.

The attitudes of PDs were also different to those of NPDs, when it comes to how they perceived their workplaces. PDs viewed their departments as more hostile or competitive while NPDs viewed them as more friendly. This is somewhat at odds with findings from a previous study that high performers preferred working in a relaxed work environment (Harris & Kaine, 1994). Another finding that came as counter-intuitive was that PDs were less inclined to do radical research when compared to NPDs.

In terms of personal attributes, PDs were mainly males and professors, which confirms conclusions from previous studies that identified gender (Kwiek, 2016; Parker et al., 2010; Patterson-Hazley & Kiewra, 2013; Prpić, 1996) and professorship (Kelchtermans & Veugelers, 2013; Kwiek, 2016) as significant predictors of high performance. A significant number of PDs in comparison to NPDs were department chairs at some point in their academic careers (after becoming professors), which is consistent with a number of studies (Kelchtermans & Veugelers, 2013; Patterson-Hazley & Kiewra, 2013). PDs also supervised a larger number of postgraduate students, which would help in generation of publications. The ability to select higher quality students would likely result in higher quality publications and more citations, and, in Egypt, department chairs have more leverage than any other academic staff member in the choice of students they will supervise. More generally, the direction of causality here is questionable. For example, given promotions in Egypt are linked to academic performance, it is likely that some of these factors are impacts of above-average research performance; perhaps more so than causes.

While our work did not set out to test particular theories, we can relate findings to all three of the ideas presented earlier. Consistent with the sacred spark notion, PDs clearly did have an internal drive and motivation for undertaking research, though more often mentioned were external rewards or drivers such as promotion, external recognition and competition that fit with utility maximisation. Perhaps most seen was a sense of cumulative advantage with, for example, researchers who undertook their PhDs overseas then building on that advantage in terms of later publications, grants, and promotions.

The majority of predictors of PD papers, resulting from the publication analysis, are in concordance with existing literature on highly cited papers. They confirm conclusions related to the length of the paper (Elgendi, 2019; Onodera & Yoshikane, 2015), abstract

(Lokker et al., 2008) and title (Haslam et al., 2008); the number of authors (Didegah & Thelwall, 2013; Elgendi, 2019; Onodera & Yoshikane, 2015), co-author affiliations from overseas institutions (Van Dalen & Henkens, 2001) and references (Davis et al., 2008; He, 2009); and the quality of the journals (Fu & Aliferis, 2010; Peng & Zhu, 2012). New predictors included the identification of topics that significantly distinguished PDs from NPDs, such as "neural networks" and "wireless sensor networks", along with publishers who were strongly associated with PD papers, such as Elsevier. This thus provides support for theories based around normative, social constructivist and natural growth factors driving publication citation rates.

Predictors relevant to global South challenges

Through this study, we were able to identify predictors of PDs in a global South context that had not been identified in previous studies, and could provide pointers to ways of overcoming challenges specific to Southern researchers. Southern researchers work in contexts of *resource limitation*, and PD researchers apply more for research grants and travel funds from international funding bodies. Some applications included partners from Northern universities, which increased the chances of securing the funds, as those partners are more familiar with grant procurement processes and more experienced in writing proposals. PDs build long-standing research collaborations with their overseas supervisors and PhD granting institutions, which may provide further access to research funds either directly or via joint grant applications. In terms of papers, the publication analysis showed that PDs published more journal articles and fewer conference papers. This choice may relate to seeking profile and citations for outputs: avoiding low-visibility local conferences, and selecting journals as more likely to deliver citations than conferences. But it also fits well as a strategy in the context of limited availability of travel funds. Tendency of PDs to publish with more authors and with foreign authors could also help pay for journal publication fees, with fees split across more authors or paid from overseas sources.

Southern researchers were seen to encounter *institutional biases* that make it harder for them to get published and cited. PDs are more likely to co-publish with foreign authors, especially US authors, which will help compensate for any such biases among editors, reviewers in single-blind or open review systems, and readers. (Seeking out foreign co-researchers and co-authors also acts as a compensation against the local contextual challenge of there being a *smaller research population* from which to draw research and publication collaborations.) PDs' preference for working on established research areas rather than on radical research topics may also help in relation to institutional barriers, with research that builds incrementally on existing ideas and literature being more likely to be accepted for publication by referees, and cited by others working in the established area. Any biases against citation of work by Southern researchers may be counteracted by PDs' publication of papers with more authors and more affiliations than NPDs. Having multiple authors and affiliations increases the likelihood of citations, as each author has their own network and bringing those networks together can increase readership (Elgendi, 2019). Multiple authorship may also enrich the paper through the integration of different perspectives and expertise, which could lead to greater citation (Peng & Zhu, 2012). Similarly, PDs publish papers with a larger number of references which increases paper visibility through citation-based search in databases that allow it, such as Google Scholar (Didegah & Thelwall, 2013), and through the "tit-for-tat" hypothesis i.e. authors tend to cite those who cite them (Webster et al., 2009).[19] By and large, then, this tends to support social constructivist views of publication citations; showing how contextual factors influence publication—but also how researchers seek to compensate when those factors may tend to reduce citation rates.

Southern researchers work in contexts of *lower English proficiency*, and PDs were shown to take scientific writing and English writing courses more than NPDs, and their greater likelihood of PhD study at a global North university may also have enhanced their command of English.

Methodological innovation

The use of six different citation metrics enabled us to evaluate performance using different dimensions while controlling for factors that could disadvantage certain groups. It also enabled us to identify and profile PD researchers into three main clusters: rising stars, high performers and highly cited researchers. It was not possible to investigate predictors specific to each cluster individually, due to their small sample size, but this could be a possible avenue for future research.

