Journal of Informetrics 5 (2011) 101–113

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

The effects and their stability of field normalization baseline on relative performance with respect to citation impact: A case study of 20 natural science departments

Cristian Colliander[a,*], Per Ahlgren[b]

[a] Department of Sociology, Inforsk, Umeå University, SE 901 87 Umeå, Sweden
[b] Department of e-Resources, University Library, Stockholm University, SE 106 91 Stockholm, Sweden

ARTICLE INFO

Article history:
Received 6 June 2010
Received in revised form 3 September 2010
Accepted 21 September 2010

Keywords:
Stability analysis
Field normalization baseline
Journal
ISI/Thomson Reuters subject category
Essential Science Indicators field
Citation impact

ABSTRACT

In this paper we study the effects of field normalization baseline on relative performance of 20 natural science departments in terms of citation impact. Impact is studied under three baselines: journal, ISI/Thomson Reuters subject category, and Essential Science Indicators field. For the measurement of citation impact, the indicators item-oriented mean normalized citation rate and Top-5% are employed. The results, which we analyze with respect to stability, show that the choice of normalization baseline matters. We observe that normalization against publishing journal is particular. The rankings of the departments obtained when journal is used as baseline, irrespective of indicator, differ considerably from the rankings obtained when ISI/Thomson Reuters subject category or Essential Science Indicators field is used. Since no substantial differences are observed when the baselines Essential Science Indicators field and ISI/Thomson Reuters subject category are contrasted, one might suggest that people without access to subject category data can perform reasonable normalized citation impact studies by combining normalization against journal with normalization against Essential Science Indicators field.

1. Introduction

The use of bibliometrics as a tool for research evaluation is widespread nowadays. Usually, scientific impact is measured on the basis of received citations, and citation-based evaluation is an area of increasing importance (for instance Steele, Butler, & Kingsley, 2006; Visser & Nederhof, 2007). This increasing importance is connected to the need for research funding entities, like universities and government offices, to assess the quality of applicants (individuals, research groups, departments, institutions, and so on). It is known, though, that citation volumes varies across scientific fields. For example, the citation volume in mathematics is considerably lower than in biology. Such field differences in citation volume render cross-field performance comparisons more difficult.

In order to deal with the citation volume discrepancy, *field normalization*, to compare the citation rates of the target publications – the publications that are evaluated with respect to received citations – with citation rates of publications that are similar from a subject point of view, can be applied (Schubert & Braun, 1993, 1996). The journals in which the target publications are published may act as a normalization baseline. Another possibility is to use journal sets categorized

* Corresponding author. Tel.: +46 (0)90 786 68 71.
E-mail address: cristian.colliander@soc.umu.se (C. Colliander).

1751-1577/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2010.09.003

according to some pre-defined classification system for normalization. There are, then, several possibilities for the choice of normalization baseline.

Zitt, Ramanana-Rahary, and Bassecoulard (2005) investigated the effects of normalization baseline on citation impact of articles indexed in Science Citation Index. Five levels of aggregation (baselines) were used, among them journal and ISI/Thomson Reuters subject category. A lack of stability across levels was observed. With regard to top-cited proportions, the results showed that the cited set was wholly dependent on the level of aggregation employed.

Citation impact of UK university departments, in the three fields Biological Sciences, Physics and Psychology, and for each field under grade awarded to the departments by external peer review, was studied by Adams, Gurney, and Jackson (2008). The performance was studied for each of three levels of article aggregation. The levels were, from narrower to broader: journal, ISI/Thomson Reuters subject category, and Research Assessment Exercise (RAE) Unit of Assessment. It was found that the performance within the departments was dependent on the level of aggregation. The performance, irrespective of grade and field, were considerably better when received citations were normalized relative to ISI/Thomson Reuters subject categories compared to normalization against journals. However, regarding rank orders of departments, citation impact normalized against ISI/Thomson Reuters subject categories was significantly correlated with citation impact normalized against RAE Units of Assessment. Glänzel, Thijs, Schubert, and Debackere (2009) studied citation impact of European research institutions in relation to three different levels of publication aggregation. The highest level consisted of 12 major fields, whereas the lowest level was ISI/Thomson Reuters subject category. The intermediate level consisted of 60 subfields. A strong correlation between impact normalized against ISI/Thomson Reuters subject categories and impact normalized against subfields was observed when the institutions were studied as a totality, whereas impact normalized against these two levels correlated less well with impact normalized against major fields.

The effects of field normalization baseline on citation impact is a main theme in the works referred to in the two preceding paragraphs. Ball, Mittermaier, and Tunger (2009) comment on the issue in question, while the importance of using appropriate sets of publications against which to field normalize citations rates has been pointed out by Kostoff (2002). Not only might the lack of stability across baseline alternatives be regarded as problematic for citation-based research evaluation, but one also faces the problem of interpreting if observed differences between analyzed units due to the choice of normalization baseline are substantial enough to warrant in-depth analysis of the cause.

The purpose of this work is twofold: (a) to study the effects of field normalization baseline on relative performance with respect to citation impact of 20 Stockholm University (SU) natural science departments, and (b) to analyze the stability of the obtained results. With regard to (a), impact is studied under three baselines. These are, from narrower to broader: journal, ISI/Thomson Reuters subject category, and Essential Science Indicators field. We measure citation impact with the indicators item-oriented mean normalized citation rate and Top-5%, and we give a more precise presentation of the latter indicator than what is typically given in the literature. With regard to (b), to our knowledge, the kind of stability analysis we perform in this work has not been performed in other studies in the area of evaluative bibliometrics.

The remainder of the paper is organized as follows. Data and methods are described in Section 2, while the results are reported in Section 3. In Section 4, the results are discussed, and conclusions are put forward.

2. Data and methods

2.1. Data

The data source of the study is Web of Science. With regard to the target publications and publications belonging to normalization baselines, SCI-EXPANDED and SSCI were utilized. For citing publications, all five citation indices were taken into account: SCI-EXPANDED, SSCI, A&HCI, CPCI-S and CPCI-SSH.

For each SU department in the study, a corresponding query was constructed, and a set of bibliographic records was retrieved and downloaded in the middle of January 2010. We worked exclusively with documents[1] of the types article, proceedings paper and review and such that their database year belongs to the interval 2005–2007. Thus, each retrieved record represents either an article, a proceedings paper or a review, and the document it represents was indexed in Web of Science during 2005–2007.

We define a retrieved record as *relevant*, for a given query, if at least one address in the record applies to the department that corresponds to the query. The 20 queries were deliberately recall-oriented in order to avoid missing relevant records. Some of the queries retrieved several non-relevant records, which were eliminated from the corresponding record sets.

In the first week of February 2010, all data needed for the generation of baseline citation values were downloaded from Web of Science. The endpoint of the citation window is thereby the first week of February 2010. We used the downloaded data to update, for each department, and each record in the record set for the department, the earlier obtained citation frequency of the document represented by the record.

Regarding the document type proceedings paper, all papers of this type considered in the study, including non-SU papers, is published in a journal. Therefore, we did not discriminate between articles and proceedings papers. Consequently, we worked with two values on the parameter document type: article/proceedings paper and review. This yielded that an SU

[1] We use the terms *document* and *publication* synonymously in this work.

article, or an SU proceedings paper, was compared to both articles and proceeding papers, whereas an SU review was compared to reviews.

In Appendix A, we list the English names of the 20 SU departments included in the study, together with corresponding abbreviations.

