A percentile rank score of group productivity: an evaluation of publication productivity for researchers from various fields

Koh Yamamoto¹ · Takuo Yasunaga²

Received: 20 January 2021 / Accepted: 13 January 2022 / Published online: 3 March 2022
© The Author(s) 2022

Abstract

The difficulty in evaluating the research performance of groups is attributable to the following two factors: 1) difference of population size or discipline of group members and 2) skewed distribution of the research performance of individuals. This study attempts to overcome this difficulty, focusing on the research performance based on publication productivity. We employ the normalized index for the number of papers, in which publication efficiency was considered and disciplinary variation in the publication intensity was corrected by the disciplinary averages, to calculate a new percentile rank score. The score was developed on the basis of the principle that a person who is rare is valuable. The score was also tested with publication data for faculty members of 17 Japanese universities. The employment of the normalized index increased the score of universities with relatively few faculty members working in the disciplines of high productivity, resulting in more plausible university rankings. The rankings show a high correlation with those for a previously established percentile rank score, which was developed for citation analysis, and they are consistent with the judgment by evaluators of several universities under study. The advantage of the new score over the previous one is that it has no room for arbitrariness in determining the scheme of rank classification and the weights given to each rank class.

Keywords Percentile rank score · Publication productivity · Normalized index for the number of papers · Efficiency

Introduction

Scientometric indicators are statistical measures based on the number and distribution of publications, authors, references and citations (Schubert et al., 1988). They are used for assessment, such as time series analysis of research performance or performance

✉ Koh Yamamoto
yamamoto-k@office.kyutech.ac.jp

¹ Institutional Research Office, Kyushu Institute of Technology, 1-1 Sensui, Tobata, Kitakyushu, Fukuoka 804-8550, Japan

² Department of Bioscience and Bioinformatics, Graduate School of Computer Science and Systems Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka, Fukuoka 820-8502, Japan

comparison with others. The data are often obtained from large abstract and citation databases, such as Web of Science by Clarivate Analytics and Scopus by Elsevier, which have specific criteria for registering scientific journals. The indicators offered by the databases can be roughly divided into two types: one for evaluating publication productivity based on the number of papers and the other for measuring the visibility of papers or journal impact. This study focuses on the former for which intensive studies are scarce compared with the latter.

Group assessment using the bibliometric indicators is a traditional practice at various levels ranging from countries or research institutions to research projects. For example, Gul et al. (2015) evaluated the difference in research performance among Middle Eastern countries by comparing the share of the number of papers in the region or the average number of citations per paper of each country. At the institutional level, Guskov et al. (2018) conducted a time-series analysis for the growth rate of the number of papers sorted by document types, covering research institutes participating in a national project for improvement of research performance. At the project level, to evaluate the research performance of 147 chemistry groups in the Netherlands, van Raan (2006) compared bibliometric indicators, such as the number of papers and citations, to the research quality judged by peers.

Two types of data set can be used to evaluate group publication productivity: one collected using a group name, such as a university name, and the other using authors' names. The two kinds of data set produce different results when the time window of data collection is set to be long. The paper count generated by the former represents the research activity of a group and is widely used because it can be directly obtained from the abstract and citation databases, whereas that by the latter represents the research potential of a group because it contains the publications that constituent members had published in other groups in the past. The latter, a bottom-up approach of data collection and integration, requires great care to unify author IDs or authorial information, which tend to split into multiple ones with time; thus, name disambiguation is essential. Although evaluating research potential is conducted less than that of research activity for this reason, it is also important for executive board members of research institutes to formulate strategies to improve research performance. Kotsemir and Shashnov (2017) employed the bottom-up approach in evaluating the research potential of departments and staff members of a Russian university to avoid underestimation of the number of publications, which is unique to some countries, due to missing correct affiliation names with the publications indexed in the Scopus database. This study also evaluates the research potential of groups using publication data collected at the individual level.

The number of papers aggregated at several levels— such as individual, department, and university— is a simple index to estimate the publication productivity of the target, but it has two deficiencies. First, efficiency is not taken into account. Schubert et al. (1988) noted the importance of evaluating a citation impact in consideration of balance between the impact as an output and publication effort as an input. The same applies in evaluating publication productivity. Second, comparing the index across disciplinary boundaries is difficult due to a large disciplinary variation stemmed from a difference in average publication intensity and average number of coauthors. For the citation impact, a number of studies have been conducted to develop normalized indices, which are provided in research performance evaluation tools associated with abstract and citation databases (e.g., CNCI by Clarivate Analytics and FWCI by Elsevier); these databases enable us to compare the citation impact across years, document types, and disciplinary boundaries. Using these indices, for example, a comparison of research quality has been made not only among faculty members but also among nations or institutions (Avanesova and Shamliyan, 2018), as

has a comparison of the effectiveness of interuniversity cooperation (Khor and Yu, 2016). Regarding publication productivity, in contrast, no such normalized index is widely perceived; evaluation and comparison within a specific discipline are considered preferable (Abramo and D'Angelo, 2014) and therefore are generally conducted in the previous studies (e.g., Vinluan, 2012; Barrot, 2017).

This study employs the normalized index for the number of papers named as Discipline Weighted Publication Productivity (DWPP) proposed by Yamamoto and Ishikawa (2017) to make a group comparison across population sizes and disciplines of its constituent members. The DWPP is calculated at the individual level as a ratio of paper count as an output to publication effort as an input to represent publication efficiency; the publication effort is here defined as an individual's effort devoted to a research field following Schubert et al. (1988). The normalization is conducted by dividing the efficiency by its average in each discipline and aggregating over all disciplines. Even with this index, however, comparing the publication productivity at the group level remains not easy. This is because most groups are composed of a small number of prolific researchers and a large number of unproductive researchers, which skews the distribution of the DWPP and renders its arithmetic mean inappropriate to know the representative value.