The majority of studies on predictors of high-performing researchers have focused on individual-level and institution-level predictors. This is one of very few studies that examined publication-level predictors along with individual-level predictors through multiple stages, angles and by triangulating different sources of data. This multi-stage process was both useful and insightful. A number of assumptions about PDs in Stage 1 turned out to be not statistically significant in Stage 2. Examples include practices related to research publishing like submitting papers in conferences followed by extended submissions in journals, paying for proofreading services, and practices related to where researchers publish their work. On the other hand, Stage 1 was crucial because it led to the discovery of PD predictors (some of which had not previously been examined in the literature) that proved to be significant in the statistical analysis that was conducted in Stages 2 and 3. Examples of those predictors include but are not limited to publishing with foreign reputable authors, taking scientific and formal writing courses, and the selection of journal publishers.

Stage 3 enabled us to better understand significant predictors that were identified in Stage 2, e.g. the discovery that teams established overseas were mainly located in the US with authors having US university affiliations. Stage 3 was also useful in quantifying the types of papers (i.e. conference paper, review paper and journal article), and the quality of journals and their different publishers. Although recent studies already demonstrated that topic-related paper features increase the predictive power of highly cited research (Hu et al., 2020), this is one of very few studies that combined topic or paper intrinsic features with extrinsic and publication outlet features to predict papers of high performers. We also explored the adoption and prevalence of topics over time in each of the PD and NPD corpora, which provided additional longitudinal insights that were not possible to capture through the regression analysis alone. It also emerged that it was possible to predict a paper of a PD from its features with an accuracy that is similar to predicting if this researcher is a PD using his/her survey response.

[19] More references might also indicate more comprehensive work, hence a better quality paper, and could mean a large related field, hence better citations (Moed et al., 1985).

Through this study, we demonstrated that application of the DPPD methodology has potential value to the scientometrics field. Advances in this field have enabled digital measurement and tracking of researchers' performance using multiple dimensions, and the open nature of their digital products (i.e. publications) enabled us to digitally quantify and identify some of their publication strategies and research directions. DPPD also provided means to reduce the qualitative search space by limiting the interviewing to a smaller sample of information rich individuals i.e. PDs, thus reducing the time needed for hypothesis generation. Finally, the "data powered" aspect of DPPD characterised by combining digital data with traditional data helped us confirm and better understand the identified predictors.

Practical implications

The key finding of this study is the identification of a set of factors that are significant predictors of PD outcomes. Our analysis cannot, of course, guarantee that applying these factors more broadly would lead to the same outcomes achieved by PDs. Additionally, although causal connections have been outlined in many instances, correlates of high performance do not necessarily imply a causal relation. The work here has only covered one academic discipline in one global South location: replication in other disciplines and countries represents a future research agenda.

One must also step back and recognise two things. First, that citation-based research performance is not the "be all and end all" that should mechanically shape research: relevance of topic to national socio-economic challenges or development of Southern-based methodology and theory could also, for example, be important criteria for Southern researchers. Second, that the findings here in part reflect structural impediments. The fact that highly-cited researchers are overwhelmingly male would not, for example, generate the practical implication that there should be greater resource flows to men in order to generate stronger research performance!

Nonetheless, there would be value in individual Southern researchers reflecting on the research- and paper-related behaviours that have been shown associated with positive-deviant research profiles. These include publishing with multiple authors from different institutions (domestic and international); establishing connections with foreign reputable authors; including a large number of references; having a comprehensive abstract; publishing in particular journals instead of conferences; and contributing to mainstream topics that build on existing work.

Higher education institutions and higher education policy makers may also reflect on the findings, and consider strategic implications for training, resource provision, collaborations, etc. For example, English and scientific/formal writing courses were associated with PD performance; such training could be part of the mandatory training that academics are required to take in order to be promoted in the Egypt's higher education system. Training could be designed around research grant writing and providing guidance on funding bodies that researchers can apply to. International research collaborations appeared as an important predictor of PDs; so university senior managers and policy makers can explore ways to reduce barriers and increase opportunities for overseas PhD study, post-PhD return, and ongoing joint research projects with global North universities.

Future research

This study has developed and tested a methodology that could be replicated in other contexts, such as other countries or other academic disciplines. However, it only covered the first three stages of the DPPD method: defining the problem, determining positive deviants, and discovering the PD practices and strategies. The last two stages of the DPPD method concerned with designing and implementing interventions, and monitoring and evaluating their effects on the intervention population, were not included in this study due to time and resource constraints. There is an opportunity for future research to apply the full DPPD method especially given that the performance indicators that were captured for the study population could relatively easily be used for monitoring and evaluating interventions.

Furthermore, this study demonstrated how the different citation metrics enabled us to cluster and profile researchers into certain groups. However, there is still an opportunity to explore cluster-specific predictors of performance if the sample size per cluster/group is big enough to infer potential hypotheses related to members of the group. Those predictors can then inform cluster-specific interventions. For example, identifying predictors specific to the ‘rising stars’ group (characterised by the low publication age) and using those findings to design interventions targeting young researchers. Additional publication-level predictors, such as the data on author contributions increasingly available via initiatives such as CRediT, can be used to understand how collaboration takes place in papers of positive deviants versus papers of non-positive deviants. Future research could also include network analysis using co-authorship data to investigate the relationship between research groups and PD performance, and to measure the magnitude of local and international collaboration that positive deviants engage in.

In this study we looked into individual-level and publication-level predictors, but we did not look into institutional-level predictors which could provide a more holistic understanding of outperformance. Looking into such supra-individual factors could uncover potential structural factors that are either conducive to or hinder outperformance within institutions. Identifying such factors could then inform higher education policies.

