RESEARCH ARTICLE

Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers

Lutz Bornmann[1], Sitaram Devarakonda[2], Alexander Tekles[1,3], and George Chacko[2]

[1]Division for Science and Innovation Studies, Administrative Headquarters of the Max Planck Society, Hofgartenstr. 8, 80539 Munich, Germany

[2]NET ESolutions Corporation, 8180 Greensboro Dr, McLean, VA 22102, USA

[3]Ludwig-Maximilians-Universität Munich, Department of Sociology, Konradstr. 6, 80801 Munich, Germany

Keywords: bibliometrics, dependence index, disruption index, novelty

ABSTRACT

Recently, Wu, Wang, and Evans (2019) proposed a new family of indicators, which measure whether a scientific publication is disruptive to a field or tradition of research. Such disruptive influences are characterized by citations to a focal paper, but not its cited references. In this study, we are interested in the question of convergent validity. We used external criteria of newness to examine convergent validity: In the postpublication peer review system of F1000Prime, experts assess papers whether the reported research fulfills these criteria (e.g., reports new findings). This study is based on 120,179 papers from F1000Prime published between 2000 and 2016. In the first part of the study we discuss the indicators. Based on the insights from the discussion, we propose alternate variants of disruption indicators. In the second part, we investigate the convergent validity of the indicators and the (possibly) improved variants. Although the results of a factor analysis show that the different variants measure similar dimensions, the results of regression analyses reveal that one variant (DI_5) performs slightly better than the others.

1. INTRODUCTION

Citation analyses often focus on counting the number of citations to a focal paper (FP). To assess the academic impact of the FP, its citation count is compared with the citation count for a similar paper (SP) that has been published in the same research field and year. If the FP receives significantly more citations than the SP, its impact is noteworthy: The FP seems to be more useful or interesting for other researchers than the SP. However, the simple counting and comparing of citations does not reveal what the reasons for the impact of publications might be. As the overviews by Bornmann and Daniel (2008) and Tahamtan and Bornmann (2019) show, various reasons exist why publications are (highly) cited. Especially for research evaluation purposes, it is very interesting to know whether certain publications have impact because they report novel or revolutionary results. These are the results from which science and society mostly profit.

In this paper, we focus on a new type of indicator family measuring the impact of publications by examining not only the number of citations received but also the references cited in publications. Recently, Funk and Owen-Smith (2017) proposed a new family of indicators that

an open access journal

Check for updates

Citation: Bornmann, L., Devarakonda, S., Tekles, A., & Chacko, G. (2020). Are disruption index indicators convergently valid? The comparison of several indicator variants with assessments by peers. *Quantitative Science Studies*, *1*(3), 1242–1259. https://doi.org/10.1162/qss_a_00068

DOI: https://doi.org/10.1162/qss_a_00068

Received: 20 November 2019
Accepted: 01 May 2020

Corresponding Author:
Lutz Bornmann
bornmann@gv.mpg.de

Handling Editor:
Vincent Larivière

The MIT Press

measure the disruptive potential of patents. Wu et al. (2019) transferred the conceptual idea to publication data by measuring whether an FP disrupts a field or tradition of research (see also a similar proposal in Bu, Waltman, & Huang, 2019). Azoulay (2019) describes the so-called disruption index proposed by Wu et al. (2019) as follows: "When the papers that cite a given article also reference a substantial proportion of that article's references, then the article can be seen as consolidating its scientific domain. When the converse is true—that is, when future citations to the article do not also acknowledge the article's own intellectual forebears—the article can be seen as disrupting its domain" (p. 331).

We are interested in the question of whether disruption indicators are convergently valid with assessments by peers. The current study has two parts: In the first part, we discuss the indicators introduced by Wu et al. (2019) and Bu et al. (2019) and identify possible limitations. Based on the insights from the discussion, we propose alternate variants of disruption indicators. In the second part, we investigate the convergent validity of the indicators proposed by Wu et al. (2019) and Bu et al. (2019) and the (possibly) improved variants. We used an external criterion of newness, which is available at the paper level for a large paper set: tags (e.g., "new finding") assigned to papers by peers expressing newness.

Convergent validity asks "to what extent does a bibliometric exercise exhibit externally convergent and discriminant qualities? In other words, does the indicator satisfy the condition that it is positively associated with the construct that it is supposed to be measuring? The criteria for convergent validity would not be satisfied in a bibliometric experiment that found little or no correlation between, say, peer review grades and citation measures" (Rowlands, 2018). The analyses are intended to identify the indicator (variant) that is more strongly related to assessments by peers (concerning newness) than other indicators.

2. INDICATORS MEASURING DISRUPTION

The new family of indicators measuring disruption has been developed based on the previous introduction of another indicator family measuring novelty. Research on the novelty indicator family is based on the view of research as a "problem solving process involving various combinatorial aspects so that novelty comes from making unusual combinations of preexisting components" (Wang, Lee, & Walsh, 2018, p. 1074). Uzzi, Mukherjee, et al. (2013) analyzed cited references, and investigated whether referenced journal pairs in papers are atypical or not. Papers with many atypical journal pairs were denoted as papers with high novelty potential. The authors argue that highly cited papers are not only highly novel but also very conventionally oriented. In a related study, Boyack and Klavans (2014) reported strong disciplinary and journal effects in inferring novelty.

In recent years, Lee, Walsh, and Wang (2015) proposed an adapted version of the novelty measure proposed by Uzzi et al. (2013), Wang, Veugelers, and Stephan (2017), and Stephan, Veugelers, and Wang (2017) introduced a novelty measure focusing on publications with great potential of being novel by identifying new journal pairs (instead of atypical pairs). A different approach is used by Boudreau, Guinan, et al. (2016) and Carayol, Lahatte, and Llopis (2017), who used unusual combinations of keywords for measuring novelty. Other studies in the area of measuring novelty have been published by Foster, Rzhetsky, and Evans (2015), Mairesse and Pezzoni (2018), Bradley, Devarakonda, et al. (2020), and Wagner, Whetsell, and Mukherjee (2019), each with a different focus. According to the conclusion by Wang et al. (2018), "prior work suggests that coding for rare combinations of prior knowledge in the publication produces a useful a priori measure of the novelty of a scientific publication" (p. 1074).