2.2. Field normalization baselines

We consider three field normalization baselines: journal (J_norm), ISI/Thomson Reuters subject category (SC_norm), and Essential Science Indicators field (ESI_norm). The number of subject categories is 227, 172 (SCI-EXPANDED) plus 55 (SSCI), while the number of ESI fields is 22.[2] For J_Norm, a document d belonging to a unit of analysis is compared to documents in its journal. For SC_Norm, d is compared to documents in the subject category (or categories) of its journal. Regarding ESI_Norm, d is compared to documents in the ESI field of its journal.

Under each normalization baseline, two indicators were used: item-oriented mean normalized citation rate (Lundberg, 2007) and Top-5%. The latter indicator is similar to the Highly Cited Papers Index put forward by Tijssen, Visser, and van Leeuwen (2002).

2.2.1. J_Norm and ESI_norm

We treat J_norm and ESI_norm separate from SC_norm since the calculations of the indicators differ between them. Let A be a unit of analysis and n the number of documents belonging to A. Let C_i be the journal or ESI field for d_i, the ith document belonging to A, and let c_i be the citation frequency for the ith document belonging to A.

The *item-oriented mean normalized citation rate* for A is given by

$$\frac{\sum_{i=1}^{n} c_i/\mu_i}{n} \tag{1}$$

$$\text{where} \quad \mu_i = \frac{\sum_{j=1}^{m_i} c_j}{m_i}$$

where m_i is the number of documents, with the same database year and of the same document type (article/proceedings paper or review) as the ith document belonging to A, in C_i, and c_j is the citation frequency for the jth of these documents. μ_i is the mean number of citations received by documents belonging to C_i, documents with the same database year and of the same document type as the ith document belonging to A. Thus, to obtain the normalized citation rate for the ith document belonging to A, the citation frequency for the document is divided by an expected frequency with regard to the journal or ESI field to which the document belongs, where database year and document type are taken into account. Eq. (1) gives the mean across these n normalized citation rates.

Besides the issue of creating reasonable normalization baselines, there is the question whether to construct the relative impact indicator by calculating a ratio of means or a mean of ratios. Traditionally, perhaps the most common approach has been to divide the mean number of citations per publication for an analyzed unit with the mean number of citations received by the publications in the normalization baseline (see, for example, Moed, De Bruin, & van Leeuwen, 1995). However, as Lundberg (2007) points out, by doing so one gives more weight to older publications and to publications in fields with dense citation traffic (which have higher normalization baseline values). He suggests therefore that normalization preferably should be carried out on the level of individual publications, instead of the aggregated level, which is the case in the ratio of means approach.

Recently, Opthof and Leydesdorff (2010) also question the rationale in weighting papers differently and proposed normalization on a publication-by-publication basis. The approach put forward by Lundberg and advocated by Opthof and Leydesdorff was commented on by van Raan, van Leeuwen, Visser, Eck, and Waltman (2010). These authors agreed that it does not seem very reasonable to weight publications differently based on the fact that they are from different fields, but that weighting could still be valid on the basis of publication age and publication type. Further, the authors demonstrate that normalization based on a publication-by-publication approach is more sensitive (i.e., less robust) to outliers compared with normalization that takes place on an aggregated level.

In this work, we use the publication-by-publication approach, which should be clear from Eq. (1) above and from Eq. (3) below.

We now turn our attention to the indicator Top-5%. For each d_i, we generated the 95th percentile of the citation distribution for the documents, with the same database year and of the same document type as d_i, in C_i. Let k_i (v_i) denote the percentile (distribution). Let $n_{5\%}$ be the number of documents d_i ($1 \leq i \leq n$) such that $c_i > k_i$. The *Top-5%* value for A is given by

$$\frac{(n_{5\%}/n)}{\mu_{5\%}} \tag{2}$$

[2] For ISI/Thomson Reuters subject categories, see http://science.thomsonreuters.com/mjl/. For Essential Science Indicators fields, see http://sciencewatch.com/about/met/journallist/.

$$\text{where} \quad \mu_{5\%} = \frac{\sum_{i=1}^{n} N_i [\mu_{5\%}]_i}{\sum_{i=1}^{n} N_i}$$

where N_i is the number of observations in v_i (equal to the number of documents underlying v_i), and $[\mu_{5\%}]_i$ the share of observations in v_i that are greater than k_i. In Eq. (2), the share of A documents with citation frequencies greater than their corresponding 95th percentiles is divided by an expected share (approximately 0.05, or 5%) with regard to the journals or ESI fields to which the A documents belongs, where database year and document type are taken into account.

In this work, where we work with discrete distributions, we did not use interpolation to generate the percentiles. Regarding J_Norm and ESI_Norm we used, for a given citation distribution, the following definition of the pth percentile: the smallest x such that $F(x) \geq p$, where F is the empirical cumulative distribution function for the distribution.

2.2.2. SC_Norm

Regarding SC_norm, we take into account that the journal of a document, and thereby the document itself, might belong to more than one subject category. Again, let A be unit of analysis and n the number of documents belonging to A. Let C_{iq} be the qth subject category for the ith document belonging to A, and let q_i (c_i) be the number of subject categories (the citation frequency) for the document.

The *item-oriented mean normalized citation rate* for A is given by

$$\frac{\sum_{i=1}^{n} \bar{x}_i}{n} \tag{3}$$

$$\text{where} \quad \bar{x}_i = \frac{\sum_{q=1}^{q_i} c_i / \mu_{iq}}{q_i}$$

$$\text{where} \quad \mu_{iq} = \frac{\sum_{j=1}^{m_{iq}} c_j / F_j}{\sum_{j=1}^{m_{iq}} 1/F_j}$$

where m_{iq} is the number of documents, with the same database year and of the same document type as the ith document belonging to A, in C_{iq}, and F_j (c_j) the number of subject categories (the citation frequency) for the jth of these documents. Note that fractionalization is applied. A document belonging to C_{iq} and such that its journal belongs to, say, three categories contributes with 1/3 to C_{iq}, and 1/3 of its citation frequency is associated with C_{iq}. To obtain the normalized citation rate for the ith document belonging to A, i.e., to obtain $\bar{x}_i$, the citation frequency for the document is first divided by one or more expected frequencies with regard to the subject categories to which the document belongs, where database year and document type are taken into account. Then the sum of the ratios is divided by the number of subject categories for the document. Eq. (3) gives the mean across these n normalized citation rates.

For Top-5%, consider the citation distribution $v_{iq} = (c_1, \ldots, c_{m_{iq}})$, where the values are ordered ascendingly, for the documents, with the same database year and of the same document type as d_i, in C_{iq} (the qth subject category for d_i). For each c_j in v_{iq}, we assign $1/F_j$, the fraction the corresponding jth document contributes to C_{iq}, to c_j as a weight. We define the *weighted* empirical cumulative distribution function for v_{iq}, $F_{iq(w)}$, as

$$F_{iq(w)}(x) = \frac{\sum_{c_j \leq x} 1/F_j}{\sum_{j=1}^{m_{iq}} 1/F_j} \tag{4}$$

Now, we define the *weighted* pth percentile for v_{iq} as the smallest x such that $F_{iq(w)}(x) \geq p$.