Appendix A: Stage 1 Interview Guide

Following Yin (2014), the interview guide was divided into two levels. Level two questions (L2Q) were questions for which answers were sought i.e. via mental inquiry; level one questions (L1Q) were questions addressed to the interviewee directly i.e. via verbal inquiry.

L2Q1: Do PDs have different motives?

L1Q1: What motivates you to publish your research?

L2Q2: Do PDs publish different types of research?

L1Q2: What kind of research do you prefer? (E.g. review, model development, coding, data analysis, etc)

L2Q3: Do PDs have different research strategies?

L1Q3.1: Where do you usually publish your research? (E.g. local conferences, international conferences, local journals, international peer reviewed journals, etc)

L1Q3.2: Are there specific conferences that you always attend?

L1Q3.3: What kind of co-authorship do you prefer the most? (E.g. student, colleague, someone from the department, someone outside the department, international co-authors)

L2Q4: How do PDs increase the chances of paper acceptance?

L1Q4.1: How do you decide on research material that qualifies for publication?

L1Q4.2: How do you decide if this is a conference paper or a journal article?

L1Q4.3: Are there specific conferences you target for paper submission?

L1Q4.4: Are there specific journals you target for paper submission?

L2Q5: Do PDs perform certain practices that increase paper citation as per the theoretical prepositions found in the literature?

L1Q5.1: How do you plan writing a paper? How long does it take to publish a paper? Are there certain steps that you perform for research publication?

L1Q5.2: What constitutes your literature review?

L1Q5.3: Do you target a minimum number of references in your articles?

L1Q5.5: How do you select the papers which you will cite or use as your literature review? Are there key individuals that you always cite?

L2Q6: Are there challenges specific to PDs?

L1Q6.1: What kind of challenges do you face in publishing (e.g. finding co-authors, journal publication fees, and conference travel expenses)?

L1Q6.2: How do you overcome those challenges?

L2Q7: Do PDs develop their research skills in ways different than NPDs ?

L2Q7.1: Did you take any type of informal education to enhance your research performance?

L2Q7.2: Do you use any tools that support your research publication process?

L2Q7.3: Did you use any of the following approaches for research publication? Please explain: Writing support groups; Structured writing courses; Provision of a writing coach

Appendix B: Stage 2 Survey Questionnaire

1. Full Name

2. Affiliated University

3. Email

4. What is your gender?
   - Male
   - Female

5. What is your marital status?
   - Single
   - Married
   - Separated/divorced
   - Widowed

How many children do you have, if any?

6. What is your last degree?
   - MSc
   - PhD

7. How long did it take you to finish it?
   - 2-3 years
   - 4-5 years
   - More than 5 years

8. From which university did you obtain your last degree?

9. What is your specific field of research?

10. What is the title of your primary current appointment?
    - Assistant Lecturer
    - Lecturer
    - Associate Professor
    - Professor or Emeritus Professor

11. Are you the department chair or were you assigned department chair before?
    - Yes
    - No

If yes, please indicate the start year and end year as department chair
Start year:
End year:

12. For how many of each of the following types of individuals do you currently serve as official advisor?
    - Undergraduate Groups (Graduation Projects):

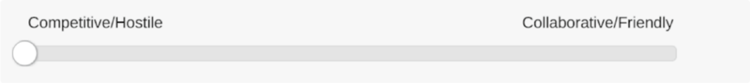
- MSc Students:
- PhD Students:

13. During the past five years, what is your average teaching and administrative work hours per week? (If in your current position for less than five years, base this on the period since your appointment)

14. Have you received any of the following resources during your academic career (Check all that apply)
    - Publication financial support
    - Research grant
    - College scholarship
    - Travel funds to attend a conference
    - None of the above

15. How many conferences did you attend in the past three years?

16. Please rate the climate of your department based on the following continuum.



17. Are you a graduate of a language school?
    - Yes
    - No

18. Do you work outside the university?
    - Yes
    - No

19. How many hours per week do you work outside the university?

20. What motivates you most to publish your research?
    - I publish research to get a promotion
    - I publish research to stay competitive
    - I publish research for international recognition
    - I publish research because I enjoy it
    - None of the above

21. The majority of your research belongs to which type of the below?
    - Studies suggesting new ways of viewing/implementing information processing systems e.g. theories, new architectures, new frameworks, ontologies, network protocols
    - Research involving the creation of new information-processing systems
    - Research involving the creation and evaluation of tools, formalisms, techniques/methods to support existing information processing systems
    - Research on social and economic issues related to information processing systems (Including studies of the social and economic impact of information systems, ethical issues, changing views of humanity, etc.)

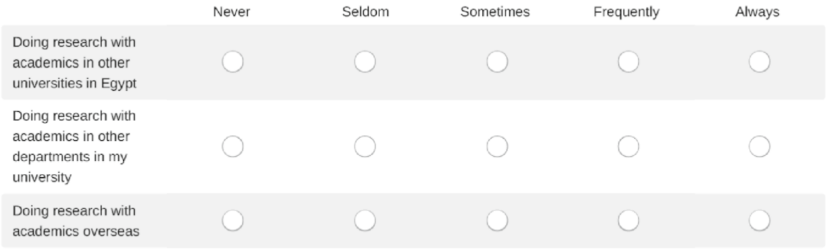
22. Which of the below research strategies reflect the majority of your research? (Please check all that apply)

- I prefer to do radical research that suggests new models / frameworks / methods / architecture that weren't implemented before
- I prefer to do incremental research that enhances existing models / frameworks / methods / architectures
- I prefer to map out broad features of important new areas, leaving detailed studies to others
- I prefer to probe deeply and thoroughly in narrow areas
- I prefer research which looks for immediate solutions to real life problems (e.g. social problem or industry need)
- I prefer purely theoretical research
- I prefer to carry out research work pretty much on my own
- I prefer to carry out research within a research team
- I prefer long-term projects to short-term ones
- I prefer short-term projects to long-term ones