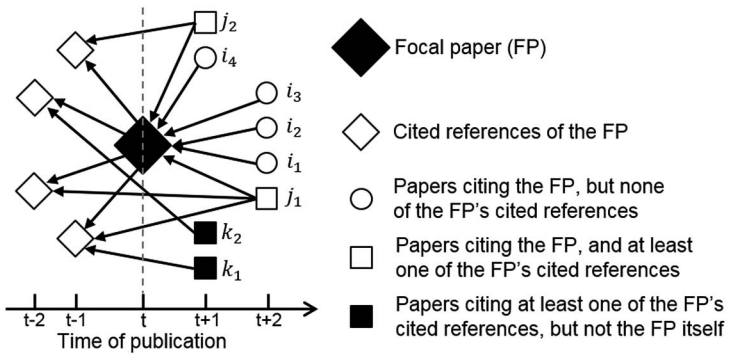
Novelty indicators have been developed against the backdrop of the desire to identify and measure creativity. How is creativity defined? According to Hemlin, Allwood, et al. (2013) "creativity is held to involve the production of high-quality, original, and elegant solutions to complex, novel, ill-defined, or poorly structured problems" (p. 10). Puccio, Mance, and Zacko-Smith (2013) claim that "many of today's creativity scholars now define creativity as the ability to produce original ideas that serve some value or solve some problem" (p. 291). The connection between the indicators measuring novelty and creativity of research is made by that stream of research viewing creativity "as an evolutionary search process across a combinatorial space and sees creativity as the novel recombination of elements" (Lee et al., 2015, p. 685). For Estes and Ward (2002) "creative ideas are often the result of attempting to determine how two otherwise separate concepts may be understood together" (p. 149), whereby the concepts may refer to different research traditions or disciplines. Similar statements on the roots of creativity can be found in the literature from other authors, as the overview by Wagner et al. (2019) shows. Bibliometric novelty indicators try to capture the combinatorial dynamic of papers (and thus, the creative potential of papers; see Tahamtan & Bornmann, 2018) by investigating lists of cited references or keywords for new or unexpected combinations (Wagner et al., 2019).

In a recent study, Bornmann, Tekles, et al. (2019) investigated two novelty indicators and tested whether they exhibit convergent validity. They used a similar design to this study and found that only one indicator is convergently valid.

In this context of measuring creativity, not only the development of indicators measuring novelty but also the introduction of indicators identifying disruptive research have occurred. These indicators are interested in exceptional research that turns knowledge formation in a field around. The family of disruption indicators proposed especially by Wu et al. (2019) and Bu et al. (2019) seizes on the concept of Kuhn (1962), who differentiated between phases of normal science and scientific revolutions. Normal science is characterized by paradigmatic thinking, which is rooted in traditions and consensus orientation; scientific revolutions follow divergent thinking and openness (Foster et al., 2015). Whereas normal science means linear accumulation of research results in a field (Petrovich, 2018), scientific revolutions are dramatic changes with an overthrow of established thinking (Casadevall & Fang, 2016). Preconditions for scientific revolutions are creative knowledge claims that disrupt linear accumulation processes in field-specific research (Kuukkanen, 2007).

Bu et al. (2019) see the development of disruption indicators in the context of a multidimensional perspective on citation impact. In contrast to simple citation counting under the umbrella of a one-dimensional perspective, the multidimensional perspective considers breadth and depth through the cited references of an FP and the cited references of its citing papers (see also Marx & Bornmann, 2016). In contrast to the family of novelty indicators, which are based exclusively on cited references, disruption indicators combine the cited references of citing papers with the cited references data of FPs. The disruptiveness of an FP is measured based on the extent to which the cited references of the papers citing the FP also refer to the cited references of the FP. According to this idea, many citing papers not referring to the FP's cited references indicate disruptiveness. In this case, the FP is the basis for new work that does not depend on the context of the FP (i.e., the FP gives rise to new research).

Disruptiveness was first described by Wu et al. (2019) and Funk and Owen-Smith (2017) and presented as a weighted index DI_1 (see Figure 1) calculated for an FP by dividing the difference between the number of publications that cite the FP without citing any of its cited references (N_i) and the number of publications that cite both the FP and at least one of its cited



$$DI_l = \frac{N_i - N_j^l}{N_i + N_j^l + N_k} \qquad DI_l^{no\,k} = \frac{N_i - N_j^l}{N_i + N_j^l} \qquad DeIn = \frac{N_{j\times cited}}{N_i + N_j^1}$$

$$N_i = |\{i_1, i_2, \ldots\}| \quad N_j^l = |\{j_m | j_m \text{ cites FP and at least } l \text{ of FP's cited references}\}|$$

$$N_k = |\{k_1, k_2, \ldots\}| \quad N_{j\times cited} = |\{(j_m, p_n) | p_n \text{ is cited by } j_m \text{ and FP}\}|$$

Figure 1. Different roles of papers in citation networks for calculating disruption indicators and formulae for different disruption indicators (see Wu & Wu, 2019).

references (N_j^1) by the sum of N_i, N_j^1, and N_k (the number of publications that cite at least one of the FP's cited reference without citing the FP itself). Simply put, this is the ratio $\frac{N_i - N_j^1}{N_i + N_j^1 + N_k}$. High positive values of this indicator should be able to point to disruptive research; high negative values should reflect developmental research (i.e., new research that continues previous research lines).

DI_1 corresponds to a certain notion of disruptiveness, according to which only few papers are disruptive (while most papers are not disruptive), and a paper needs to have a large citation impact to score high on DI_1. DI_1 detects only a few papers as disruptive due to the term N_k, which is often very large compared to the other terms in the formula (Wu & Yan, 2019). A large N_k produces disruption values of small magnitude, as N_k only occurs in the denominator of the formula. As a result, the disruption index is very similar for many papers, and only a few papers get high disruption values. However, Funk and Owen-Smith (2017), who originally defined the formula for the disruption index, designed the indicator to measure disruptiveness on a continuous scale from –1 to 1: "the measure is continuous, able to capture degrees of consolidation and destabilization" (p. 793). This raises the question of whether different nuances of disruptions can be adequately captured by DI_1, or if the term N_k is too dominant for this purpose.

Including N_k in the formula can also be questioned with regard to its function for assessing the disruptiveness of a paper. The basic idea that all disruption indicators share is to distinguish between citing papers of an FP indicating disruptiveness, and citing papers indicating consolidation. N_k does not refer to this distinction. Instead, it captures the citation impact of an FP compared to other papers in a similar context (all papers citing the same papers as the FP). Assuming a notion of disruptiveness that aims at detecting papers that have a large-scale disruptive effect, this idea seems reasonable. However, this form of considering citation impact can be problematic, as it strongly depends on the citation behavior of the FP, and small changes in the FP's cited references can have a large effect on the disruption score.

A more general issue regarding the function of N_k is whether citation impact should be considered at all when measuring disruptiveness, which depends on the underlying notion of disruptiveness. Bu et al. (2019) suggest separating disruptiveness (depth of citation impact in their terminology) and citation impact in terms of number of forward citations. This perspective assumes that disruptiveness is a quality of papers that can also be observed on a small impact level. From this perspective, N_k should not be included in the formula for measuring disruptiveness, especially as it is the dominant factor in most cases. Consequently, an alternative indicator would be simply to drop the term N_k, which corresponds to DI_1^{nok} according to the formula in Figure 1. This variant of the disruption indicator has been proposed by Wu and Yan (2019). With $\frac{N_i}{N_i+N_j}$, a very similar approach for calculating a paper's disruption has been proposed by Bu et al. (2019). This indicator can be defined as a function of DI_1^{nok}, such that differences between papers just change by the factor 0.5, so that both variants allow identical conclusions. In our analyses, we will consider DI_1^{nok} because it has the same range of output values as the original disruption index DI_1.