For each d_i we generated the weighted 95th percentile of each distribution v_{iq} ($1 \leq q \leq q_i$). Let k_{iq} denote the 95th percentile of a given v_{iq}. The *Top-5%* value for A is given by

$$\frac{(y_{5\%}/n)}{\mu_{5\%}} \tag{5}$$

$$\text{where} \quad y_{5\%} = \sum_{i=1}^{n} \sum_{q=1}^{q_i} a_{iq}$$

$$\text{where} \quad a_{iq} = \begin{cases} 1/q_i & \text{if } c_i > k_{iq} \\ 0 & \text{if } c_i \leq k_{iq} \end{cases}$$

$$\text{and} \quad \mu_{5\%} = \frac{\sum_{i=1}^{n} \sum_{q=1}^{q_i} X_{iq} [\mu_{5\%}]_{iq}}{\sum_{i=1}^{n} \sum_{q=1}^{q_i} X_{iq}}$$

where $X_{iq} = \sum_{j=1}^{m_{iq}} 1/F_j$, and $[\mu_{5\%}]_{iq}$ the sum of weights $1/F_j$ ($1 \leq j \leq m_{iq}$) such that $c_j > k_{iq}$, as a proportion of the totality of these weights (i.e., as a proportion of X_{iq}). $y_{5\%}$ is such that a given A document d_i contributes with l/q_i, where $l \in \{0, 1, \ldots, q_i\}$, to the sum. The sum is divided by n, which gives the share of *fractionalized* documents belonging to A with citation frequencies greater than their corresponding 95th percentiles. This share is divided by an expected share (approximately 0.05, or 5%) with regard to the subject categories to which the A documents belongs, where database year and document type are taken into account.

2.3. Stability analysis

We use subsample descriptive inference (Lunneborg, 2000) to evaluate the stability of our results. Subsampling is a resampling technique, which can be applied when the data is neither randomly sampled nor randomly allocated (i.e., neither population nor causal inference is feasible). Briefly, instead of talking about statistical significance (or lack thereof) we talk about stability, and a stable result is one that is not materially influenced by including or excluding specific cases (here documents) in the analysis.

Following Lunneborg (2000) we assess the stability of the indicator values by repeating the calculations on a series of subsamples. The subsamples are created by selecting, at random and without replacement, 90% of the documents for a given department. Now, if there are N documents attributed to a specific department and we let m be the result of rounding $N \times 0.9$ to the nearest integer, there are $N!/[m!(N-m)!]$ distinct subsamples. Depending on the size of N, it is generally not practicable (or necessary) to enumerate all possible subsamples; in this study we use 5000 randomly selected subsamples to create distributions of indicator values for each department and under each normalization baseline. Further, we can now use the variability among indicator values generated by the subsample procedure to construct stability intervals, e.g., by equating the lower (upper) bound with the 5th (95th) percentile in these distributions. To some extent stability intervals resembles confidence intervals. However, the latter reflects uncertainty about a population parameter, whereas the former reflects our uncertainty about the calculated indicators of the data set at hand, here based on the variability of the documents citation impact.

The length and symmetry of the stability intervals depend on properties of the observed citation distributions. For example, while the degree of right skewness of the observed citation distribution is negatively correlated with the quotient between the distance from the upper bound to the indicator value and the distance from the indicator value to the lower bound, the standard deviation of the observed citation distribution is positively correlated with the length of the interval.

The choice of percentiles for constructing the stability intervals is somewhat arbitrary. A 95% stability interval, for example, is more conservative than, say, an 80% interval. In this study we make use of 90% stability intervals. However, we also report 85% and 95% intervals for comparisons. These comparisons show that the choice is not vital: the conclusions do not change and so the patterns derived from the data are robust in this respect.

If we take a conservative approach, then one might say that – for a given indicator and normalization baseline – if two departments have overlapping stability intervals this indicates that there is no substantial difference between these departments. In essence, when we compare departments, we have little ground for stating that one performs better than another if the difference can be attributed to one or a few values in the underlying empirical citation distribution, i.e., given that we are interested in the overall citation impact of the departments, we do not consider differences of this kind as stable.

Moreover, we let the maximum rank (max) of a given department with regard to a specific indicator under a given normalization baseline be equal to the rank that this department is attributed when we take its indicator value to be equal to its upper bound, and the indicator value for every other department to be equal to their lower bounds. The minimum rank (min) is defined analogously. We utilize mid-rank assignment for handling potential ties (i.e., using the mean rank for tied observations) so under each normalization baseline and each indicator we have for every department an interval of ranks:

$$[\min, \quad \max] = \left\{ \frac{n}{2} : n \in \mathbb{N}, \quad (\min \times 2) \leq n \leq (\max \times 2) \right\}$$

We say that the ranking of a specific department, with respect to a given indicator, differ in a substantial way (based on the notion of stability) between two normalization baselines if and only if the intersection between the two sets of ranks equals the empty set.

3. Results

In this section, we report the effects of using different normalization baselines when ranking the departments with respect to item-oriented mean normalized citation rate and Top-5%. The degree of correspondence between rankings is measured by Kendall's tau-b, while substantial differences in rank are identified by applying subsample descriptive inference. The effects of normalization baseline on the absolute performance are also briefly illustrated.

Fig. 1 depicts the association between the values of item-oriented mean normalized citation rate under ESI_norm versus SC_norm.

The association between the department rankings under the two given baselines is rather high with a value of 0.82 on Kendall's tau-b. However, several shifts in rank can be observed as illustrated in Figs. 2 and 3.

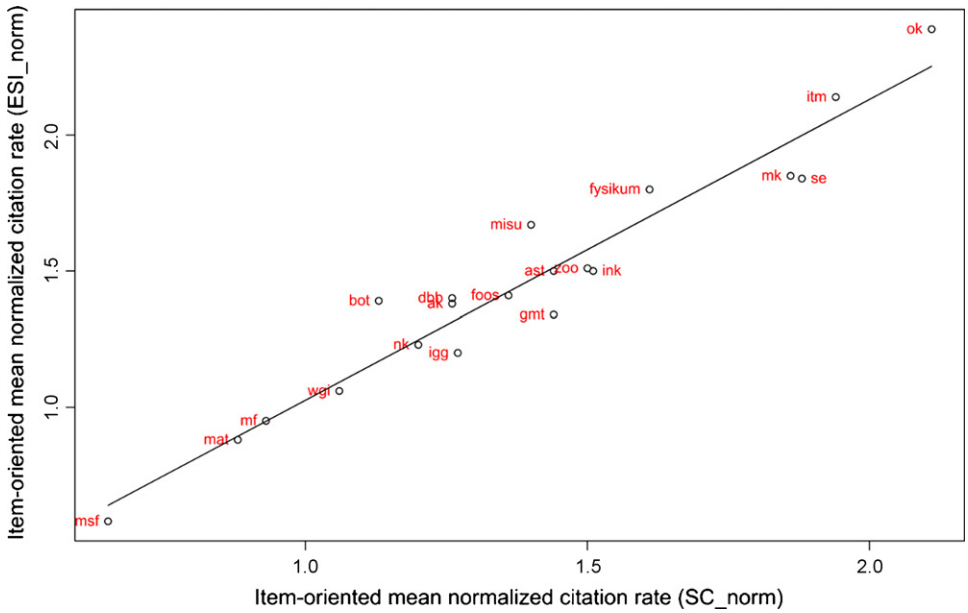

Fig. 1. Association between item-oriented mean normalized citation rate under ESI_norm and SC_norm.