Although calculating the representative value as the median value is possible, information obtained from the distribution shape should be lost. The same is true of citation impact, and attempts have been made to evaluate researchers or groups more precisely on the basis of relative rankings. To compare researchers with different publication sets, Leydesdorff et al. (2011) developed the mean percentile rank score, which is calculated by classifying the researchers' publications by the percentile rank based on the number of citations, obtaining the relative frequencies in every rank class, and summing the values after weighting. The authors proposed two schemes to give the weights, which increase in a different manner as the percentile rank rises; in their method, any other scheme can be chosen depending on the purposes of the evaluation. The "degree of freedom" for choosing the scheme is advantageous for evaluators from the aspect of great flexibility, while a disadvantage is that the evaluation is liable to be arbitrary. The present study therefore aims at developing a novel index called the Better Performance Index (BPI), which is derived from the relationship between relative ranks in the whole sample and in each group, without room for arbitrariness.

This paper first describes the outline and calculation method of the DWPP and BPI, and then shows the values calculated from publication data collected at the individual level for nearly 15,000 faculty members of 17 Japanese universities. Finally, the following are discussed: interpretation of the DWPP and BPI, the effects of normalizing the number of papers on the BPI, and the degree of agreement between university rankings for the BPI and median value or scores by Leydesdorff et al. (2011).

Materials and methods

Data

The Scopus database employs an original author ID system, in which users can unify multiple IDs for a single author manually or using an optional service of Elsevier, i.e., Profile Refinement Service. The publication data used in the present study were collected from Scopus at the individual level, based on a list of researchers in 2019, for 9 years from

2010 to 2018 inclusive. The long time window was chosen to ensure the sufficient sample size required for accurate determination of the DWPP following Yamamoto and Ishikawa (2017). It also enables us to minimize the effect of non-publishing authors within a given period, the so-called “potential authors” (e.g., Schubert and Telcs, 1986; Koski et al., 2016) because it would be reasonable to assume that those who do not publish over such long periods are unlikely to publish after the defined period, which means they are not counted as the potential authors to be included in the analysis. The data are confined to the faculty members of 17 Japanese universities having conducted the ID unification to improve the accuracy of name disambiguation. The faculty members with the ID, hereinafter simply referred to as researchers, total 14,912, corresponding to ~8% of all the faculty members who belong to any Japanese university as of 2017. The number of papers to be analyzed amounts to 267,946; only the document types “article,” “review,” and “conference paper” are included.

The data were not separated by publication year in this analysis because the number of papers does not tend to increase as time passes, unlike the number of citations, setting aside the effect of a few years of registration delay after publication. Although it is conceivable that the long period of analysis may be disadvantageous for early-career faculty, they were treated in the same way as the others as the effects would be negligible on the whole.

Table 1 summarizes the 17 universities under study: 12 comprehensive (i.e., multidisciplinary) universities and 5 specialized universities. Among the universities, a large variety is shown in the number of researchers from just over 100 to just under 3,000, as well as in the level of universities suggested by the Quacquarelli Symonds (QS) World University

Table 1 Outline of the universities under study

University ^a	Category	Researchers ^b	QS Ranking
U1	Comprehensive	2,921	1–100
U2	Comprehensive	2,040	101–300
U3	Comprehensive	1,741	101–300
U4	Comprehensive	1,175	101–300
U5	Comprehensive	1,053	101–300
U6	Comprehensive	1,210	501–750
U7	Comprehensive	946	501–750
U8	Comprehensive	873	501–750
U9	Comprehensive	600	501–750
U10	Comprehensive	488	501–750
U11	Specialized (mainly engineering)	314	751–1000
U12	Comprehensive	270	751–1000
U13	Specialized (mainly engineering)	226	751–1000
U14	Comprehensive	550	Out of Ranking
U15	Specialized (mainly engineering)	251	Out of Ranking
U16	Specialized (mainly engineering)	134	Out of Ranking
U17	Specialized (mainly agriculture)	120	No Entry

^a The universities are sorted in descending order of the QS rankings and the number of researchers.

^b The value is the number of researchers who published one or more papers in journals listed in Scopus between 2010 and 2018.

Rankings from being ranked in the top 100 to being out of rank. At the request of some universities, university names are not disclosed in this article.

Disciplinary classification of researchers

Several studies have analyzed publication productivity across disciplinary boundaries with consideration for efficiency, which requires determination of researcher population in each discipline as an input. The Times Higher Education (THE) World University Rankings includes the research performance evaluation based on the normalized publication productivity calculated from self-reported data of universities for the number of faculty members and the number of papers in each THE discipline (Times Higher Education, 2019). The normalization is, however, very rough as the classification is conducted into only 11 disciplines. In contrast, in the Italian higher education system, every researcher is classified into a single research field of 370; Abramo and D'Angelo (2014) employed the normalized publication productivity to determine university rankings and compare the results with those based on the normalized citation index. Although fine classification is utilized in the normalization, comparing the researchers with those outside of Italy in the same framework is not possible. Besides, determining the discipline uniquely for some researchers working in interdisciplinary field can present difficulty.

Thus, in the present study, the All Science Journal Classification of Scopus assigned to source titles was applied to the researcher's discipline. The classification has two levels: 27 major disciplines and 334 minor disciplines. We employed the major disciplines (hereinafter, ASJC27) in light of the sample size available. For example, one researcher has published two papers in different journals. If ENGI and MATE are assigned to one journal and ENGI and PHYS to the other, his/her discipline is represented as ENGI 50%, MATE 25%, and PHYS 25% (see the appendix for the correspondence of ASJC27 disciplines and its abbreviations). In calculating DWPP, an individual's publication effort is allocated to each related discipline on the basis of this ratio.