23. From where do you get research ideas? (Please check all that apply)

- Publications of researchers I follow on academic platforms (e.g. Google Scholar)
- Live or recorded webinars (e.g. IEEE webinars)
- Papers citing my work
- Conference attendance
- Future work section of papers
- Other (please specify)

24. For each of the below approaches please rate how often do you apply them?

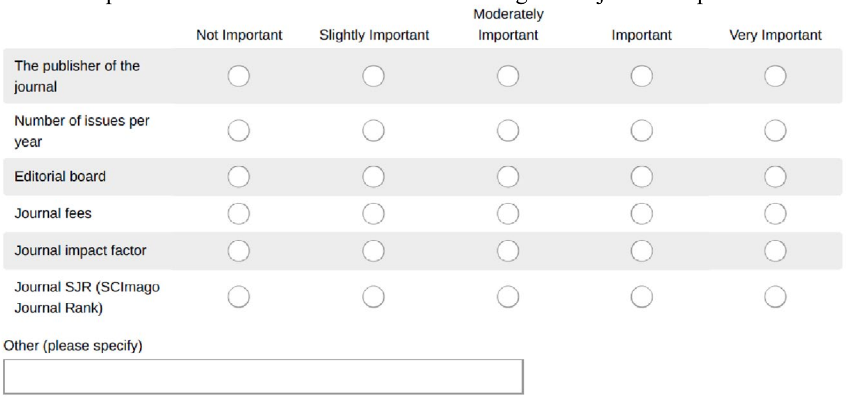


	Never	Seldom	Sometimes	Frequently	Always
Doing research with academics in other universities in Egypt	○	○	○	○	○
Doing research with academics in other departments in my university	○	○	○	○	○
Doing research with academics overseas	○	○	○	○	○

25. Where do you publish your research? (Check all that apply)

- Journals indexed in Scopus
- Journals indexed in Thomson ISI (Clarivate Analytics)
- International Conferences with Proceedings indexed in Scopus
- International Conferences with Proceedings indexed in Thomson ISI (Clarivate Analytics)
- Local Indexed Conferences
- Non-indexed Journals
- Non-indexed Conferences

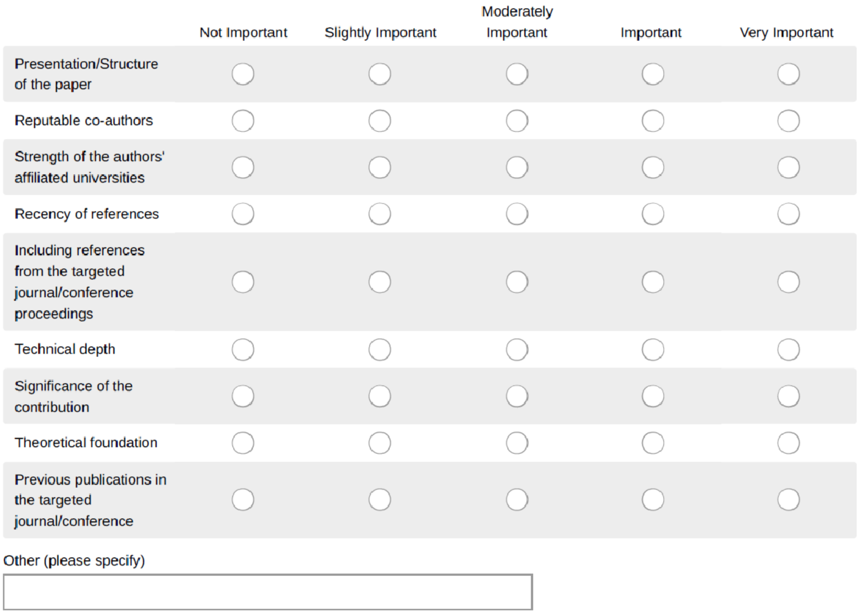
26. How important are the below factors in determining which journal to publish in?



	Not Important	Slightly Important	Moderately Important	Important	Very Important
The publisher of the journal	○	○	○	○	○
Number of issues per year	○	○	○	○	○
Editorial board	○	○	○	○	○
Journal fees	○	○	○	○	○
Journal impact factor	○	○	○	○	○
Journal SJR (SCImago Journal Rank)	○	○	○	○	○

Other (please specify)

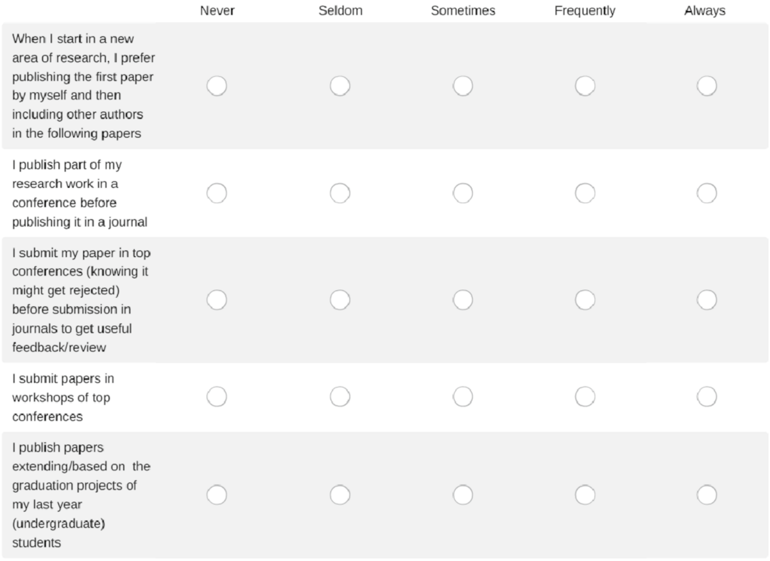
27. How important are the below factors in increasing the chances of acceptance of a paper in a
journal/conference?