For example, department gtm reach rank 12.5 under SC_norm but rank 7 when the citation rate is normalized with respect to ESI fields. Nonetheless, we observe no substantial differences in rank between the departments when we compare these two normalization baselines: for each department *D*, the intersection of the two sets of ranks for *D*, one set for SC_norm and one for ESI_norm, is non-empty (cf. Section 2.2). Continuing with comparing the rank order of departments based on item-oriented mean normalized citation rate under SC_norm and ESI_norm contra J_norm, a slightly different picture emerges. Kendall's tau-b between rankings under SC_norm and J_norm drops to 0.76 and even further, to 0.67, when the association between rankings under ESI_norm and J_norm is considered. Moreover, there now emerge a few differences in rank which are to be considered substantial. Department nk have a substantial higher rank under J_norm compared to the case when SC_norm or ESI_norm is applied, i.e., the intersection of the set of ranks for J_norm and the set of ranks for SC_norm (the set of ranks for ESI_norm) is equal to the empty set. The opposite is true for department misu when the rankings according to ESI_norm and J_norm are contrasted. Here misu have a substantially higher rank under ESI_norm.

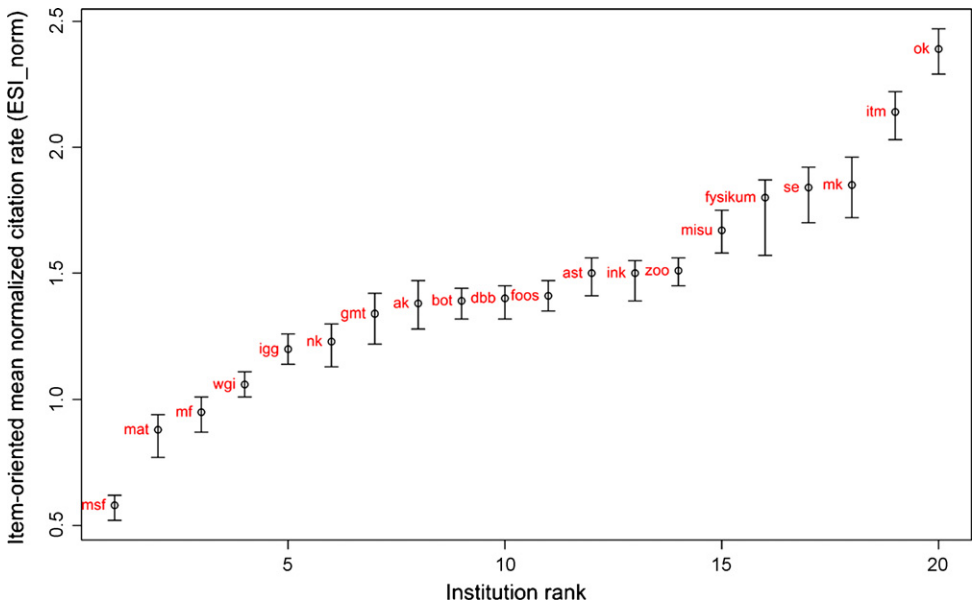

Fig. 2. Department ranking and 90%-stability bars under ESI_norm.

Table 1
Item-oriented mean normalized citation rate for the 20 SU departments under each normalization baseline.

Department	J_norm (rank); 85%, **90%**, 95% lb-ub	SC_norm (rank); 85%, **90%**, 95% lb-ub	ESI_norm (rank); 85%, **90%**, 95% lb-ub
ak	1.17 (11); 1.09–1.24, **1.08–1.25**, 1.06–1.26	1.26 (7.5); 1.18–1.32, **1.16–1.33**, 1.15–1.34	1.38 (8); 1.29–1.46, **1.28–1.47**, 1.25–1.48
ast	1.29 (16); 1.24–1.34, **1.22–1.34**, 1.21–1.35	1.44 (12.5); 1.36–1.49, **1.35–1.50**, 1.33–1.50	1.50 (12.5); 1.42–1.56, **1.41–1.56**, 1.39–1.57
bot	0.98 (6); 0.95–1.01, **0.95–1.01**, 0.94–1.01	1.13 (5); 1.09–1.17, **1.08–1.17**, 1.08–1.18	1.39 (9); 1.33–1.43, **1.32–1.44**, 1.31–1.45
dbb	1.03 (7); 1.00–1.06, **0.99–1.06**, 0.99–1.07	1.26 (7.5); 1.21–1.30, **1.21–1.31**, 1.19–1.31	1.40 (10); 1.33–1.45, **1.32–1.45**, 1.30–1.46
foos	0.96 (5); 0.93–0.98, **0.92–0.99**, 0.92–0.99	1.36 (10); 1.31–1.40, **1.30–1.40**, 1.29–1.41	1.41 (11); 1.36–1.46, **1.35–1.47**, 1.34–1.47
fysikum	1.34 (18); 1.15–1.39, **1.14–1.40**, 1.12–1.40	1.61 (16); 1.43–1.68, **1.41–1.68**, 1.39–1.69	1.80 (16); 1.60–1.86, **1.57–1.87**, 1.55–1.88
gmt	1.12 (10); 1.06–1.18, **1.05–1.19**, 1.04–1.19	1.44 (12.5); 1.33–1.53, **1.31–1.53**, 1.28–1.54	1.34 (7); 1.23–1.42, **1.22–1.42**, 1.20–1.43
igg	1.05 (9); 1.00–1.08, **1.00–1.09**, 0.99–1.09	1.27 (9); 1.21–1.32, **1.20–1.32**, 1.18–1.33	1.20 (5); 1.15–1.25, **1.14–1.26**, 1.13–1.26
ink	1.26 (14.5); 1.21–1.30, **1.20–1.31**, 1.19–1.31	1.51 (15); 1.44–1.56, **1.43–1.57**, 1.41–1.57	1.50 (12.5); 1.41–1.55, **1.39–1.55**, 1.37–1.56
itm	1.41 (19); 1.34–1.46, **1.33–1.46**, 1.32–1.47	1.94 (19); 1.85–2.01, **1.84–2.02**, 1.81–2.03	2.14 (19); 2.05–2.21, **2.03–2.22**, 2.00–2.23
mat	0.90 (3); 0.82–0.95, **0.80–0.96**, 0.78–0.96	0.88 (2); 0.77–0.93, **0.75–0.94**, 0.73–0.94	0.88 (2); 0.79–0.93, **0.77–0.94**, 0.75–0.94
mf	0.71 (2); 0.66–0.75, **0.66–0.76**, 0.64–0.76	0.93 (3); 0.87–0.98, **0.86–0.98**, 0.84–0.99	0.95 (3); 0.89–1.00, **0.87–1.01**, 0.86–1.01
misu	1.04 (8); 1.00–1.09, **0.99–1.09**, 0.98–1.10	1.40 (11); 1.33–1.47, **1.31–1.48**, 1.29–1.49	1.67 (15); 1.60–1.75, **1.58–1.75**, 1.56–1.77
mk	1.24 (13); 1.17–1.29, **1.16–1.30**, 1.14–1.31	1.86 (17); 1.75–1.96, **1.74–1.97**, 1.71–1.99	1.85 (18); 1.74–1.95, **1.72–1.96**, 1.69–1.97
msf	0.70 (1); 0.65–0.75, **0.64–0.75**, 0.63–0.76	0.65 (1); 0.59–0.69, **0.59–0.70**, 0.58–0.70	0.58 (1); 0.53–0.62, **0.52–0.62**, 0.51–0.63
nk	1.23 (12); 1.17–1.28, **1.16–1.29**, 1.15–1.29	1.20 (6); 1.11–1.26, **1.10–1.27**, 1.08–1.28	1.23 (6); 1.15–1.29, **1.13–1.30**, 1.11–1.31
ok	1.53 (20); 1.48–1.57, **1.47–1.57**, 1.46–1.58	2.11 (20); 2.04–2.17, **2.03–2.18**, 2.01–2.19	2.39 (20); 2.30–2.46, **2.29–2.47**, 2.27–2.49
se	1.33 (17); 1.26–1.38, **1.25–1.38**, 1.23–1.39	1.88 (18); 1.74–1.97, **1.72–1.97**, 1.67–1.98	1.84 (17); 1.72–1.92, **1.70–1.92**, 1.67–1.94
wgi	0.94 (4); 0.90–0.97, **0.90–0.97**, 0.89–0.98	1.06 (4); 1.02–1.10, **1.01–1.10**, 1.00–1.11	1.06 (4); 1.02–1.10, **1.01–1.11**, 1.00–1.11
zoo	1.26 (14.5); 1.22–1.29, **1.21–1.30**, 1.19–1.30	1.50 (14); 1.45–1.55, **1.44–1.55**, 1.42–1.56	1.51 (14); 1.46–1.55, **1.45–1.56**, 1.44–1.57