Discipline weighted publication productivity

The DWPP employs the fractional counting method by the number of coauthors to evaluate a degree of contribution in the publication process following Moed (2005), which referred that the integer counting method represents participation, whereas the fractional counting method represents a kind of credit to a study. The different coauthors of a publication usually have not equally contributed (Waltman and Van Eck, 2015), but it would present a problem with the evaluation at the individual level, not at the level of large group like a university.

The following is an overview of the DWPP proposed by Yamamoto and Ishikawa (2017). To take the work efficiency into consideration, a unit of publication effort is defined as manpower, with each researcher having a total of one manpower in a given time period. According to Abramo and D'Angelo (2014), productivity is the quintessential indicator of efficiency and defining research productivity as the number of publications per researcher is a norm in bibliometrics. The DWPP is equivalent to the indicator, where each researcher is classified into a single research discipline. In the present study, as researchers are classified into one or more disciplines (see “Disciplinary classification of researchers” section), the separable value of manpower is employed in exchange for the number of researchers. Furthermore, the following simplifying assumptions are adopted:

- 1) Coauthors equally contribute to the publication.
- 2) The paper count can be allocated evenly if the paper is assigned to multiple disciplines.
- 3) A given researcher's publication efforts devoted to a specific discipline are proportional to the ratio of the number of papers in the discipline to the total number of his/her papers.
- 4) Non-publishing faculty members, whose fields of study are not identifiable, can be ignored.

The method for calculating the DWPP and its components in the formula are summarized in Table 2. The DWPP values of all researchers average one; if a given researcher has a DWPP larger than one, he/she has above-average publication productivity.

Percentile rank score

The DWPP makes it possible to compare the publication productivity of individuals across disciplinary boundaries. Meanwhile, its arithmetic mean and standard deviation are considered unsuitable for the comparison of groups. Bibliometric indicators, such as the number of papers and citations, have long been known to have a highly skewed distribution; normality cannot be assumed (e.g., Schubert and Glänzel, 1984; Seglen, 1992; Baccini et al., 2014). As is to be expected, in the present study, null hypothesis that the DWPP values are normally distributed was rejected for all universities by the Shapiro-Wilk normality test at the 5% significance level. As percentile rank is less affected by outliers and can be used for highly skewed data to obtain the representative value, several studies have employed it to evaluate the difference of citation impact among institutions or researchers (e.g., Leydesdorff et al., 2011; Bornmann et al., 2013; Bornmann and Marx, 2014).

This study also adopts the percentile rank score to evaluate the publication productivity of groups. First, the relative frequency of researcher rankings $p(k)$, where k is the class number, is calculated following the steps below:

Table 2 Summary of the method for calculating the normalized number of papers (DWPP)

Index	Formula	Description
Publication effort ratio of researcher i in discipline j (β_{ij})	$\frac{\alpha_{ij}}{\alpha_i}$	α is the number of papers and subscripts i and j indicate the identity numbers of the researcher and discipline, respectively. The symbol α_i represents the total number of papers for researcher i , being integrated over all disciplines.
Publication productivity of discipline j (P_j)	$\frac{\sum_{i=1}^N \alpha_{ij}}{\sum_{i=1}^N \beta_{ij}}$	N is the total number of researchers included in the analysis. The numerator and denominator represent the total number of papers and the total effort ratio, respectively, in discipline j .
DWPP of researcher i (γ_i)	$\sum_{j=1}^M \frac{\alpha_{ij}}{P_j}$	M is the total number of disciplines. The γ_i is calculated by dividing the number of papers for researcher i by the reference value P_j in each discipline and integrating the ratios over all disciplines.

- 1) Calculate individuals' DWPP and ranking in the whole sample based on the value.
- 2) Divide researchers' rankings into l percentile rank classes.
- 3) For each university, sum up the number of researchers in each class and divide it by the university population.

When classifying the percentile ranks at regular intervals, the k -th class includes researchers with percentile ranks greater than $100(k-1)/l$ and less than or equal to $100k/l$.

Next, three percentile rank scores are calculated using the $p(k)$. Two are the scores proposed by Leydesdorff et al. (2011), for which $p(k)$ is calculated in classes divided into various intervals, weighted with the class number k , and then summed up over all classes as follows:

$$R(l) = \sum_{k=1}^l k \cdot p(k) . \quad (1)$$

Leydesdorff et al. (2011) calculated $R(6)$ and $R(100)$ by employing two different schemes of rank classification. For $R(6)$, the ranking data are divided into six percentile rank classes at the 50th, 75th, 90th, 95th, and 99th percentile following National Science Board (2010). Because of the uneven classification, the weight to the $p(k)$ rapidly increases at higher ranks. As for $R(100)$, the data are evenly divided into 100 classes, and the weight is proportional to the percentile rank. Bearing in mind that their target of analysis is the number of citations, which is different from the number of papers, we employ the $R(6)$ and $R(100)$ for comparison with our original score introduced next.