In contrast to the aforementioned variants of indicators measuring disruptiveness, Bu et al. (2019) also proposed an indicator that considers how many cited references of the FP are cited by an FP's citing paper. This approach takes into account how strongly the FP's citing papers rely on the cited references of the FP, instead of just considering if this relationship exists (in the form of at least one citation of a cited reference of the FP). The corresponding indicator proposed by Bu et al. (2019) (denoted as *DeIn* in Figure 1) is defined as the average number of cited references of the FP that its citing papers cite. In contrast to the other indicators mentioned earlier, *DeIn* is supposed to decrease with the disruptiveness of a paper, because it measures the dependency of the paper on earlier work (as opposed to disruptiveness). Another difference to the other indicators is that the range of *DeIn* has no upper bound, because the average number of citation links from a paper citing the FP to the FP's cited references is not limited. This makes it more difficult to compare the results of *DeIn* and the other indicators.

By considering only those citing papers of the FP that cite at least l ($l > 1$) of the FP's cited references, it becomes possible to follow the idea of taking into account how strongly the FP's citing papers rely on the cited references of the FP, but also get values that are more comparable to the other indicators. The probability that a citing paper of the FP cites a highly cited reference of the FP is higher than it is for a less frequently cited reference of the FP. Therefore, the fact that a paper cites both the FP and at least one of its cited references is not equally indicative for a developmental FP in all cases. Only considering those of the FP's citing papers that also cite at least a certain number of the FP's cited references mitigates the problem, because the focus on the most reliable cases of citing papers indicates a developmental FP.

This is formalized in the formulae in Figure 1, where the subscripts of DI_l and DI_l^{nok} correspond to the threshold for the number of cited references of the FP which a citing paper must cite to be considered. With a threshold of $l = 1$ (i.e., without any restriction on the number of the FP's cited references that the FP's citing papers must cite), the indicator is identical to the indicator originally proposed by Wu et al. (2019). To analyze how well these different strategies are able to measure the disruptiveness of a paper, we compare the following indicators in our analyses: DI_1, DI_5, DI_1^{nok}, DI_5^{nok}, *DeIn*. The subscript in four variants indicates the minimum number of cited references that are cited along with the FP. The superscript "no k" in two variants indicates that N_k is excluded from the calculation.

3. METHODS

3.1. F1000Prime

F1000Prime is a database including important papers from biological and medical research (see https://f1000.com/prime/home). The database is based on a postpublication peer review system: Peer-nominated faculty members (FMs) select the best papers in their specialties and assess these papers for inclusion in the F1000Prime database. FMs write brief reviews explaining the importance of papers and rate them as "good" (1 star), "very good" (2 stars) or "exceptional" (3 stars). Many papers in the database are assessed by more than one FM. To rank the papers in the F1000Prime database, the individual scores are summed up to a total score for each paper.

FMs also assign the following tags to the papers, if appropriate:[1]

- Confirmation: article validates published data or hypotheses
- *Controversial*: article challenges established dogma
- Good for teaching: key article in a field and/or is well written
- **Hypothesis**: article presents an interesting hypothesis
- Negative/null results: article has null or negative findings
- **New finding**: article presents original data, models or hypotheses
- **Novel drug target**: article suggests new targets for drug discovery
- Refutation: article disproves published data or hypotheses
- **Technical advance**: article introduces a new practical/theoretical technique, or novel use of an existing technique

The tags in bold reflect aspects of novelty in research. As disruptive research should include elements of novelty, we expect that the tags are positively related to the disruption indicator scores. For instance, we assume that a paper receiving many "new finding" tags from FMs will have a higher disruption index score than a paper receiving only a few tags (or none at all). The tags not printed in bold are not related to newness (e.g., confirmation of published hypotheses), so that the expectations for these tags are zero or negative correlations with disruption index scores. In terms of measures that are likely to be inversely correlated with disruptiveness, the one that seems most plausible is the "confirmation" tag. The tag "controversial" is printed in italics. It is not clear whether the tag is able to reflect novelty or not. FMs further assign the tags "clinical trial," "systematic review/meta-analysis," and "review/commentary" to papers that are not relevant for this study (and not used thus).

We interpret the empirical results in Section 4.3 against the backdrop of the above assumptions. In the interpretations of the results, however, it should be considered that the allocations of tags by the FMs are subjective decisions associated with (more or less) different meanings. In other words, the tags data are affected by noise (uncertainties) covering the signal (clear-cut judgments). Another point which should be considered in the interpretation of the empirical results is the fact that the above assumptions can be formulated in another way. For example, we anticipate that papers that are "good for teaching" would be inversely correlated with disruptiveness. The opposite could be true as well. Papers that introduce new topics, perspectives, and ways of thinking—that shift the conversation—would be most useful for teaching. Many factors play a role in the interpretation of the "good for teaching" tag: How complex is the paper assessed by the FMs? Is it a landmark paper published decades ago or a recent research front paper? Is the paper intended for teaching of bachelor, masters or doctoral students?

[1] The definitions of the tags are adopted from https://f1000.com/prime/about/whatis/how

Many other studies have already used data from the F1000Prime database for correlating them with metrics. Most of these studies are interested in the relationship between quantitative (metrics-based) and qualitative (human-based) assessments of research. The analysis of Anon (2005) shows that "papers from high-profile journals tended to be rated more highly by the faculty; there was a tight correlation (R^2 = .93) between average score and the 2003 impact factor of the journal" (see also Jennings, 2006). Bornmann and Leydesdorff (2013) correlated several bibliometric indicators and F1000Prime recommendations. They found that the "percentile in subject area achieves the highest correlation with F1000 ratings" (p. 286). Waltman and Costas (2014) report "a clear correlation between F1000 recommendations and citations. However, the correlation is relatively weak" (p. 433). Similar results were published by Mohammadi and Thelwall (2013). Bornmann (2015) investigated the convergent validity of F1000Prime assessments. He found that "the proportion of highly cited papers among those selected by the FMs is significantly higher than expected. In addition, better recommendation scores are also associated with higher performing papers" (p. 2415). The most recent study by Du, Tang, and Wu (2016) shows that "(a) nonprimary research or evidence-based research are more highly cited but not highly recommended, while (b) translational research or transformative research are more highly recommended but have fewer citations" (p. 3008).