	Not Important	Slightly Important	Moderately Important	Important	Very Important
Presentation/Structure of the paper	○	○	○	○	○
Reputable co-authors	○	○	○	○	○
Strength of the authors' affiliated universities	○	○	○	○	○
Recency of references	○	○	○	○	○
Including references from the targeted journal/conference proceedings	○	○	○	○	○
Technical depth	○	○	○	○	○
Significance of the contribution	○	○	○	○	○
Theoretical foundation	○	○	○	○	○
Previous publications in the targeted journal/conference	○	○	○	○	○

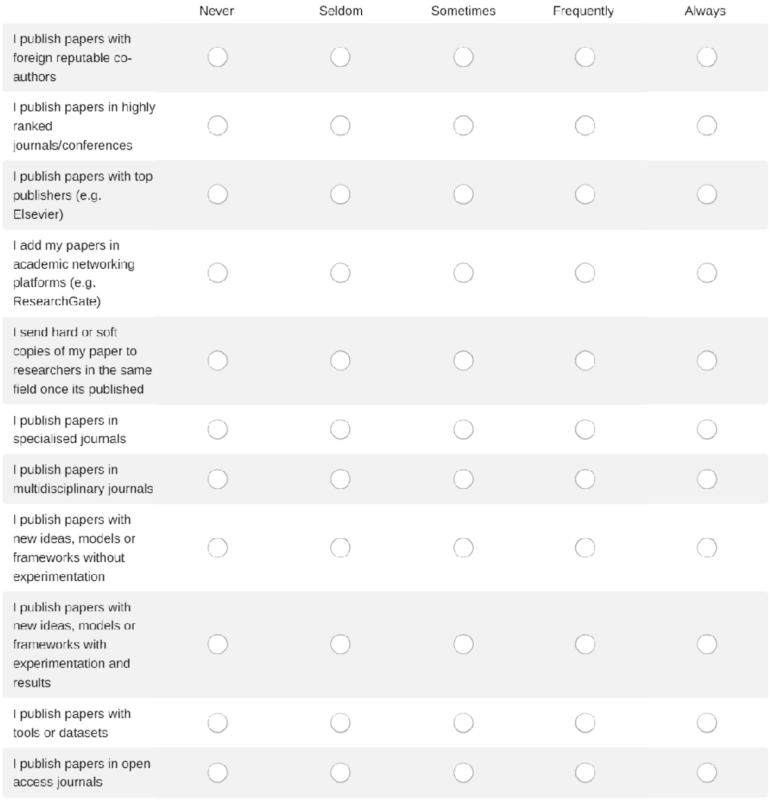
Other (please specify)

28. Please rate the below publication strategies based on how often you apply them.



	Never	Seldom	Sometimes	Frequently	Always
When I start in a new area of research, I prefer publishing the first paper by myself and then including other authors in the following papers	○	○	○	○	○
I publish part of my research work in a conference before publishing it in a journal	○	○	○	○	○
I submit my paper in top conferences (knowing it might get rejected) before submission in journals to get useful feedback/review	○	○	○	○	○
I submit papers in workshops of top conferences	○	○	○	○	○
I publish papers extending/based on the graduation projects of my last year (undergraduate) students	○	○	○	○	○

29. Please rate the below publication approaches based on how often you apply them.

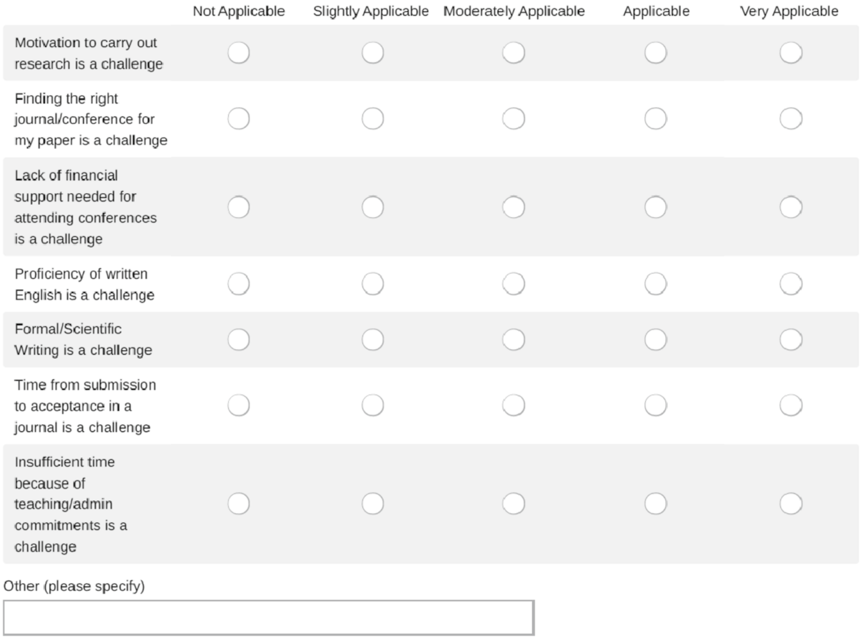


	Never	Seldom	Sometimes	Frequently	Always
I publish papers with foreign reputable co-authors	○	○	○	○	○
I publish papers in highly ranked journals/conferences	○	○	○	○	○
I publish papers with top publishers (e.g. Elsevier)	○	○	○	○	○
I add my papers in academic networking platforms (e.g. ResearchGate)	○	○	○	○	○
I send hard or soft copies of my paper to researchers in the same field once its published	○	○	○	○	○
I publish papers in specialised journals	○	○	○	○	○
I publish papers in multidisciplinary journals	○	○	○	○	○
I publish papers with new ideas, models or frameworks without experimentation	○	○	○	○	○
I publish papers with new ideas, models or frameworks with experimentation and results	○	○	○	○	○
I publish papers with tools or datasets	○	○	○	○	○
I publish papers in open access journals	○	○	○	○	○