The third score we propose, the BPI, is calculated as follows with ranks evenly divided into l classes:

$$BPI(l) = \frac{1}{l-1} \sum_{k=1}^{l-1} \frac{1 - q_s(k)}{1 - q_e(k)}$$

where $q(k)$ denotes cumulative relative frequency in the k -th class, and the subscripts s and e indicate single university and equilibrium conditions, respectively; the equilibrium conditions are met when the distribution of researcher rankings coincides with that for the whole sample. Here $q_s(k)$ can be calculated as $\sum_{i=1}^k p_s(i)$, and $q_e(k)$ as k/l . This equation can be transformed into the following form using deviation $\delta(k) = [1 - q_s(k)] - [1 - q_e(k)] = q_e(k) - q_s(k)$, which represents an excess of researcher ratio in above the k -th classes for a specific university when compared to that for the whole sample:

$$\begin{aligned} BPI(l) &= \frac{1}{l-1} \sum_{k=1}^{l-1} \frac{1 - q_e(k) + \delta(k)}{1 - q_e(k)} \\ &= 1 + \frac{1}{l-1} \sum_{k=1}^{l-1} \frac{\delta(k)}{1 - q_e(k)} . \end{aligned} \quad (2)$$

The equation suggests two outcomes. First, the deviation $\delta(k)$ receives weights inversely proportional to $1 - q_e(k)$; the weight nonlinearly increases as k increases. Second, the deviation of BPI from one represents the weighted $\delta(k)$ averaged over all classes. To be specific, one researcher in the highest class (100th percentile class; top-1%) weighs 99 times as much as one in the lowest class (1st percentile class; 99%-bottom). In the case of $l = 10$,

Fig. 1 The cumulative distribution function of researcher rankings in a model university. The function is drawn by linearly interpolating the values of adjacent decile classes. The dashed line denotes the equilibrium conditions between a specific university and all universities, i.e., the whole sample. Researchers ranked in the 10th decile class have the highest publication productivity, whereas those in the 1st have the lowest. The closed square and circle show the relative frequency cumulated up to the 5th decile for equilibrium conditions ($q_e(5)$) and for a specific university ($q_s(5)$), respectively; $\delta(5)$ is calculated as $q_e(5) - q_s(5) = 0.5 - 0.75 = -0.25$

the weight in the highest (10th decile class; top–10%) is nine times larger than the lowest (1st decile class; 90%–bottom). Although a small interval should be taken when the analysis includes small groups, this study focuses on universities, and no significant difference was observed between $BPI(10)$ and $BPI(100)$; hereinafter, $BPI(10)$ is referred to as BPI and will be analyzed further.

Cumulative distribution function of researcher rankings in a model university is illustrated for visual understanding of the method (Fig. 1). The university is assumed to have $p(k)$ decreasing from 0.19 by 0.02 at each step with increasing k . The function shows the relation between relative ranks in the whole sample and in the model university; specifically, it reads that a given researcher is at the 5th decile, i.e., the 50th percentile, in the whole, whereas he/she is at the 75th percentile in the university. If the distribution of researcher rankings for the university is similar to that for the whole sample, the function coincides with the equilibrium line, and he/she is at the 50th percentile in the university as well. In Fig. 1, an example is shown for the derivation of $\delta(k)$ in Equation (2), which is calculated for the 5th class as $q_e(5) - q_s(5) = 0.5 - 0.75 = -0.25$. The BPI of the model university comes to 0.5 when cumulating the $\delta(k)$ over all classes after weighted. In summary, the BPI is an indicator of difference between researchers' relative ranks in the whole sample and those in a specific group. A BPI value of less than one can be interpreted as researcher distribution inclined toward the lower ranks from overall average.

Fig. 2 Disciplinary variation of nine-year average publication productivity (P_j) across 17 studied Japanese universities over 2010–2018. The dotted line and error bars on the P_j bars indicate the average value of P_j over all disciplines and 95% confidence intervals calculated by the bootstrap method, respectively. In order to facilitate comparisons of P_j between research areas consisting of closely related disciplines, the data are arranged in the same order as in the Wheel of Science provided by Elsevier

Results

Variation of publication productivity among disciplines and individuals

Figure 2 shows the nine-year average publication productivity P_j calculated for each discipline (see Table 2); the data are arranged so that relevant disciplines are placed adjacent to each other following the way of the “Wheel of Science” provided by Elsevier (Elsevier 2020). Whereas researchers engaged in COMP, MATE, and ENGI disciplines have the highest P_j , those engaged in ARTS and DENT disciplines have the lowest. In this study, disciplines with the P_j significantly higher than its overall average are referred to as high- P_j disciplines and otherwise as low- P_j disciplines, except for insignificant ones. The high- P_j disciplines were COMP, MATH, PHYS, CHEM, CENG, MATE, ENGI, ENER, and EART; the P_j was consistently high in these closely related disciplines. Regarding the low- P_j disciplines, these encompassed not only initially-anticipated ones in soft science, such as ARTS, PSYC, SOCI, BUSI, and ECON but also some in medical science, such as HEAL, NURS, and DENT (see the appendix for abbreviations and the P_j values). The P_j values were used to normalize the number of papers of individuals in calculating the normalized index DWPP.

The variation of the DWPP among individuals is presented in the form of boxplot in each decile class (Fig. 3). The DWPP is approximately linear in relation to decile rank in the lower classes, whereas it exponentially increases over the 5th class, resulting in a skewed distribution overall. A notable feature is the extremely large variation observed in the highest 10th class.