3.2. Data Set Used and Variables

The study is based on a data set from F1000Prime including 207,542 assessments of papers. These assessments refer to 157,020 papers (excluding papers with duplicate DOIs, missing DOIs, missing expert assessments, etc.). The bibliometric data for these papers are from an in-house database (Korobskiy, Davey, et al., 2019), which utilizes Scopus data (Elsevier Inc.). To increase the validity of the indicators included in this study, we considered only papers with at least 10 cited references and at least 10 citations. Furthermore, we included only papers from 2000 to 2016 to have reliable data (some publications are from 1970 or earlier) and a citation window for the papers of at least 3 years (since publication until the end of 2018). The reduced paper set consists of 120,179 papers published between 2000 and 2016 (see Table 1).

We included several variables in the empirical part of this study: the disruption index proposed by Wu et al. (2019) (DI_1) and the dependence indicator proposed by Bu et al. (2019) (*DeIn*). The alternative disruption indicators described in Section 2 considered were: DI_5, DI_1^{nok}, and DI_5^{nok}. For the comparison with the indicators reflecting disruption, we included the sum (ReSc.sum) and the average (ReSc.avg) of reviewer scores (i.e., scores from FMs). Besides the qualitative assessments of research, quantitative citation impact scores are also considered: number of citations until the end of 2018 (Citations) and percentile impact scores (Percentiles).

As publication and citation cultures are different in the fields, it is standard in bibliometrics to field- and time-normalize citation counts (Hicks, Wouters, et al., 2015). Percentiles are field- and time-normalized citation impact scores (Bornmann, Leydesdorff, & Mutz, 2013) that are between 0 and 100 (higher scores reflect more citation impact). For the calculation of percentiles, the papers published in a certain subject category and publication year are ranked in decreasing order. Then the formula $(i - 0.5)/n \times 100$ (Hazen, 1914) is used to calculate percentiles (i is the rank of a paper and n the number of papers in the subject category). Impact percentiles of papers published in different fields can be directly compared (despite possibly differing publication and citation cultures).

Table 1. Number and percentage of papers included in the study

Publication year	Number of papers	Percentage of papers
2000	196	0.16
2001	1,530	1.27
2002	3,229	2.69
2003	3,717	3.09
2004	5,185	4.31
2005	6,711	5.58
2006	8,765	7.29
2007	8,824	7.34
2008	10,046	8.36
2009	10,368	8.63
2010	11,074	9.21
2011	10,934	9.1
2012	10,536	8.77
2013	9,903	8.24
2014	7,261	6.04
2015	6,121	5.09
2016	5,779	4.81
Total	120,179	100.00

Table 2 shows the key figures for citation impact scores, reviewer scores, and variants measuring disruption. As the percentiles reveal, the paper set includes especially papers with a considerable citation impact. Table 3 lists the papers that received the maximum scores in Table 2. The maximum DI_1 with the value 0.677 has been reached by the paper entitled "Cancer statistics, 2010" published by Jemal, Siegel, et al. (2010). This publication is one of an annual series published on incidence, mortality, and survival rates for cancer and its high score may be an artifact of the DI_1 formula because it is likely that the report is cited much more than its cited references. In fact, this publication may make the case for the DI_5 formulation. Similarly, the 2013 edition of the cancer statistics report was found to have the maximum percentile. The maximum number of citations was seen for the well-known review article by Hanahan and Weinberg (2011), "Hallmarks of cancer: the next generation."

3.3. Statistics Applied

The statistical analyses in this study have three steps:

1. We investigated the correlations between citation impact scores, reviewer scores, and the scores of the indicators measuring disruption. All variables are not normally distributed and affected by outliers. To tackle this problem, we logarithmized the scores by

Table 2. Key figures of the included variables (n = 120,179)

Variable	Mean	Median	Standard deviation	Minimum	Maximum
DI_1	−0.007	−0.004	0.013	−0.322	0.677
DI_5	0.089	−0.007	0.278	−0.800	1.000
DI_1^{nok}	−0.521	−0.579	0.294	−0.998	0.975
DI_5^{nok}	−0.008	−0.053	0.545	−0.990	1.000
DeIn	3.327	2.970	1.871	0.013	43.059
ReSc.sum	2.028	2.000	1.808	1.000	55.000
ReSc.avg	1.486	1.000	0.586	1.000	3.000
Citations	149.848	73.000	298.467	10.000	20446.000
Percentiles	87.246	91.947	13.248	23.659	100.000

using the formula log(x + 1). This logarithmic transformation approximates the distributions to normal distributions[2]. As perfectly normally distributed variables cannot be achieved with the transformation, Spearman rank correlations have been calculated (instead of Pearson correlations). We interpret the correlation coefficients against the backdrop of the guidelines proposed by Cohen (1988) and Kraemer, Morgan, et al. (2003): small effect = 0.1, medium effect = 0.3, large effect = 0.5, and very large effect = 0.7.

2. We performed an exploratory factor analysis (FA) to analyze the variables. FA is a statistical method for data reduction (Gaskin & Happell, 2014); it is an exploratory technique to identify latent dimensions in the data and to investigate how the variables are related to the dimensions (Baldwin, 2019). We expected three dimensions, because we have variables with citation impact scores, reviewer scores, and indicators' scores measuring disruption. As the (logarithmized) variables do not perfectly follow the normal distribution, we performed the FA using the robust covariance matrix following Verardi and McCathie (2012). Thus, the results of the FA are based on not the variables but on a covariance matrix. The robust covariance matrix has been transformed into a correlation matrix (StataCorp, 2017), which has been analyzed by the principal component factor method (the communalities are assumed to be 1). We interpreted the factor loadings for the orthogonal varimax rotation; the factor loadings have been adjusted "by dividing each of them by the communality of the correspondence variable. This adjustment is known as the Kaiser normalization" (Afifi, May, & Clark, 2012, p. 392). In the interpretation of the results, we focused on factor loadings with values greater than 0.5.
3. We investigated the relationship between the dimensions (identified in the FA) and F1000Prime tags (as proxies for newness or not). We expected a close relationship between the dimension reflecting disruption and tags reflecting newness. The tags are count variables including the sum of the tags assignments from F1000Prime FMs for

[2] We additionally performed the statistical analyses with scores that are not logarithmized and received very similar results.