30. What is the primary reason for presenting in conferences?

- Interaction with peers and getting feedback
- To be known among my research community
- To publicise my research and attract paper citation
- To gain knowledge about new research areas and trends
- To search for academic posts, possible grants and project collaborations
- Other (please specify)

31. To what extent are the below research publication challenges applicable on you?



	Not Applicable	Slightly Applicable	Moderately Applicable	Applicable	Very Applicable
Motivation to carry out research is a challenge	○	○	○	○	○
Finding the right journal/conference for my paper is a challenge	○	○	○	○	○
Lack of financial support needed for attending conferences is a challenge	○	○	○	○	○
Proficiency of written English is a challenge	○	○	○	○	○
Formal/Scientific Writing is a challenge	○	○	○	○	○
Time from submission to acceptance in a journal is a challenge	○	○	○	○	○
Insufficient time because of teaching/admin commitments is a challenge	○	○	○	○	○

Other (please specify)

32. Do you use any of the below approaches to overcome these challenges? (Please check all that apply)

- I use journal finder online tools
- I pay for proofreading and editing services for my paper
- I seek external funding agencies (e.g. ITIDA, ASRT, TIEC) to cover the costs of travelling to attend conferences
- I use the financial support provided by the university to cover my travel and publication fees
- I apply for research grants (e.g. Erasmus)
- I establish research teams overseas
- Other (please specify)

33. Do you use any of the below tools in writing and publishing your research? (Please check all that
apply)

- Grammarly
- Reference managers (e.g. Mendeley)
- Latex (e.g. Sharelatex)
- Other (please specify)

34. Do you enhance your publication quality using any of the below approaches? (Please check all that apply)

- Observing highly cited papers to see how they are written and structured
- English writing courses
- Scientific writing / Formal writing courses
- Technical courses related to the field
- Using a graphic designer to represent results in an attractive manner
- Sending papers to friends/relatives for proof editing
- Other (please specify)

35. Do you check the papers of researchers who cited your work

- Yes
- No

36. Why do you track citations mainly?

- To check the geographical distribution of the citing papers
- See the impact of the paper after removing self-citation
- Get ideas on future research areas / improvement areas
- Other (please specify)

37. Do you have an account on any of the below research platforms? (Please check all that apply)

- Academia
- Semantic Scholar
- ResearchGate
- Google Scholar profile
- Arxiv
- DBLP
- ORCID
- ResearcherID
- ACM
- Other (please specify)

38. Please state at least two actions/strategies that you believe could increase paper citation rates.

Action 1:

Action 2:

References

Abdi, H. (2003). Partial least square regression (PLS regression). *Encyclopedia for Research Methods for the Social Sciences, 6*(4), 792–795.

Albanna, B., & Heeks, R. (2019). Positive deviance, big data, and development: A systematic literature review. *Electronic Journal of Information Systems in Developing Countries*, *85*(1), e12063.

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics, 3*(4), 273–289.

Altanopoulou, P., Dontsidou, M., & Tselios, N. (2012). Evaluation of ninety-three major Greek university departments using Google Scholar. *Quality in Higher Education, 18*(1), 111–137.

Baldi, S. (1998). Normative versus social constructivist processes in the allocation of citations: A network-analytic model. *American Sociological Review, 63*(6), 829–846.

Bastien, P., Vinzi, V. E., & Tenenhaus, M. (2005). PLS generalised linear regression. *Computational Statistics & Data Analysis, 48*(1), 17–46.

Batista, P. D., Campiteli, M. G., & Kinouchi, O. (2006). Is it possible to compare researchers with different scientific interests? *Scientometrics, 68*(1), 179–189.

Belew, R. K. (2005). Scientific impact quantity and quality: Analysis of two sources of bibliographic data. arXiv:cs.IR/0504036v1.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*, 993–1022.

Blicharska, M., Smithers, R. J., Kuchler, M., Agrawal, G. K., Gutiérrez, J. M., Hassanali, A., Huq, S., Koller, S. H., Marjit, S., Mshinda, H. M., & Masjuki, H. (2017). Steps to overcome the North-South divide in research relevant to climate change policy and practice. *Nature Climate Change, 7*(1), 21–27.

Cole, J. R., & Cole, S. (1973). *Social Stratification in Science*. University of Chicago Press.

Confraria, H., Godinho, M. M., & Wang, L. (2017). Determinants of citation impact: A comparative analysis of the Global South versus the Global North. *Research Policy, 46*(1), 265–279.

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G. & Connolly, M. J. L. (2008). Open access publishing, article downloads, and citations: randomised controlled trial. *British Medical Journal*, *337*, a568.

Didegah, F., & Thelwall, M. (2013). Determinants of research citation impact in nanoscience and nanotechnology. *Journal of the American Society for Information Science and Technology, 64*(5), 1055–1064.

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

Elgendi, M. (2019). Characteristics of a highly cited article: A machine learning perspective. *IEEE Access, 7*, 87977–87986.

Flanigan, A. E., Kiewra, K. A., & Luo, L. (2018). Conversations with four highly productive German educational psychologists: Frank Fischer, Hans Gruber, Heinz Mandl, and Alexander Renkl. *Educational Psychology Review, 30*(1), 303–330.

Franceschet, M. (2010). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics, 83*(1), 243–258.