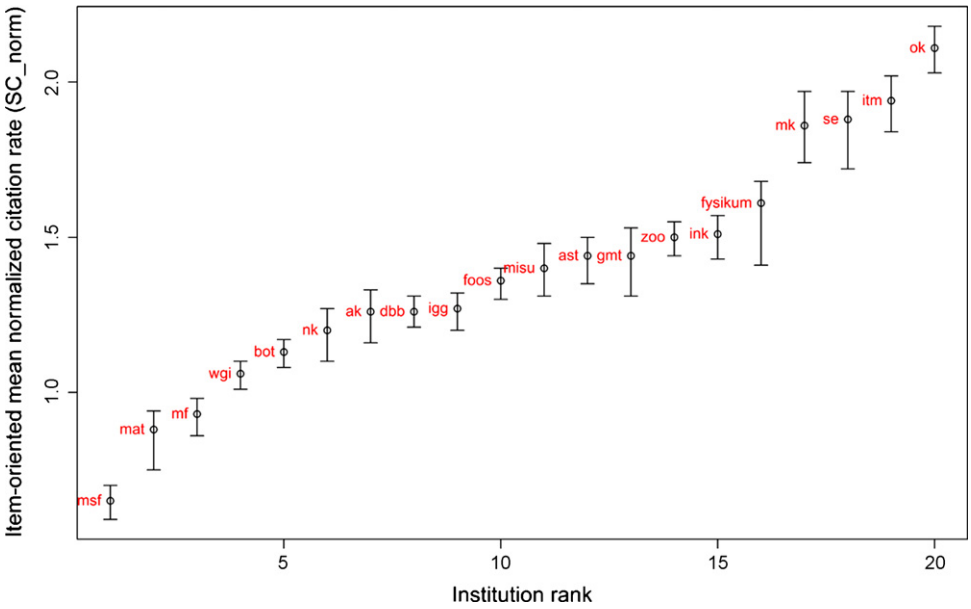

Fig. 3. Department ranking and 90%-stability bars under SC_norm.

Table 1 reports, for every department and under each normalization baseline, the attained value on item-oriented mean normalized citation rate with corresponding rank and with corresponding lower and upper bound (lb and ub, respectively).

Moving over to the Top-5% indicator, we see greater effects of the choice of normalization baseline. Consider Fig. 4, which depicts the association between the values of Top-5% under ESI_norm versus J_norm.

A weak rank order association is observed with 0.49 as the value on Kendall's tau-b. Further, several substantial differences in the two rankings can be observed by inspecting Figs. 5 and 6. As was the case when the rankings were based on item-oriented mean normalized citation rate, nk has a substantially higher rank under J_norm than under ESI_norm, and misu achieves a substantially higher rank under ESI_norm compared to J_norm. Department ak also has a substantially higher rank under J_norm, whereas mk is attributed a substantially lower rank under J_norm.

Further, comparing the rank order of departments based on Top-5% under the other combinations of normalization baselines, we observe the following effects: values on Kendall's tau-b increases (slightly) to 0.51 when comparing rankings under SC_norm and J_norm and increases (noticeably) to 0.88 when association between rankings under SC_norm and ESI_norm are

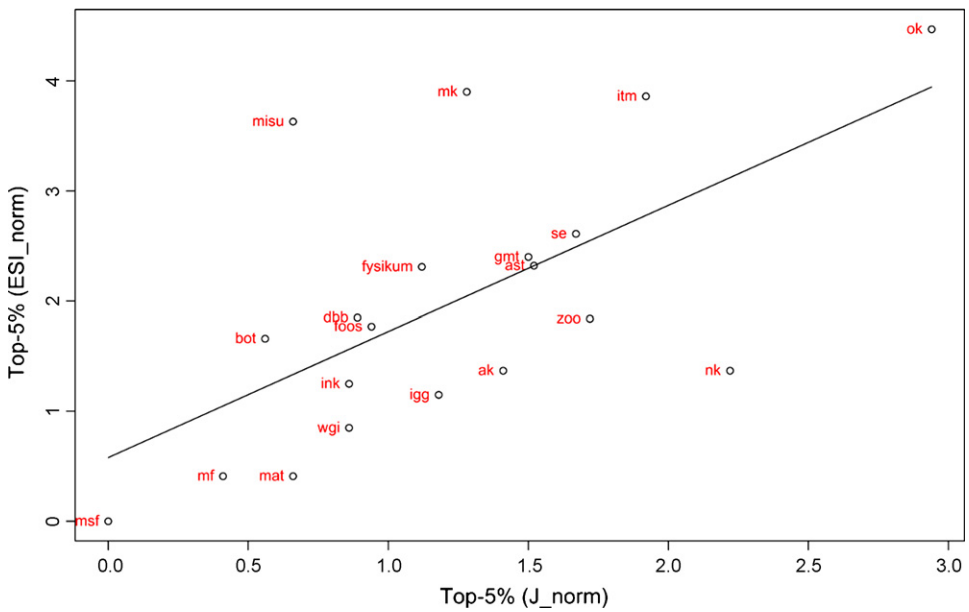

Fig. 4. Association between Top-5% under ESI_norm and J_norm.

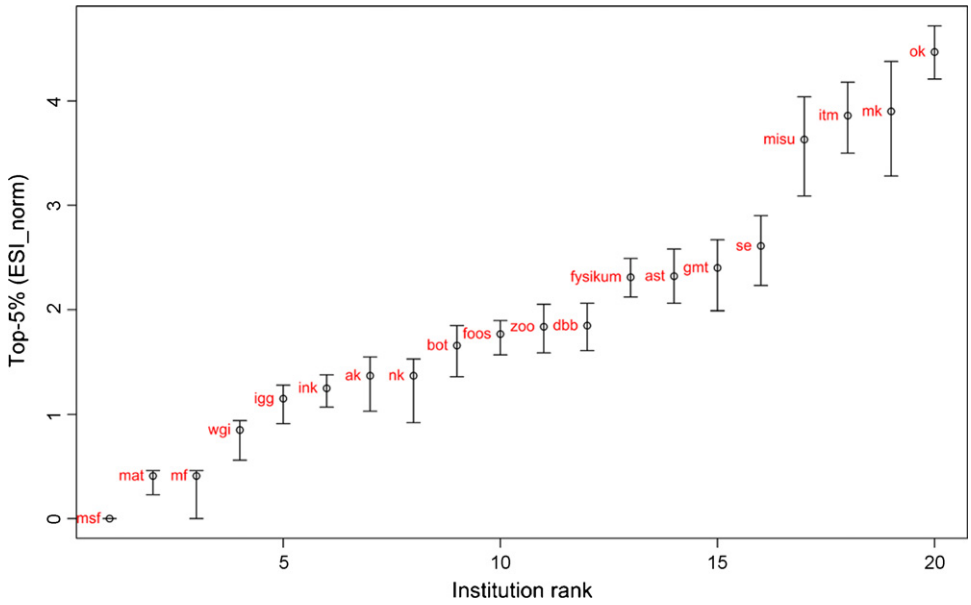

Fig. 5. Department ranking and 90%-stability bars under ESI_norm.

considered. Substantial differences between SC_norm and J_norm rankings are the same as between ESI_norm and J_norm (ESI_norm can be substituted by SC_norm in the result description in the preceding paragraph) except for ak, which no longer has a substantial change in rank. As were the case with item-oriented mean normalized citation rate, no substantial differences in rank were identified when contrasting SC_norm to ESI_norm.