Table 3. Example papers with maximum scores

Variable	Paper (authors, publication year, and title)
DI_1	Jemal et al. (2010). Cancer statistics, 2010
DI_5	Mohan and Shellard (2014). Providing family planning services to remote communities in areas of high biodiversity through a Population-Health-Environment programme in Madagascar
DI_1^{nok}	Kourtis, Nikoletopoulou, and Tavernarakis (2012). Small heat-shock proteins protect from heat-stroke-associated neurodegeneration
DI_5^{nok}	Frank (2009). The common patterns of nature
DeIn	Kincaid, Murata, et al. (2016). Specialized proteasome subunits have an essential role in the thymic selection of $CD8^+$ T cells
ReSc.sum	Lolle, Victor, et al. (2005). Genome-wide non-Mendelian inheritance of extra-genomic information in *Arabidopsis*
ReSc.avg	McEniery, Yasmin, et al. (2005). Normal vascular aging: Differential effects on wave reflection and aortic pulse wave velocity: The Anglo-cardiff Collaborative Trial (ACCT)
Citations	Hanahan and Weinberg (2011). Hallmarks of cancer: The next generation
Percentiles	Siegel, Naishadham, and Jemal (2013). Cancer statistics, 2013

single papers. To calculate the relationship between dimensions and tags, we performed a robust Poisson regression (Hilbe, 2014; Long & Freese, 2014). The Poisson model is recommended to be used in cases of count data as dependent variable. Robust methods are recommended when the distributional assumptions for the model are not completely met (Hilbe, 2014). Because we are interested in identifying indicators for measuring disruption that might perform better than the other variants, we tested the correlation between each variant and the tag assignments using several robust Poisson regressions. Citations, disruptiveness, and tag assignments are dependent on time (Bornmann & Tekles, 2019). Thus, we included the number of years between 2018 and the publication year as exposure time in the models (Long & Freese, 2014, pp. 504–506).

4. RESULTS

4.1. Correlations Between Citation Impact Scores, Reviewer Scores, and Variants Measuring Disruption

Figure 2 shows the matrix including the coefficients of the correlations between reviewer scores, citation impact indicators, and variants measuring disruption. DI_1 is correlated at a medium level with the other indicators measuring disruption, whereby these indicators correlate among themselves at a very high level. Very high positive correlations are visible between citations and percentiles and between the average and sum of reviewer scores.

The correlation between DI_1 and citation impact (citations and percentiles) is at least at the medium level, but it is negative ($r = -.46$, $r = -.37$). Thus, the original DI_1 seems to measure another dimension than citation impact. This result is in agreement with results reported by Wu et al. (2019, Figure 2a). However, the situation changed with the other indicators measuring disruption to small positive (negative in the case of *DeIn*) correlation coefficients.

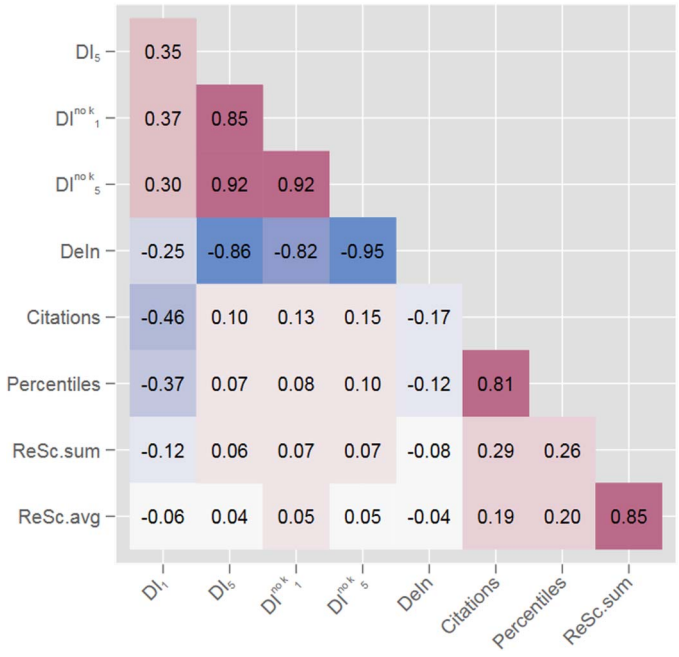


Figure 2. Spearman rank correlations based on logarithmized variables [log($y + 1$)]. The following abbreviations are used: different indicators measuring disruption (DI_1, DI_5, DI_1^{nok}, DI_5^{nok}, *DeIn*), the sum (ReSc.sum), and the average (ReSc.avg) of reviewer scores.

4.2. Factor Analysis to Identify Latent Dimensions

We calculated an FA including reviewer scores, citation impact indicators, and variants measuring disruption to investigate the underlying dimensions (latent variables). Most of the results shown in Table 4 agree with expectations: We found three dimensions, which we labeled as *disruption* (factor 1), *citations* (factor 2), and *reviewers* (factor 3). However, contrary to what was expected, DI_1 loads negatively on the citation dimension revealing that (a) high DI_1 scores are related to low citation impact scores (see above) and (b) all other indicators measuring disruption are independent of DI_1. Thus, the other indicators (at least one) seem to be promising developments compared to the originally proposed indicator DI_1.

4.3. Relationship Between Tag Mentions and FA Dimensions

Using Poisson regression models including the tags, we calculated correlations between the tags and the three FA dimensions (disruption, citations, and reviewers). We are especially interested in the correlation between the tags (measuring newness of research or not) and the disruption dimension from the FA. We also included the citation impact and reviewer dimensions into the analyses to see the corresponding results for comparison. In the analyses, we considered the FA scores for the three dimensions (predicted values from FA that are not correlated by definition) as independent variables and the various tags (the sum of the tags assignments) as dependent variables in nine Poisson regressions (one regression model for each tag).

The results of the regression analyses are shown in Table 5. We do not focus on the statistical significance of the results, because they are more or less meaningless against the backdrop of the high case numbers. The most important information in the table is the signs of the

Table 4. Rotated factor loadings from a factor analysis using logarithmized variables [log(y + 1)]

Variable	Factor 1	Factor 2	Factor 3	Uniqueness
DI_1	0.24	**−0.69**	0.05	0.46
DI_5	**0.90**	−0.07	0.00	0.19
DI_1^{nok}	**0.90**	−0.10	0.02	0.17
DI_5^{nok}	**0.97**	−0.03	0.01	0.05
DeIn	**−0.91**	−0.01	0.01	0.17
Citations	0.05	**0.91**	0.04	0.16
Percentiles	0.04	**0.84**	0.12	0.29
ReSc.sum	0.00	0.05	**1.00**	0.00
ReSc.avg	0.00	0.05	**1.00**	0.00

Notes: Three eigenvalues > 1. The following abbreviations are used: different indicators measuring disruption (DI_1, DI_5, DI_1^{nok}, DI_5^{nok}, and *DeIn*), the sum (ReSc.sum), and the average (ReSc.avg) of reviewer scores.

coefficients and the percentage change coefficients. The percentage change coefficients are counterparts to odds ratios in regression models, which measure the percentage changes in the dependent variable if the independent variable (FA score) increases by one standard deviation (Deschacht & Engels, 2014; Hilbe, 2014; Long & Freese, 2014). The percentage change coefficient for the model based on the "technical advance" tag and the disruption dimension can be interpreted as follows: For a standard deviation increase in the scores for disruption, a paper's expected number of new finding tags increases by 10.93%, holding other variables in the regression analysis constant. This increase is as expected and substantial. However, the results of the other tags expressing newness have a negative sign and are against expectations.