Fig. 3 The boxplot of the normalized index for the number of papers (DWPP) from 17 studied Japanese universities. The plot in each decile class is drawn from the values of 1,491 researchers. The whiskers of the boxplot except for the upper one in the 10th class indicate minimum and maximum values; although not shown here, some outliers are in the 10th class at a maximum value of 18.1

Fig. 4 The relative frequency distribution of researcher rankings based on the DWPP for three universities: U1, U2, and U7

Percentile rank score of each university

Figure 4 shows the relative frequency distribution of researcher rankings based on the DWPP for three universities— U1, U2, and U7 (see details in Table 1), which are sorted out because of the characteristic distribution pattern. The relative frequency increases as the rank becomes higher for U1, whereas the opposite is true for U7. No obvious tendency is present for U2; the values are almost constant over all classes. The distribution patterns suggest that the research performance is superior for U1 dominated by high-ranking researchers, medium for U2 with balanced researcher rankings, and inferior for U7 dominated by low-ranking researchers. The cumulative distribution function, which is generated by summing up the relative frequency at each step from the lowest class, can manifest the difference more clearly (Fig. 5). It becomes convex for universities such as U1 with the values smaller than 0.1 in the lower classes, whereas it becomes concave for universities such as U7 with the values larger than 0.1 there. Regarding universities such as U2 with the values comparable to 0.1 in all classes, the function becomes close to the equilibrium line. The function visually represents correspondence between a given researcher's relative rank in the whole sample and that in each university. For

Fig. 5 Cumulative distribution function of researcher rankings based on the DWPP for three universities: U1, U2, and U7

example, if a researcher is at the 50th percentile in the whole, he/she is at the 39th in U1, 49th in U2, and 62nd in U7.

Table 3 shows the percentile rank scores of all universities calculated for three patterns described in “[Percentile rank score](#)” section; the table includes the median of the DWPP for each university as a benchmark score. For the later discussion, also shown is the score calculated from rankings based on publication productivity before normalization (PP) obtained in the same way as the DWPP, except for setting the P_j values of every discipline equal to unity (see the equation for the DWPP in Table 2).

Discussion

Interpretation of the DWPP and BPI

Comparing the publication productivity across the disciplinary boundaries is undesirable, and this discourages attempts to measure research activity of groups. In this study, therefore, the new percentile rank score, BPI, based on the normalized index for the number of papers, DWPP, was developed and tested with real publication data of faculty members from 17 Japanese universities. In calculating the DWPP, the disciplinary average of publication productivity P_j is very important for achieving accuracy. As described in “[Variation of publication productivity among disciplines and individuals](#)” section, the variation in the P_j values is consistent with the empirical knowledge that the publication productivity is higher than the overall average in hard science, while the opposite is true in soft science. The results provide assurance for further examination of the BPI calculated from the researcher rankings for the DWPP.

For the soft science disciplines, however, the P_j is ambiguous concerning its real meaning. The outputs are frequently written in non-English native languages because the necessity of writing in English does not become widely accepted; furthermore, they are published in the form of a book or monograph (e.g., Kyvik, 2003; Nederhof, 2006). As a result, they are prone to be excluded from registration to worldwide bibliometric databases; accordingly, the evaluation based on such databases tends to be disadvantageous to these disciplines (Mongeon and Paul-Hus, 2016) Although the normalization of the number of

Table 3 The percentile rank scores of all universities calculated for three patterns

University	Median	Percentile rank score				ROC	β_{high}^b	QS Ranking
		$R(6)^a$	$R(100)^a$	BPI_{dwpp}	BPI_{pp}			
U1	0.85 [1]	2.21 [1]	0.58 [1]	1.25 [1]	1.24 [1]	1%	46%	1–100
U2	0.66 [4]	1.93 [4]	0.57 [2]	1.02 [4]	1.03 [6]	– 1%	43%	101–300
U3	0.80 [2]	2.09 [2]	0.51 [5]	1.16 [2]	1.19 [3]	– 2%	49%	101–300
U4	0.65 [5]	1.99 [3]	0.52 [3]	1.04 [3]	0.98 [7]	6%	28%	101–300
U5	0.51 [13]	1.80 [7]	0.51 [4]	0.90 [9]	0.90 [9]	0%	54%	101–300
U6	0.55 [12]	1.73 [11]	0.49 [6]	0.87 [13]	0.83 [13]	5%	29%	501–750
U7	0.45 [15]	1.63 [14]	0.46 [13]	0.76 [14]	0.68 [16]	13%	18%	501–750
U8	0.58 [8]	1.78 [9]	0.49 [7]	0.90 [8]	0.88 [11]	2%	33%	501–750
U9	0.60 [7]	1.79 [8]	0.48 [8]	0.93 [7]	0.90 [8]	3%	33%	501–750
U10	0.56 [11]	1.73 [13]	0.47 [12]	0.88 [12]	0.86 [12]	2%	36%	501–750
U11	0.65 [6]	1.85 [6]	0.47 [11]	0.95 [6]	1.21 [2]	– 21%	86%	751–1000
U12	0.69 [3]	1.90 [5]	0.48 [10]	1.00 [5]	1.17 [4]	– 15%	76%	751–1000
U13	0.57 [9]	1.73 [12]	0.48 [9]	0.88 [10]	1.09 [5]	– 19%	79%	751–1000
U14	0.34 [17]	1.42 [17]	0.42 [14]	0.55 [17]	0.52 [17]	6%	21%	Out of Ranking
U15	0.45 [14]	1.55 [15]	0.40 [15]	0.69 [15]	0.88 [10]	– 22%	84%	Out of Ranking
U16	0.40 [16]	1.43 [16]	0.38 [16]	0.60 [16]	0.82 [14]	– 27%	90%	Out of Ranking
U17	0.57 [10]	1.73 [10]	0.35 [17]	0.88 [11]	0.71 [15]	23%	8%	No Entry

Notes. Median and all score except BPI_{pp} are based on the normalized publication productivity; the subscript pp denotes the score based on the publication productivity before normalization. Square brackets denote the university rankings based on the score. ROC represents the rate of change calculated as $(BPI_{dwpp} - BPI_{pp})/BPI_{pp} \times 100(\%)$ and β_{high} the manpower ratio obtained by dividing the total effort ratio in the high- P_j disciplines by the number of researchers for each university.

^a $R(6)$ and $R(100)$ were calculated following the method employed by Leydesdorff et al. (2011).