Fu, L. D., & Aliferis, C. F. (2010). Using content-based and bibliometric features for machine learning models to predict citation counts in the biomedical literature. *Scientometrics, 85*(1), 257–270.

Gibbs, W. W. (1995). Lost science in the third world. *Scientific American, 273*(2), 92–99.

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

Glänzel, W., & Schoepflin, U. (1995). A bibliometric study on ageing and reception processes of scientific literature. *Journal of Information Science, 21*(1), 37–53.

Goldemberg, J. (1998). What is the role of science in developing countries? *Science, 279*, 1140–1141.

Gonzalez-Brambila, C. N., Reyes-Gonzalez, L., Veloso, F. & Perez-Angón, M. A. (2016). The scientific impact of developing nations. *PLoS One*, *11*(3), e0151328.

Hagstrom, W. (1965). *The Scientific Community*. Basic Books.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association, 69*, 383–393.

Harris, G., & Kaine, G. (1994). The determinants of research performance: A study of Australian university economists. *Higher Education, 27*(2), 191–201.

Harzing, A. W. (2007). *Publish or Perish*. http://www.harzing.com/pop.htm

Haslam, N., Ban, L., Loughnan, S., Peters, K., Whelan, J., & Wilson, S. (2008). What makes an article influential? Predicting impact in social and personality psychology. *Scientometrics, 76*(1), 169–185.

He, Z. (2009). International collaboration does not have greater epistemic authority. *Journal of the American Society for Information Science and Technology, 60*(10), 2151–2164.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Scientometrics, 85*(3), 741–754.

Hu, Y., Tai, C., Ernest, K. & Cai, C. (2020). Identification of highly-cited papers using topic-model-based and bibliometric features: the consideration of keyword popularity. *Journal of Informetrics*, *14*(1), 101004.

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

Kao, A., & Poteet, S. R. (2007). *Natural Language Processing and Text Mining*. Springer Science & Business Media.

Kaplan, N. (1965). The norms of citation behavior: Prolegomena to the footnote. *American Documentation, 16*(3), 179–184.

Karlsson, S., Srebotnjak, T., & Gonzales, P. (2007). Understanding the North-South knowledge divide and its implications for policy: A quantitative analysis of the generation of scientific knowledge in the environmental sciences. *Environmental Science and Policy, 10*(7–8), 668–684.

Kelchtermans, S., & Veugelers, R. (2013). Top research productivity and its persistence. *Review of Economics and Statistics, 95*(1), 273–285.

Kiewra, K. A., & Creswell, J. W. (2000). Conversations with three highly productive educational psychologists: Richard Anderson, Richard Mayer, and Michael Pressley. *Educational Psychology Review, 12*(1), 135–161.

King, D. A. (2004). The scientific impact of nations. *Nature, 430*(6997), 311–316.

Knorr-Cetina, K. (1981). *The Manufacture of Knowledge: An Essay on the Constructivist and Contextual Nature of Science*. Pergamon Press.

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multidiscipline exploratory analysis. *Journal of the American Society for Information Science and Technology, 58*(7), 1055–1065.

Kwiek, M. (2016). The European research elite: A cross-national study of highly productive academics in 11 countries. *Higher Education, 71*(3), 379–397.

Kwiek, M. (2018). High research productivity in vertically undifferentiated higher education systems: Who are the top performers? *Scientometrics, 115*(1), 415–462.

Kyvik, S. (1990). Age and scientific productivity. Differences between fields of learning. *Higher Education*, *19*(1), 37–55.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

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

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. (2008). Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: Retrospective cohort study. *British Medical Journal, 336*(7645), 655–657.

Mahanty, S., Boons, F., Handl, J., & Batista-Navarro, R. (2019). Studying the evolution of the 'circular economy' concept using topic modelling. *International Conference on Intelligent Data Engineering and Automated Learning* (pp. 259–270). Springer.

Man, J. P., Weinkauf, J. G., Tsang, M., & Sin, D. D. (2004). Why do some countries publish more than others? An international comparison of research funding, English proficiency and publication output in highly ranked general medical journals. *European Journal of Epidemiology, 19*(8), 811–817.

Mann, G. S., Mimno, D. & McCallum, A. (2006). Bibliometric impact measures leveraging topic analysis. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital libraries - JCDL '06*. New York, NY: ACM Press, 65–74.

Mansournia, M. A., Geroldinger, A., Greenland, S., & Heinze, G. (2018). Separation in logistic regression: Causes, consequences, and control. *American Journal of Epidemiology, 187*(4), 864–870.

Martínez, R. S., Floyd, R. G., & Erichsen, L. W. (2011). Strategies and attributes of highly productive scholars and contributors to the school psychology literature: Recommendations for increasing scholarly productivity. *Journal of School Psychology, 49*(6), 691–720.

Mayrath, M. C. (2008). Attributions of productive authors in educational psychology journals. *Educational Psychology Review, 20*(1), 41–56.

Merton, R. K. (1968). The Matthew effect in science: The reward and communication systems of science are considered. *Science, 159*(3810), 56–63.

Merton, R. K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. University of Chicago Press.

Moed, H. F., Burger, W. J. M., Frankfort, J. G., & Van Raan, A. F. J. (1985). The use of bibliometric data for the measurement of university research performance. *Research Policy, 14*(3), 131–149.

National Science Board (2018). *Science and Engineering Indicators 2018*. NSB-2018–1. Alexandria, VA: National Science Foundation.

O'Boyle, E., & Aguinis, H. (2012). The best and the rest: Revisiting the norm of normality of individual performance. *Personnel Psychology, 65*(1), 79–119.

Onodera, N., & Yoshikane, F. (2015). Factors affecting citation rates of research articles. *Journal of the Association for Information Science and Technology, 66*(4), 739–764.