The percentage change coefficients for the citation dimension are significantly higher than for the disruption dimension (especially for the new finding tag) and positive. This result is against our expectations, because the disruption variants should measure newness in a better way than citations. However, one should consider in the interpretation of the results that DI_1 correlates negatively with the citation indicators. Thus, the dimension also measures disruptiveness (as originally proposed), whatever the case may be. If we interpret the results for the dimension against this backdrop, at least the results for the tags not representing newness seem to accord with our expectations. The results for the reviewer dimension are similar to the citations dimension results. The consistent positive coefficients for the citations and reviewers dimensions in Table 5 might result from the fact that the tags are from the same FMs as the recommendations, and the FMs probably use citations to find relevant papers for reading, assessing, and including in the F1000Prime database.

Table 6 reports the results from some additional regression analyses. Because we are interested in not only correlations between dimensions (reflecting disruptiveness) and tags (the sum of the tags assignments) but also in correlations between the various variants measuring disruption and tags, we calculated 45 additional regression models. We are interested in the question of which variant measuring disruption reflects newness better than other variants: Are the different variants differently or similarly related to newness, as expressed by the tags? Table 6 only shows percentage change coefficients (see above) from the regression models

Table 5. Results of nine Poisson regression analyses (n = 120,179 papers). The models have been adjusted for exposure time (different publication years): How long was the time that the papers have been at risk of being tagged and cited (number of years between publication and counting of citations or tags, respectively)?

	Disruption		Citations		Reviewers		
Tag	Coefficient	Percentage change	Coefficient	Percentage change	Coefficient	Percentage change	Constant
Tags expressing newness (expecting positive signs)							
Hypothesis	−.06*** (−10.22)	−6.11	.20*** (36.91)	28.62	.32*** (33.24)	41.28	−3.96*** (−36.32)
New finding	−.09*** (−41.77)	−9.92	.23*** (111.06)	34.02	.27*** (76.41)	33.48	−3.35*** (−83.59)
Novel drug target	−.01 (−1.45)	−1.66	.27*** (27.33)	39.91	.34*** (22.96)	43.93	−5.87*** (−34.41)
Technical advance	.09*** (13.67)	**10.93**	.22*** (32.56)	32.22	.25*** (24.34)	30.74	−4.58*** (−39.50)
Tags not expressing newness (expecting negative signs)							
Confirmation	−.03*** (−6.41)	**−3.85**	.14*** (28.93)	19.69	.13*** (21.86)	14.30	−4.56*** (−68.84)
Good for teaching	−.06*** (−6.73)	**−7.08**	.14*** (14.71)	19.00	.38*** (19.75)	49.61	−4.06*** (−21.12)
Negative/Null results	−.14** (−3.25)	**−14.72**	.17*** (5.09)	23.23	.34*** (10.02)	44.25	−7.86*** (−19.94)
Refutation	−.19*** (−8.14)	**−19.00**	.23*** (12.05)	32.88	.28*** (12.64)	34.47	−7.71*** (−28.50)
Tag without expectations							
Controversial	−.04*** (−4.54)	−4.78	.20*** (24.85)	28.11	.27*** (27.92)	33.54	−5.30*** (−47.46)

Notes: t statistics in parentheses; * $p < .05$, ** $p < .01$, *** $p < .001$. Percentage change coefficients in bold are as expected, otherwise the results are not as expected (or, for the tag "Controversial," there are no clear expectations).

Table 6. Results (percentage change coefficients) of 45 Poisson regressions with tags as dependent variables and different variants measuring disruption as independent variables each. The models have been adjusted for exposure time (different publication years): How long have the papers been at risk of being tagged and cited?

Tag	DI_1	DI_5	DI_1^{nok}	DI_5^{nok}	*DeIn*
Tags expressing newness (expecting positive signs)					
Hypothesis	**4.32**	**3.01**	**4.66**	**2.75**	0.25
New finding	–2.71	–0.62	–2.13	–2.13	1.92
Novel drug target	**6.89**	**6.74**	**14.85**	**15.19**	**–7.91**
Technical advance	**6.72**	**20.65**	**18.34**	**19.80**	**–18.11**
Tags not expressing newness (expecting negative signs)					
Confirmation	**–5.08**	**–0.41**	0.08	1.45	–1.88
Good for teaching	12.24	0.37	8.82	3.73	**6.41**
Negative/Null results	3.45	**–11.46**	**–13.35**	**–5.20**	–2.10
Refutation	**–9.28**	**–9.59**	**–14.83**	**–10.37**	**8.06**
Tag without expectations					
Controversial	–0.33	4.42	1.46	1.46	–3.62

Note: The following abbreviations are used: different indicators measuring disruption (DI_1, DI_5, DI_1^{nok}, DI_5^{nok}, *DeIn*). Percentage change coefficients in bold are as expected, otherwise the results are not as expected (or, for the tag "Controversial," there are no clear expectations).

(because of the great number of models). In other words, percentage changes in expected counts (of the tag) for a standard deviation increase in the variant measuring disruption are listed. For example, a standard deviation change in DI_1 on average increases a paper's expected number of technical advance tags by 6.72%. This result agrees with expectation, because the technical advance tag reflects newness. It seems that DI_5 reflects the assessments by FMs at best; the lowest number of results in agreement is visible for *DeIn*.

5. DISCUSSION

For many years, scientometrics research has focused on improving the way of field-normalizing citation counts or developing improved variants of the *h*-index. However, this research is rooted in a relatively one-dimensional way of measuring impact. With the introduction of the new family of disruption indicators, the one-dimensional method of impact measurement may now give way to multidimensional approaches. Disruption indicators consider not only times-cited information but also cited references data (of FPs and citing papers). High indicator values should be able to point to published research disrupting traditional research lines. Disruptive papers catch the attention of citing authors (at the expense of the attention devoted to previous research); disruptive research enters unknown territory, which is scarcely consistent with known territories handled in previous papers (and cited by disruptive papers). Thus, the citing authors exclusively focus on the disruptive papers (by citing them) and do not reference previous papers cited in the disruptive papers.

Starting from the basic approach of comparing cited references of citing papers with cited references of FPs, different variants of measuring disruptiveness have been proposed recently. An overview of many possible variants can be found in Wu and Yan (2019). In this study, we

included some variants that sounded reasonable and/or followed different approaches. For example, *DeIn* proposed by Bu et al. (2019) is based on the number of citation links from an FP's citing papers to the FP's cited references (without considering N_k). We were interested in the convergent validity of these new indicators following the basic analyses approach by Bornmann et al. (2019). The convergent validity can be tested by using an external criterion measuring a similar dimension. Although we did not have an external criterion at hand measuring disruptiveness specifically, we used tags from the F1000Prime database reflecting newness. FMs assess papers using certain tags and some tags reflect newness. We assumed that disruptive research is assessed as new. Based on the F1000Prime data, we investigated whether the tags assigned by the FMs to the papers correspond with indicator values measuring disruptiveness.