^bThe values less than 20% or more than 75% are highlighted in bold.

papers favors researchers in these disciplines thanks to smaller P_j , it does not resolve the problem with a large number of unindexed outputs. Furthermore, the potential authors pointed out by Koski et al. (2016) are expected to be more common in these disciplines than in the hard-science disciplines; although inclusion of their manpower to the calculation will cause a decrease in P_j and consequently an increase in the DWPP, a limit exists to the accuracy of the estimation for their manpower.

Considering these limitations, the DWPP can represent the publication productivity of only researchers in hard science. Fortunately, the 17 universities under study do not include specialized universities in the field of soft science, and 92% of the total manpower is devoted to hard science. Thus, considering that the scores in Table 3 represent the publication productivity of hard-science faculty would be reasonable, and the error stemmed from including soft-science faculty in the analysis is insignificant.

Figure 3 illustrates that the median and variation in the DWPP are remarkably higher in the highest class than those in the second-highest class. The results suggest that the top-ranked researchers largely contribute to the overall publication productivity of universities even if they are in a minority, which is consistent with findings from previous studies (e.g., Kwiek, 2016). The behavior of the DWPP is considered to justify the application of the

nonlinearly increasing weight to the difference between q_s and q_e in the calculation of the BPI (see “[Percentile rank score](#)” section).

Effects of normalizing the number of papers

Next, this study examines how the normalization of the number of papers affects the BPI. As shown in Table 3, the rate of change in the BPI, i.e., $(BPI_{dwpp} - BPI_{pp})/BPI_{pp} \times 100$ (%), exceeded 10% only for U7 and U17; the two universities enjoyed an advantage by the normalization. In contrast, the normalization brought disadvantages to five universities; the rate of change fell below -10% for U11, U12, U13, U15, and U16. All except U12 of the latter universities are institutes of technology (see Table 1). Although U12 is a comprehensive university, it is well known for high research performance in engineering disciplines. The ratio of manpower devoted to the high- P_j disciplines is less than 20% for the universities with a substantial increase in the score, whereas it is more than 75% for the universities with a substantial decrease. Considering the fact that the high- P_j disciplines correspond to the science and engineering disciplines (see “[Variation of publication productivity among disciplines and individuals](#)” section), the normalization naturally brings in better scores to the universities with a small ratio of researchers who work in these disciplines.

Figure 6 shows the change in the relative frequency distribution of researcher rankings based on the publication productivity before and after normalization, i.e., the PP and the DWPP, for three universities with characteristic changes: U5, U11, and U17. U11 is one of the universities with a substantial decrease in the BPI due to the normalization, and the distribution is shifted from the higher to lower ranks on the whole (Table 3). This is probably because the manpower devoted to the high- P_j disciplines accounts for 86% of the total, and the publication productivity of most researchers is decreased by the normalization. No significant change is seen in the distribution before and after normalization for U5, where 54% of manpower is devoted to the high- P_j disciplines; elevated and lowered rankings can be inferred as balanced. The largest improvement of the BPI is observed for the agricultural specialized university, U17, where only 8% of manpower is devoted to the high- P_j disciplines; one can see that the overall distribution has shifted upward after normalization.

Fig. 6 The change in the relative frequency of researcher rankings based on the publication productivity before and after normalization, i.e., the PP and the DWPP, for three universities. The open and solid bars denote the values for the PP and the DWPP, respectively

A certain correlation can be found between difficulty of entrance exams and research performance in Japanese universities. In light of this correlation, 3rd place of U4, 6th of U11, and 10th of U13 in the rankings for BPI_{dwpp} are probable, but 7th place of U4, 2nd of U11, and 5th of U13 for BPI_{pp} are implausible (see Table 3). The results suggest that the normalization of the number of papers is crucial for fair comparison of publication productivity of groups that incorporate researchers of a large variety of disciplines.

University rankings

In this section, rankings were compared based on the median of the DWPP and three percentile rank scores (see Table 3). To examine the similarities, the Spearman's rank correlation coefficients are calculated (Table 4). Although the correlation coefficients are generally high, the BPI has the highest correlation with the $R(6)$, followed by the median, and the least with the $R(100)$. The $R(100)$ had low correlations with both the $R(6)$ and median. As mentioned in “[Percentile rank score](#)” section, nonlinear weights with respect to the percentile rank are used in the calculation of the $R(6)$ and BPI, whereas linear weights are employed for the $R(100)$. By interviewing evaluators of several universities under study, a high degree of agreement was found between the rankings based on the $R(6)$ or BPI and their subjective evaluation on the 17 universities. The results suggest that applying the nonlinearly increasing weights as rank increases is reasonable when calculating the percentile rank scores.

In this analysis, the correlation of the rankings for the $R(6)$ with those for the BPI is quite high; the ranking difference for each university is two at a maximum, and the standard deviation is 0.97. The group size can, however, make a difference in the results. When evaluating small groups such as departments and research units, instead of universities, setting the number of classes small should make it difficult for the score to represent difference among groups; the problem can be avoided by making the evaluation at fine rank-intervals. In adopting the $R(6)$, reconsidering the scheme of rank classification and weighting is necessary.

The BPI in Equation (2) does not require such an adjustment; this is a merit of this score in exchange for flexibility on the weighting. Based on the principle that a person who is rare is valuable and the empirical fact that publication productivity of individuals exponentially increases with increasing rank (see “[Variation of publication productivity among disciplines and individuals](#)” section), the weight to each class is approximated by the reciprocal number of the cumulative relative frequency of researchers ranked above the class in the whole sample. The score is simple to apply also to the case of small groups or to the case without clear policy of evaluators on how to classify the ranks and how to weight the relative frequency in each class.