Ortega, J. L. (2015). How is an academic social site populated? A demographic study of Google Scholar Citations population. *Scientometrics, 104*(1), 1–18.

Parker, J. N., Lortie, C., & Allesina, S. (2010). Characterizing a scientific elite: The social characteristics of the most highly cited scientists in environmental science and ecology. *Scientometrics, 85*(1), 129–143.

Parker, J. N., Allesina, S., & Lortie, C. J. (2013). Characterizing a scientific elite (B): Publication and citation patterns of the most highly cited scientists in environmental science and ecology. *Scientometrics, 94*(2), 469–480.

Pasgaard, M., & Strange, N. (2013). A quantitative analysis of the causes of the global climate change research distribution. *Global Environmental Change, 23*(6), 1684–1693.

Patterson-Hazley, M., & Kiewra, K. A. (2013). Conversations with four highly productive educational psychologists: Patricia Alexander, Richard Mayer, Dale Schunk, and Barry Zimmerman. *Educational Psychology Review, 25*(1), 19–45.

Peng, T. Q., & Zhu, J. J. H. (2012). Where you publish matters most: A multilevel analysis of factors affecting citations of internet studies. *Journal of the American Society for Information Science and Technology, 63*(9), 1789–1803.

Peters, H. P. F., & van Raan, A. F. J. (1994). On determinants of citation scores: A case study in chemical engineering. *Journal of the American Society for Information Science, 45*(1), 39–49.

Positive Deviance Initiative. (2010). *Basic Field Guide to the Positive Deviance Approach*. Tufts University.

Postiglione, G. A., & Jung, J. (2013). World-class university and Asia's top tier researchers. In Q. Wang, C. Ying, & N. Cai Liu (Eds.), *Building World-Class Universities: Different Approaches to a Shared Goal* (pp. 161–179). SensePublishers.

Prpić, K. (1996). Characteristics and determinants of eminent scientists' productivity. *Scientometrics, 36*(2), 185–206.

Ranganathan, P., Pramesh, C., & Aggarwal, R. (2017). Common pitfalls in statistical analysis: Logistic regression. *Perspectives in Clinical Research, 8*(3), 148–151.

Salager-Meyer, F. (2008). Scientific publishing in developing countries: Challenges for the future. *Journal of English for Academic Purposes, 7*(2), 121–132.

'Scientometrics' (2020). *Wikipedia*. Available at: https://en.wikipedia.org/wiki/Scientometrics

'SCImago Journal Rank' (2020). *Wikipedia*. Available at: https://en.wikipedia.org/wiki/SCImago_Journal_Rank

Sidiropoulos, A., Katsaros, D., & Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics, 72*(2), 1–34.

Sternin, J. (2002). Positive deviance: A new paradigm for addressing today's problems today. *The Journal of Corporate Citizenship, 5*, 57–62.

Sternin, M., Sternin, M. D., & Marsh, D. (1997). Rapid, sustained childhood malnutrition alleviation through a positive deviance approach in rural Vietnam: Preliminary findings. In Wollinka O, Keeley E, Burkhalter RB, Bashir N, eds. *The Hearth Nutrition Model: Applications in Haiti, Vietnam, and Bangladesh*. Arlington, VA: BASICS, 49–61.

Stewart, J. A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces, 62*(1), 166–189.

Thesee, G. (2006). A tool of massive erosion: Scientific knowledge in the neo-colonial enterprise. In G. J. Sefa Dei & A. Kempf (Eds.), *Anti-Colonialism and Education* (pp. 25–42). Sense Publishers.

Tobias, R. D. (1995). An introduction to partial least squares regression. *Proceedings of the Twentieth Annual SAS Users Group International Conference* (Vol. 20). Cary, NC: SAS Institute, 1250–1257.

Van Noorden, R. (2010). A profusion of measures: Scientific performance indicators are proliferating–leading researchers to ask afresh what they are measuring and why. *Nature, 465*(7300), 864–866.

Van Dalen, H. P., & Henkens, K. (2001). What makes a scientific article influential? *The Case of Demographers. Scientometrics, 50*(3), 455–482.

Van Dalen, H. P., & Henkens, K. (2005). Signals in science—On the importance of signaling in gaining attention in science. *Scientometrics, 64*(2), 209–233.

Walfish, S. (2006). A review of statistical outlier methods. *Pharmaceutical Technology, 30*(11), 82.

Walters, G. D. (2006). Predicting subsequent citations to articles published in twelve crime-psychology journals: Author impact versus journal impact. *Scientometrics, 69*(3), 499–510.

Webster, G. D., Jonason, P. K., & Schember, T. O. (2009). Hot topics and popular papers in evolutionary psychology: Analyses of title words and citation counts in evolution and human behavior, 1979–2008. *Evolutionary Psychology, 7*(3), 147470490900700300.

White, C. S., James, K., Burke, L. A., & Allen, R. S. (2012). What makes a 'research star'? Factors influencing the research productivity of business faculty. *International Journal of Productivity and Performance Management, 61*(6), 584–602.

World Bank. (2020). *Science & Technology Indicators*. World Bank.

Yair, G., Gueta, N., & Davidovitch, N. (2017). The law of limited excellence: Publication productivity of Israel Prize laureates in the life and exact sciences. *Scientometrics, 113*(1), 299–311.

Authors and Affiliations

Basma Albanna[1] · Julia Handl[2] · Richard Heeks[1]

Julia Handl
julia.handl@manchester.ac.uk

Richard Heeks
richard.heeks@manchester.ac.uk

[1] Centre for Digital Development, Global Development Institute, University of Manchester, Manchester, UK

[2] Alliance Manchester Business School, University of Manchester, Manchester, UK