Table 2 reports, for every department and under each normalization baseline, the attained value on Top-5% with corresponding rank and with corresponding lower and upper bound (lb and ub, respectively).

Finally, there is a clear tendency that an increasing number of documents used for normalization, i.e., moving from J_norm over SC_norm to ESI_norm, is associated with higher indicator values. This is especially evident when contrasting J_norm with ESI_norm and with SC_norm. For example, 17 out of the 20 departments perform better with respect to item-oriented mean normalized citation rate when ESI_norm or SC_norm are used as baselines rather than publishing journal.

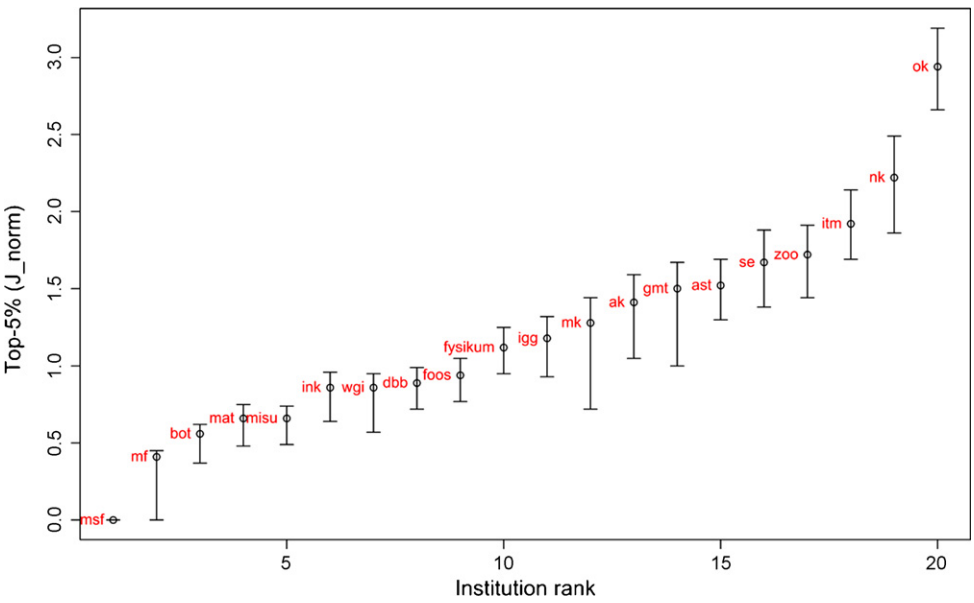

Fig. 6. Department ranking and 90%-stability bars under J_norm.

Table 2
Top-5% for the 20 SU departments under each normalization baseline.

Department	J_norm (rank); 85%, **90%**, 95% lb-ub	SC_norm (rank); 85%, **90%**, 95% lb-ub	ESI_norm (rank); 85%, **90%**, 95% lb-ub
ak	1.41 (13); 1.05–1.59, **1.05–1.59**, 0.53–1.60	1.86 (11); 1.31–2.09, **1.31–2.09**, 1.05–2.09	1.37 (7.5); 1.03–1.55, **1.03–1.55**, 0.51–1.55
ast	1.52 (15); 1.30–1.69, **1.30–1.69**, 1.17–1.69	2.19 (14); 1.93–2.44, **1.92–2.44**, 1.80–2.44	2.32 (14); 2.07–2.58, **2.06–2.58**, 1.94–2.58
bot	0.56 (3); 0.37–0.62, **0.37–0.62**, 0.37–0.62	1.35 (9); 1.14–1.50, **1.14–1.50**, 1.08–1.50	1.66 (9); 1.48–1.85, **1.36–1.85**, 1.35–1.85
dbb	0.89 (8); 0.72–0.99, **0.72–0.99**, 0.72–0.99	1.32 (7); 1.15–1.47, **1.14–1.47**, 1.11–1.47	1.85 (12); 1.70–1.97, **1.61–2.06**, 1.61–2.06
foos	0.94 (9); 0.83–1.04, **0.77–1.05**, 0.77–1.05	1.86 (11); 1.71–2.01, **1.67–2.01**, 1.64–2.04	1.77 (10); 1.64–1.90, **1.57–1.90**, 1.57–1.97
fysikum	1.12 (10); 1.02–1.25, **0.95–1.25**, 0.95–1.25	1.86 (11); 1.69–1.99, **1.67–2.02**, 1.62–2.06	2.31 (13); 2.12–2.49, **2.12–2.49**, 2.05–2.49
gmt	1.50 (14); 1.00–1.67, **1.00–1.67**, 1.00–1.67	2.22 (15); 1.81–2.47, **1.81–2.47**, 1.70–2.47	2.40 (15); 2.00–2.67, **1.99–2.67**, 1.67–2.67
igg	1.18 (11); 0.93–1.32, **0.93–1.32**, 0.93–1.32	1.29 (6); 1.07–1.44, **1.07–1.44**, 0.98–1.44	1.15 (5); 0.91–1.28, **0.91–1.28**, 0.91–1.28
ink	0.86 (6.5); 0.64–0.96, **0.64–0.96**, 0.64–0.96	1.33 (8); 1.15–1.48, **1.09–1.48**, 1.02–1.48	1.25 (6); 1.07–1.38, **1.07–1.38**, 0.92–1.38
itm	1.92 (18); 1.69–2.14, **1.69–2.14**, 1.58–2.14	3.48 (18); 3.21–3.74, **3.19–3.76**, 3.13–3.81	3.86 (18); 3.61–4.10, **3.50–4.18**, 3.49–4.18
mat	0.66 (4.5); 0.48–0.74, **0.48–0.75**, 0.25–0.75	0.42 (2); 0.23–0.47, **0.23–0.47**, 0.23–0.47	0.41 (2.5); 0.23–0.46, **0.23–0.46**, 0.23–0.46
mf	0.41 (2); 0.00–0.45, **0.00–0.45**, 0.00–0.45	0.81 (3); 0.45–0.90, **0.45–0.90**, 0.45–0.90	0.41 (2.5); 0.00–0.46, **0.00–0.46**, 0.00–0.46
misu	0.66 (4.5); 0.49–0.74, **0.49–0.74**, 0.25–0.74	2.31 (16); 1.98–2.58, **1.86–2.58**, 1.86–2.58	3.63 (17); 3.32–4.04, **3.09–4.04**, 3.09–4.04
mk	1.28 (12); 1.07–1.44, **0.72–1.44**, 0.72–1.44	3.87 (19); 3.42–4.34, **3.24–4.34**, 3.23–4.35	3.90 (19); 3.28–4.37, **3.28–4.38**, 3.27–4.38
msf	0.00 (1); 0.00–0.00, **0.00–0.00**, 0.00–0.00	0.00 (1); 0.00–0.00, **0.00–0.00**, 0.00–0.00	0.00 (1); 0.00–0.00, **0.00–0.00**, 0.00–0.00
nk	2.22 (19); 1.86–2.49, **1.86–2.49**, 1.55–2.50	1.09 (5); 0.91–1.22, **0.71–1.22**, 0.61–1.22	1.37 (7.5); 0.92–1.53, **0.92–1.53**, 0.92–1.53
ok	2.94 (20); 2.74–3.18, **2.66–3.19**, 2.65–3.19	4.09 (20); 3.86–4.33, **3.82–4.37**, 3.77–4.38	4.47 (20); 4.21–4.72, **4.21–4.72**, 4.13–4.80
se	1.67 (16); 1.39–1.87, **1.38–1.88**, 1.17–1.89	2.76 (17); 2.41–3.07, **2.40–3.08**, 2.27–3.08	2.61 (16); 2.23–2.90, **2.23–2.90**, 2.01–2.91
wgi	0.86 (6.5); 0.57–0.95, **0.57–0.95**, 0.57–0.96	0.87 (4); 0.68–0.96, **0.59–0.96**, 0.59–0.96	0.85 (4); 0.57–0.94, **0.56–0.94**, 0.56–0.94
zoo	1.72 (17); 1.55–1.91, **1.44–1.91**, 1.43–1.92	2.01 (13); 1.81–2.19, **1.78–2.23**, 1.72–2.23	1.84 (11); 1.60–2.05, **1.59–2.05**, 1.59–2.05