Table 4 Correlation matrix of university rankings based on the median and three percentile rank scores calculated from the normalized index DWPP. The values represent Spearman's rank correlation coefficients

	Median	$R(6)$	$R(100)$	BPI
Median	–			
$R(6)$	0.92	–		
$R(100)$	0.66	0.79	–	
BPI	0.96	0.98	0.77	–

Finally, the difference between rankings based on the BPI and simplest score, the median of the DWPP, is discussed. Although the standard deviation of the ranking differences, 1.33, is unexpectedly low, the largest difference appears for U5, i.e., 9th place for the BPI and 13th place for the median. U5 has the relative frequency of researcher rankings over 10% in 10th as well as 4th decile class or below (Fig. 6). Only universities U1–U5 have more than 10% of researchers in the highest class; the BPI places significant value on the existence of the exceptional researchers, and the median does not, resulting in the wide difference. The results suggest that the BPI or $R(6)$ is preferable to the median of the DWPP because weights can be adopted as a function of rank.

In this study, we examined the availability of the BPI; in the practices of research performance evaluation based on the score, statistical significance testing of differences would be crucial as discussed by Leydesdorff et al. (2011).

Future work

When comparing the QS World University Rankings to the rankings for the BPI, universities ranked in the range of 501–750 and 751–1000 in the former are often inversely correlated with the latter (see Tables 1 and 3). In particular, for U5, the QS ranking is in the second highest class, being in the range of 101–300, whereas the ranking for the BPI is in the middle of the 17 universities, being 9th. In light of the fact that universities with comparable value of BPI (i.e., U8, U9, U10, and U13) are in the range of 501–750 or below, U5 receives an unexpectedly high evaluation. Such differences are inevitable because the QS rankings are based on a reputation from academics or employers, faculty/student ratio, citations per paper, and international faculty or student ratio (Quacquarelli Symonds, 2020); the evaluation does not incorporate a factor of publication productivity. Regarding U5, the university is large in scale, having the second largest value of the institutional average of the normalized citation index FWCI after U1.

Improving the metrics of reputation used for the world university rankings, such as those reported by the THE or QS, is difficult due to their vagueness. Benchmarking for purely research performance is absolutely imperative for university executive board members to decide future policies. Undoubtedly, research performance cannot be measured only by publication productivity; rather, citation impact tends to be preferred in the evaluation. To the authors' knowledge, however, there exists no normalized citation index in which efficiency is considered in a similar way to the DWPP. Since the number of citations depends on publication year, more data are required for the normalization. A major future challenge is evaluating the research performance of groups in terms of both activity and impact by developing the normalized index for the number of citations in a similar way to the DWPP with an expanded data set.

Another issue to be solved is which document types should be included in the analysis. As mentioned above, some universities are placed at lower ranks in the QS rankings but at higher ranks in the rankings for the BPI. The feature of these universities, specifically U11, U12, and U13, is that the ratio of researchers related to the science and engineering disciplines is high, and that the number of conference papers accounts for over 35% of the total. As a number of researchers of the 17 Japanese universities specialized in computer science, which places importance on the conference papers (Vrettas and Sanderson, 2015), we decided not to exclude this document type from the analysis to evaluate as many researchers as possible. In the context of subjects or purposes of the evaluation, however,

deep deliberation would be needed on the inclusion of the conference paper or on giving the same weight to it as to the article and review.

Conclusion

The difficulty in evaluating the research performance of groups is attributable to two factors: 1) difference of population size or discipline of group members and 2) skewed distribution of the research performance of individuals. This study attempted to overcome this difficulty, focusing on research performance based on publication productivity. We employed the normalized index for the number of papers, DWPP, to calculate a new percentile rank score, BPI, developed on the basis of the principle that a person who is rare is valuable. In calculating the DWPP, publication efficiency was considered by introducing the concept of manpower devoted to the publication, and the disciplinary variation in the publication intensity was corrected by the disciplinary averages of publication productivity P_j .

The BPI based on the DWPP was tested with publication data collected at the individual level for faculty members of 17 universities in Japan. Regarding the P_j , significantly high values were observed in the science and engineering disciplines, except for biological disciplines, while low values were distinct in soft science and some medical disciplines. The normalization of the number of papers increased the BPI by over 10% for the universities with the manpower devoted to the high- P_j disciplines less than 20% of the total, resulting in more plausible university rankings. The results suggest that the DWPP can alleviate the problem in which universities with a small portion of faculty members working in the high- P_j disciplines tend to be disadvantaged in the research performance evaluation. The university rankings for the BPI showed a high correlation with those for the $R(6)$ developed by Leydesdorff et al. (2011), and they were consistent with the subjective judgment by evaluators of several universities under study. The advantage of the BPI over the $R(6)$ is that it has no room for arbitrariness in determining the scheme of rank classification and the weights given to each rank class.

Appendix

See Table 5.

Table 5 Abbreviation of ASJC27 and the P_j values shown in Fig. 2

Discipline	Abbreviation	P_j
Computer Science	COMP	9.50
Mathematics	MATH	6.86
Physics and Astronomy	PHYS	8.41
Chemistry	CHEM	7.88
Chemical Engineering	CENG	8.19
Materials Science	MATE	9.22
Engineering	ENGI	9.01
Energy	ENER	8.70
Environmental Science	ENVI	5.09
Earth and Planetary Sciences	EART	6.03
Agricultural and Biological Sciences	AGRI	4.47
Biochemistry, Genetics and Molecular Biology	BIOC	4.62
Immunology and Microbiology	IMMU	4.63
Veterinary	VETE	4.39
Medicine	MEDI	4.95
Pharmacology, Toxicology and Pharmaceutics	PHAR	5.45
Health Professions	HEAL	3.66
Nursing	NURS	3.56
Dentistry	DENT	2.77
Neuroscience	NEUR	4.88
Arts and Humanities	ARTS	2.46
Psychology	PSYC	3.53
Social Sciences	SOCI	2.89
Business, Management and Accounting	BUSI	3.20
Economics, Econometrics and Finance	ECON	3.75
Decision Sciences	DECI	5.44
Multidisciplinary	MULT	5.03

Acknowledgements The authors would like to thank several universities for not only providing us the author IDs of their faculty members, but also for their helpful comments on the results. This work was supported by JSPS KAKENHI Grant Number JP18K18333.