In the first step of the statistical analyses, we calculated an FA for inspecting whether the various indicators measuring disruptiveness load on a single "disruptiveness" dimension. As the results reveal, this was partly the case: All variants of the DI_1—the original disruption index proposed by Wu et al. (2019)—loaded on one dimension—the "disruptiveness" dimension. However, the original disruption index itself loaded on a dimension which reflects citation impact; however, it loaded negatively. These results might be interpreted as follows: The proposed disruption index variants measure the same construct, which might be interpreted as "disruptiveness." DI_1 is related to citation impact whereby negative values—the developmental index manifestation of this indicator (see Section 2)—correspond to high citation impact levels. As all variants of DI_1 loaded on the same factor in the FA, the results do not show which variant should be preferred (if any). Thus, we considered a second step of analyses in this study.

In this step, we tested the correlation between each variant (including the original) and the external "newness" criterion. The results showed that DI_5 reflects the FMs' assessments at best (corresponding with our expectations more frequently than the other indicators); the lowest number of results that demonstrated an agreement between tag and indicator scores is visible for *DeIn*. The difference between the variants is not very large; however, the results can be used to guide the choice of a variant if disruptiveness is measured in a scientometric study. Although the authors of the paper introducing DI_1 (Wu et al., 2019) performed analyses to validate the index (e.g., by calculating the indicator for Nobel-Prize-winning papers), they did not report on evaluating possible variants of the original, which might perform better.

We noted that while a single publication was the most highly disruptive for the DI_1 (0.6774), and DI_1^{nok} (0.9747), 703 and 3816 publications respectively scored the maximum disruptiveness value of 1.0 for variants DI_5 and DI_5^{nok}. We also reviewed examples of the most highly disruptive publications as measured by all four variants and observed that instances of an annual Cancer Statistics report published by the American Cancer Society received maximal disruptiveness scores for all four variants, presumably because this report is highly cited in each year of its publication without its references being cited. A publication from the *Journal of Global Environmental Change* (https://doi.org/10.1016/j.gloenvcha.2008.10.009) was also noteworthy and may reflect the focus on climate change.

All empirical research has limitations. This study is no exception. We assumed in this study that novelty is necessarily a (or the) defining feature of disruptiveness. There are plenty of existing bibliometric measures of novelty (e.g., new keyword or cited references combinations; see Bornmann et al., 2019). Although novelty may be necessary for disruptiveness, it is not necessarily sufficient to make something disruptive. We cannot completely exclude the possibility that many nondisruptive discoveries are novel. "Normal science" (Kuhn, 1962)

discoveries do not-necessarily lack novelty; they make novel contributions (e.g., hypotheses, findings, or technical advances) in a way that builds on and enhances prior work (e.g., within the paradigm). A second limitation of the study might be that the F1000Prime data are affected by possible biases. We know from the extensive literature on (journal and grant) peer review processes that indications for possible biases in the assessments by peers exist (Bornmann, 2011). The third limitation refers to the F1000Prime tags that we used in this study. None of them directly captures disruptiveness (as we acknowledge above). Most of the tags capture novelty, which is related, but conceptually different from disruptiveness, and for which there are already existing metrics (see Section 2). Because disruption indicators propose measuring disruption (and not novelty), we can't directly make claims on whether disruption indicators measure what they propose to measure.

It would be interesting to follow up in future studies that use mixed-methods approaches to evaluate the properties of N_i, N_j, and N_k variants more systematically against additional gold standard data sets. The F1000 data set is certain to feature its own bias (e.g., it is restricted to biomedicine and includes disproportionately many high-impact papers) and the variants we describe may exhibit different properties when evaluated against multiple data sets.

ACKNOWLEDGMENTS

We would like to thank Tom Des Forges and Ros Dignon from F1000 for providing us with the F1000Prime data set. We thank two anonymous reviewers for very helpful suggestions to improve a previous version of the paper.

AUTHOR CONTRIBUTIONS

Lutz Bornmann: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing—original draft. Sitaram Devarakonda: Data curation, Formal analysis, Investigation, Methodology, Writing—review & editing. Alexander Tekles: Conceptualization, Formal analysis, Investigation, Writing—review & editing. George Chacko: Conceptualization, Formal analysis, Investigation, Methodology, Writing—review & editing.

COMPETING INTERESTS

The authors do not have any competing interests. Citation data used in this paper relied on Scopus (Elsevier Inc.) as implemented in the ERNIE project (Korobskiy et al., 2019), which is collaborative between NET ESolutions Corporation and Elsevier Inc. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health, NET ESolutions Corporation, or Elsevier Inc.

FUNDING INFORMATION

Research and development reported in this publication was partially supported by funds from the National Institute on Drug Abuse, National Institutes of Health, US Department of Health and Human Services, under Contract No HHSN271201800040C (N44DA-18-1216).

DATA AVAILABILITY

Access to the Scopus bibliographic data requires a license from Elsevier; we cannot therefore make the data used in this study publicly available.

REFERENCES

Afifi, A., May, S., & Clark, V. A. (2012). *Practical multivariate analysis* (5th ed.). Boca Raton, FL: CRC Press.

Anon. (2005). Revolutionizing peer review? *Nature Neuroscience*, *8*(4), 397–397.

Azoulay, P. (2019). Small research teams "disrupt" science more radically than large ones. *Nature*, *566*, 330–332.

Baldwin, S. (2019). *Psychological statistics and psychometrics using stata*. College Station, TX: Stata Press.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*, 199–245.

Bornmann, L. (2015). Inter-rater reliability and convergent validity of F1000Prime peer review. *Journal of the Association for Information Science and Technology*, *66*(12), 2415–2426.

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

Bornmann, L., & Leydesdorff, L. (2013). The validation of (advanced) bibliometric indicators through peer assessments: A comparative study using data from InCites and F1000. *Journal of Informetrics*, *7*(2), 286–291. https://doi.org/10.1016/j.joi.2012.12.003

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. https://doi.org/10.1016/j.joi.2012.10.001

Bornmann, L., & Tekles, A. (2019). Disruption index depends on length of citation window. *El profesional de la información*, *28*(2), e280207. https://doi.org/10.3145/epi.2019.mar.07

Bornmann, L., Tekles, A., Zhang, H. H., & Ye, F. Y. (2019). Do we measure novelty when we analyze unusual combinations of cited references? A validation study of bibliometric novelty indicators based on F1000Prime data. *Journal of Informetrics*, *13*(4), 100979.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., & Riedl, C. (2016). Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, *62*(10), 2765–2783. https://doi.org/10.1287/mnsc.2015.2285

Boyack, K., & Klavans, R. (2014). Atypical combinations are confounded by disciplinary effects. In E. Noyons (Ed.), *Proceedings of the Science and Technology Indicators Conference 2014 Leiden: "Context Counts: Pathways to Master Big and Little Data"* (pp. 64–70). Leiden: Universiteit Leiden.