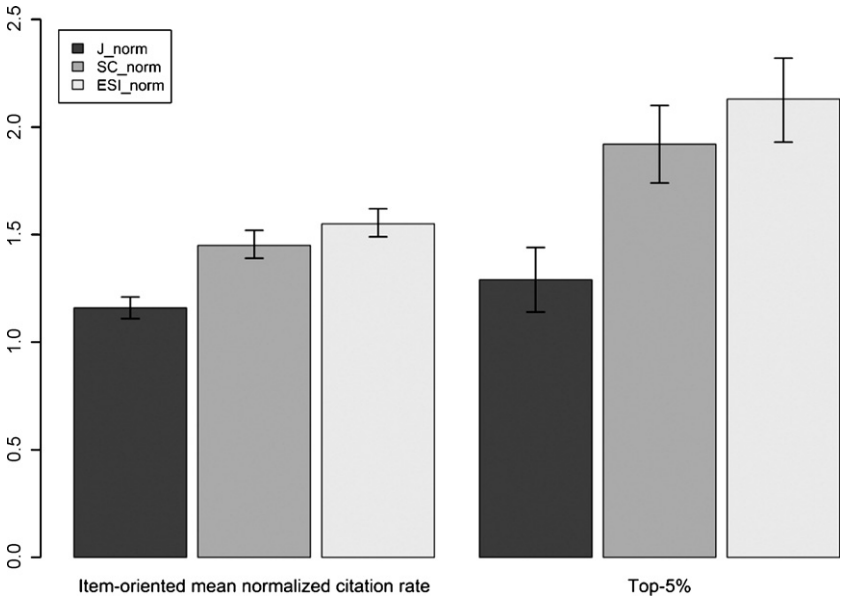

Fig. 7. Absolute performance under different normalization baselines with 90%-stability bars, based on the documents constituting the complete set of SU documents. (From left to right, lower bound-upper bound: 1.12–1.22, 1.41–1.53, 1.50–1.64 and 1.15–1.45, 1.84–2.19, 2.00–2.38. The corresponding 85%-stability bars: 1.13–1.22, 1.42–1.53, 1.51–1.63 and 1.16–1.43, 1.86–2.17, 2.03–2.36, whereas the 95%-stability bars equals: 1.11–1.23, 1.40–1.54, 1.49–1.65 and 1.12–1.48, 1.81–2.22, 1.96–2.42.)

The effects of normalization baseline on the absolute performance are summarized in Fig. 7. The indicators are calculated on the complete set of (unique) SU documents, i.e., treating the aggregated departments as the unit of analysis.[3] The performance is clearly better when normalization is carried out with ESI_norm or SC_norm rather than with publishing journal, i.e., higher indicator values and non-overlapping stability bars with respect to J_norm. While the performance is slightly better under ESI_norm compared to SC_norm, the differences are not such as they can be considered as substantial.

4. Discussion and conclusions

We have studied the effects of field normalization baseline on citation impact of 20 SU natural science departments. Three baselines were used: J_norm (journal), SC_norm (ISI/Thomson Reuters subject category), and ESI_norm (Essential Science Indicators field). Citation impact was measured by the indicators item-oriented mean normalized citation rate and Top-5%.

The results show that the choice of normalization baseline matters. For item-oriented mean normalized citation rate, we observed a fairly weak association, measured by Kendall's tau-b, between the rankings of the departments under SC_norm (ESI_norm) and J_norm, 0.76 and 0.67, respectively, and a few differences in rank were substantial. The association between the rankings under SC_norm and ESI_norm was higher (0.82), though, and it is worth to underlining that no substantial differences were observed when contrasting ESI_norm with SC_norm, whether we look at the effect on rank or the effect on absolute values. Greater effects of the choice of normalization baseline were observed when the indicator Top-5% was applied (thus, with regard to our data, there is an interaction between normalization baseline and indicator). The associations between the rankings under SC_norm (ESI_norm) and J_norm were as weak as 0.51 and 0.49, respectively, and several differences in rank were substantial. However, as was the case with item-oriented mean normalized citation rate, a relatively high association (0.88) was observed, and no substantial differences (with respect to rank and absolute values) were identified, when contrasting SC_norm with ESI_norm.

One might expect that using only 22 macro-fields compared to the considerably larger numbers of subject categories would have a greater effect. However, somewhat similar results are briefly indicated by Glänzel et al. (2009), where baseline values based on a custom classification of journals into 60 (non-overlapping) broad fields are used to create normalized impact values. These values are shown to be highly correlated with impact values based on the non-mutually exclusive classification of journals according to the more numerous subject categories of ISI/Thomson Reuters.

[3] Due to the aggregation, the indicator values are calculated on the basis of a much larger number of documents (compared to when each department is treated separately). Hence, it is feasible to reduce the size of the subsamples. For this aggregated data set we used 10,000 randomly selected half-samples, i.e., the size of each sample equals 50% of the aggregated set.

We note that normalization against J_norm is particular. The rankings of the departments obtained when J_norm is used, irrespective of indicator, differ considerably from the rankings obtained when SC_norm or ESI_norm are used. The observation that J_norm is particular is in line with findings reported by Zitt et al. (2005). Regarding absolute performance, where indicators are calculated on the complete set of (unique) SU documents, the performance is clearly better when normalization is carried out using ESI_norm or SC_norm compared to normalization using J_norm. With regard to absolute performance, SC_norm and J_norm, Adams et al. (2008) made a similar observation as we did: the citation impact was considerably better when received citations were normalized against to SC_norm compared to normalization against J_norm.

One might object, with regard to the stability analysis, that the approach to use of 90%-stability intervals is too conservative. With the use of a lower percentage, i.e., to use a higher lower bound and a lower upper bound, substantial differences in rank might be identified when SC_norm is compared with ESI_norm. We admit that our choice of percentage is somewhat arbitrary, but we note that the use of a lower percentage yields that more sampling data are discarded. Moreover, a research evaluation entity may adjust the involved parameters in accordance with its needs.