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Abramo, G., & D'Angelo, C. A. (2014). How do you define and measure research productivity? *Scientometrics*, 101(2), 1129–1144.
- Avanesova, A. A., & Shamliyan, T. A. (2018). Comparative trends in research performance of the Russian universities. *Scientometrics*, 116(3), 2019–2052.

- Baccini, A., Barabesi, L., Cioni, M., & Pisani, C. (2014). Crossing the hurdle: The determinants of individual scientific performance. *Scientometrics*, 101(3), 2035–2062.
- Barrot, J. S. (2017). Research impact and productivity of Southeast Asian countries in language and linguistics. *Scientometrics*, 110(1), 1–15.
- Bornmann, L., Leydesdorff, L., & Mutz, R. (2013). The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits. *Journal of Informetrics*, 7(1), 158–165.
- Bornmann, L., & Marx, W. (2014). How to evaluate individual researchers working in the natural and life sciences meaningfully? A proposal of methods based on percentiles of citations. *Scientometrics*, 98(1), 487–509.
- Elsevier (2020). Topic prominence in science. <https://www.elsevier.com/solutions/scival/releases/topic-prominence-in-science>. Accessed 17 September 2020.
- Gul, S., Nisa, N. T., Shah, T. A., Gupta, S., Jan, A., & Ahmad, S. (2015). Middle East: Research productivity and performance across nations. *Scientometrics*, 105(2), 1157–1166.
- Guskov, A. E., Kosyakov, D. V., & Selivanova, I. V. (2018). Boosting research productivity in top Russian universities: The circumstances of breakthrough. *Scientometrics*, 117(2), 1053–1080.
- Khor, K. A., & Yu, L. G. (2016). Influence of international co-authorship on the research citation impact of young universities. *Scientometrics*, 107(3), 1095–1110.
- Koski, T., Sandström, E., & Sandström, U. (2016). Towards field-adjusted production: Estimating research productivity from a zero-truncated distribution. *Journal of Informetrics*, 10(4), 1143–1152.
- Kotsemir, M., & Shashnov, S. (2017). Measuring, analysis and visualization of research capacity of university at the level of departments and staff members. *Scientometrics*, 112(3), 1659–1689.
- Kwiek, M. (2016). The European research elite: a cross-national study of highly productive academics in 11 countries. *Higher Education*, 71(3), 379–397.
- Kyvik, S. (2003). Changing trends in publishing behaviour among university faculty, 1980–2000. *Scientometrics*, 58(1), 35–48.
- Leydesdorff, L., Bornmann, L., Mutz, R., & Opthof, T. (2011). Turning the tables on citation analysis one more time: Principles for comparing sets of documents. *Journal of the American Society for Information Science and Technology*, 62(7), 1370–1381.
- Moed, H. F. (2005). *Citation analysis in research evaluation* (Chap. 22, pp. 271–284). Springer.
- Mongeon, P., & Paul-Hus, A. (2016). The journal coverage of Web of Science and Scopus: A comparative analysis. *Scientometrics*, 106(1), 213–228.
- National Science Board. (2010). *Science and engineering indicators 2010*. VA: Arlington.
- Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, 66(1), 81–100.
- Quacquarelli Symonds (2020). QS World University Rankings 2020: methodology. <https://www.topuniversities.com/qs-world-university-rankings/methodology>. Accessed 17 December 2019.
- Schubert, A., & Glänzel, W. (1984). A dynamic look at a class of skew distributions. A model with scientometric applications. *Scientometrics*, 6(3), 149–167.
- Schubert, A., Glänzel, W., & Braun, T. (1988). Against absolute methods. Relative scientometric indicators and relational charts as evaluation tools. In: van Raan, A. F. J., editor, *Handbook of Quantitative Studies of Science and Technology*, chapter 5, pages 137–176. North-Holland Publishing Company.
- Schubert, A., & Telcs, W. (1986). Publication potential—an indicator of scientific strength for cross-national comparisons. *Scientometrics*, 9(5–6), 231–238.
- Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, 43(9), 628–638.
- Times Higher Education (2019). The world university rankings 2020: methodology. <https://www.timeshighereducation.com/world-university-rankings/world-university-rankings-2020-methodology>. Accessed 17 December 2019.
- van Raan, A. F. J. (2006). Comparison of the hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, 67(3), 491–502.
- Vinluan, L. R. (2012). Research productivity in education and psychology in the Philippines and comparison with ASEAN countries. *Scientometrics*, 91(1), 277–294.
- Vrettas, G., & Sanderson, M. (2015). Conferences versus journals in computer science. *Journal of the Association for Information Science and Technology*, 66(12), 2674–2684.
- Waltman, L., & Van Eck, N. J. (2015). Field-normalized citation impact indicators and the choice of an appropriate counting method. *Journal of Informetrics*, 9, 872–894.
- Yamamoto, K., & Ishikawa, M. (2017). A normalized index to quantify publication productivity across disciplinary boundaries. *International Journal of Institutional Research and Management*, 1(1), 67–81.