In the light of the typically right-skewed nature of the underlying citation distribution, the use of cardinal measures, like item-oriented mean normalized citation rate, could be questioned. Here subsample stability analysis has a clear merit in that it reveals the effect a few documents might have on the indicator value, and ward off over-interpretation by adding an interval to statements such as "unit *A* is cited *x*% above expectation", an interval that indicates how stable the observed indicator value is. This seems especially relevant for the applied and recently suggested indicator which normalizes on item level and thus is especially sensitive to outliers (van Raan et al., 2010). Further, rankings are increasingly popular and perhaps the part of an evaluation that most often obtain broad attention. Stability intervals are one good way of highlighting the instability inherent in many rankings. Observed differences in rank might turn out to be non-substantial in the sense of overlapping intervals.

Is one of the three baselines to be preferred to the other two? A drawback with J_norm is that a collection of documents, with low citation frequencies and published in journals with low citations volumes, might have a similar value on item-oriented mean citation rate (or on Top-5%) as a collection of highly cited documents, published in journals with high citation volumes. On the other hand, normalization against ESI_norm might be regarded as a case of under-normalization: target documents are compared to documents that substantially deviate from the target documents with respect to subject. Under-normalization also occurs when SC_norm is applied, but to a lesser extent. However, there are no unambiguous reasons why one of the three baselines should be preferred over the other two, it largely depends on which point of view the evaluation entity wants to illuminate. However, a reasonable choice, in our view, is to construct normalized impact indicators based on several baselines to get a more comprehensive portrait of the citation impact of a research unit. By calculating normalized impact based on J_norm and contrasting the result with indicators normalized on a higher aggregated level, a more informative picture can emerge, compared to the case where only one baseline is used. The use of multiple baselines is also valuable for indicating how robust, with respect to the choice of baseline, the rank of a particular unit is.

J_norm baseline values are more or less readily available from Web of Science, and ESI_norm values are calculated and made available in ESI.[4] Based on the observations in this paper that no substantial differences were observed contrasting ESI_norm with SC_norm, one might suggest that people without access to SC_norm data (which for many people still are laborious and hard to obtain) can perform reasonable normalized citation impact studies by combining J_norm and ESI_norm.

Finally, one might question the rationale of using pre-defined journals sets of the type used in this paper as the basis for normalization. In particular, papers published in multidisciplinary journals constitute a problem (Glänzel, Schubert, Schoepflin, & Czerwon, 1999). Additionally one might question if baseline values calculated on the basis of journal sets, such as the ones associated with SC_norm or ESI_norm, or even individual journals are reasonable comparison values for publications in very specialized sub-fields (Bornmann, 2010). When possible, one should probably consider creating baselines based on classification of papers on the individual level (Neuhaus & Daniel, 2009; Strotmann & Zhao, 2010) or use a different approach altogether and carry out the normalization process on the citing side (i.e., in essence based on the length of the citing articles reference lists), similar to what was recently proposed by Moed (2010) for measuring the normalized citation impact of journals. For future research, it would be interesting to compare traditional field normalization with the alternatives sketched above.

Appendix A.

In Table 3, we give the English names of the SU departments included in the study, together with corresponding abbreviations and number of publications.[5]

[4] Albeit they are calculated somewhat different than in this paper.

[5] Recently, foos and mk have been merged to Department of Materials and Environmental Chemistry, and igg has changed name to Department of Geological Sciences.

Table 3
The English names of the departments with corresponding abbreviations.

Abbreviation	English name	Number of publications
ak	Department of Analytical Chemistry	45
ast	Department of Astronomy	178
bot	Department of Botany	197
dbb	Department of Biochemistry and Biophysics	258
foos	Department of Physical, Inorganic and Structural Chemistry	347
fysikum	Department of Physics	315
gmt	Department of Genetics, Microbiology and Toxicology	70
igg	Department of Geology and Geochemistry	129
ink	Department of Physical Geography and Quaternary Geology	154
itm	Department of Applied Environmental Science	210
mat	Department of Mathematics	104
mf	Department of Molecular Biology and Functional Genomics	51
misu	Department of Meteorology	99
mk	Department of Environmental Chemistry	65
msf	Medical Radiation Physics	33
nk	Department of Neurochemistry	76
ok	Department of Organic Chemistry	268
se	Department of Systems Ecology	109
wgi	Wenner-Gren Institute	123
zoo	Department of Zoology, including Population Genetics	210

References

Adams, J., Gurney, K., & Jackson, L. (2008). Calibrating the zoom—A test of Zitt's hypothesis. *Scientometrics*, *75*(1), 81–95.

Ball, R., Mittermaier, B., & Tunger, D. (2009). Creation of journal-based publication profiles of scientific institutions—A methodology for the interdisciplinary comparison of scientific research based on the J-factor. *Scientometrics*, *81*(2), 381–392.

Bornmann, L. (2010). Towards an ideal method of measuring research performance: Some comments to the Opthof and Leydesdorff (2010) paper. *Journal of Informetrics*, doi:10.1016/j.joi.2010.04.004

Glänzel, W., Schubert, A., Schoepflin, U., & Czerwon, H. J. (1999). An item-by-item subject classification of papers published in journals covered by the SSCI database using reference analysis. *Scientometrics*, *46*(3), 431–441.

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics*, *78*(1), 165–188.

Kostoff, R. N. (2002). Citation analysis of research performer quality. *Scientometrics*, *53*(1), 49–71.

Lundberg, J. (2007). Lifting the crown-citation z-score. *Journal of Informetrics*, *1*(2), 145–154.

Lunneborg, C. E. (2000). *Data analysis by resampling: Concepts and applications*. Pacific Grove, CA: Duxbury Press.

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, doi:10.1016/j.joi.2010.01.002

Moed, H. F., De Bruin, R. E., & van Leeuwen, T. N. (1995). New bibliometric tools for the assessment ofnational research performance: Database description, overview of indicators and first applications. *Scientometrics*, *33*(3), 381–422.

Neuhaus, C., & Daniel, H. D. (2009). A new reference standard for citation analysis in chemistry and related fields based on the sections of Chemical Abstracts. *Scientometrics*, *78*(2), 219–229.

Opthof, T., & Leydesdorff, L. (2010). Caveats for the journal and field normalizations in the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, doi:10.1016/j.joi.2010.02.003

Schubert, A., & Braun, T. (1993). Reference-standards for citation based assessments. *Scientometrics*, *26*(1), 21–35.

Schubert, A., & Braun, T. (1996). Cross-field normalization of scientometric indicators. *Scientometrics*, *36*(3), 311–324.

Steele, C., Butler, L., & Kingsley, D. (2006). The publishing imperative: the pervasive influence of publication metrics. *Learned Publishing*, *19*(4), 277–290.

Strotmann, A., & Zhao, D. (2010). Combining commercial citation indexes and open-access bibliographic databases to delimit highly interdisciplinary research fields for citation analysis. *Journal of Informetrics*, *4*(2), 194–200.

Tijssen, R. J. W., Visser, M. S., & van Leeuwen, T. N. (2002). Benchmarking international scientific excellence: Are highly cited research papers an appropriate frame of reference? *Scientometrics*, *54*(3), 381–397.

van Raan, A. F. J., van Leeuwen, T. N., Visser, M. S., van Eck, N. J., & Waltman, L. (2010). Rivals for the crown: Reply to Opthof and Leydesdorff. *Journal of Informetrics*, doi:10.1016/j.joi.2010.03.008

Visser, M. S., & Nederhof, A. J. (2007). Bibliometric study of the Uppsala University, Sweden, 2002–2006. In J. Nordgren (Ed.), *Quality and renewal 2007: An overall evaluation of research at Uppsala University 2006/2007*. Uppsala: Uppsala University.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics*, *63*(2), 373–401.