Bradley, J., Devarakonda, S., Davey, A., Korobskiy, D., Liu, S., Lakhdar-Hamina, D., … Chacko, G. (2020). Co-citations in context: Disciplinary heterogeneity is relevant. *Quantitative Science Studies*, *1*(1), 264–276. https://doi.org/10.1162/qss_a_00007

Bu, Y., Waltman, L., & Huang, Y. (2019). A multidimensional perspective on the citation impact of scientific publications. Retrieved February 6, 2019, from https://arxiv.org/abs/1901.09663

Carayol, N., Lahatte, A., & Llopis, O. (2017). *Novelty in science Proceedings of the Science, Technology, & Innovation Indicators Conference Open indicators: Innovation, participation and actor-based STI indicators*. Paris, France.

Casadevall, A., & Fang, F. C. (2016). Revolutionary science. *mBio*, *7*(2), e00158. https://doi.org/10.1128/mBio.00158-16

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Deschacht, N., & Engels, T. E. (2014). Limited dependent variable models and probabilistic prediction in informetrics. In Y. Ding, R. Rousseau & D. Wolfram (Eds.), *Measuring scholarly impact* (pp. 193–214). Berlin: Springer.

Du, J., Tang, X., & Wu, Y. (2016). The effects of research level and article type on the differences between citation metrics and F1000 recommendations. *Journal of the Association for Information Science and Technology*, *67*(12), 3008–3021.

Estes, Z., & Ward, T. B. (2002). The emergence of novel attributes in concept modification. *Creativity Research Journal*, *14*(2), 149–156. https://doi.org/10.1207/S15326934crj1402_2

Foster, J. G., Rzhetsky, A., & Evans, J. A. (2015). Tradition and innovation in scientists' research strategies. *American Sociological Review*, *80*(5), 875–908. https://doi.org/10.1177/0003122415601618

Frank, S. A. (2009). The common patterns of nature. *Journal of Evolutionary Biology*, *22*(8), 1563–1585. https://doi.org/10.1111/j.1420-9101.2009.01775.x

Funk, R. J., & Owen-Smith, J. (2017). A dynamic network measure of technological change. *Management Science*, *63*(3), 791–817. https://doi.org/10.1287/mnsc.2015.2366

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*(3), 511–521. https://doi.org/10.1016/j.ijnurstu.2013.10.005

Hanahan, D., & Weinberg, Robert A. (2011). Hallmarks of cancer: The next generation. *Cell*, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013

Hazen, A. (1914). Storage to be provided in impounding reservoirs for municipal water supply. *Transactions of American Society of Civil Engineers*, *77*, 1539–1640.

Hemlin, S., Allwood, C. M., Martin, B., & Mumford, M. D. (2013). Introduction: Why is leadership important for creativity in science, technology, and innovation. In S. Hemlin, C. M. Allwood, B. Martin & M. D. Mumford (Eds.), *Creativity and leadership in science, technology, and innovation* (pp. 1–26). New York, NY: Taylor & Francis.

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden Manifesto for research metrics. *Nature*, *520*(7548), 429–431.

Hilbe, J. M. (2014). *Modelling count data*. New York, NY: Cambridge University Press.

Jemal, A., Siegel, R., Xu, J., & Ward, E. (2010). Cancer statistics, 2010. *CA: A Cancer Journal for Clinicians*, *60*(5), 277–300. https://doi.org/10.3322/caac.20073

Jennings, C. G. (2006). Quality and value: the true purpose of peer review. What you can't measure, you can't manage: The need for quantitative indicators in peer review. Retrieved July 6, 2006, from http://www.nature.com/nature/peerreview/debate/nature05032.html

Kincaid, E. Z., Murata, S., Tanaka, K., & Rock, K. L. (2016). Specialized proteasome subunits have an essential role in the thymic selection of CD8+ T cells. *Nature Immunology*, *17*(8), 938–945. https://doi.org/10.1038/ni.3480

Korobskiy, D., Davey, A., Liu, S., Devarakonda, S., & Chacko, G. (2019). Enhanced Research Network Informatics (ERNIE) (Github Repository). Retrieved November 11, 2019, from https://github.com/NETESOLUTIONS/ERNIE

Kourtis, N., Nikoletopoulou, V., & Tavernarakis, N. (2012). Small heat-shock proteins protect from heat-stroke-associated

neurodegeneration. *Nature*, *490*(7419), 213–218. https://doi.org/10.1038/nature11417
Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*(12), 1524–1529. https://doi.org/10.1097/01.chi.0000091507.46853.d1
Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.
Kuukkanen, J.-M. (2007). Kuhn, the correspondence theory of truth and coherentist epistemology. *Studies In History and Philosophy of Science Part A*, *38*(3), 555–566.
Lee, Y.-N., Walsh, J. P., & Wang, J. (2015). Creativity in scientific teams: Unpacking novelty and impact. *Research Policy*, *44*(3), 684–697. https://doi.org/10.1016/j.respol.2014.10.007
Lolle, S. J., Victor, J. L., Young, J. M., & Pruitt, R. E. (2005). Genome-wide non-Mendelian inheritance of extra-genomic information in Arabidopsis. *Nature*, *434*(7032), 505–509. https://doi.org/10.1038/nature03380
Long, J. S., & Freese, J. (2014). *Regression models for categorical dependent variables using Stata* (3rd ed.). College Station, TX: Stata Press.
Mairesse, J., & Pezzoni, M. (2018). Novelty in science: The impact of French physicists' novel articles. In P. Wouters (Ed.), *Proceedings of the Science and Technology Indicators Conference 2018 Leiden, "Science, technology and innovation indicators in transition"* (pp. 212–220). Leiden: University of Leiden.
Marx, W., & Bornmann, L. (2016). Change of perspective: Bibliometrics from the point of view of cited references-a literature overview on approaches to the evaluation of cited references in bibliometrics. *Scientometrics*, *109*(2), 1397–1415. https://doi.org/10.1007/s11192-016-2111-2
McEniery, C. M., Yasmin, Hall, I. R., Qasem, A., Wilkinson, I. B., & Cockcroft, J. R. (2005). Normal vascular aging: Differential effects on wave reflection and aortic pulse wave velocity. *The Anglo-Cardiff Collaborative Trial (ACCT)*, *46*(9), 1753–1760. https://doi.org/10.1016/j.jacc.2005.07.037
Mohammadi, E., & Thelwall, M. (2013). Assessing non-standard article impact using F1000 labels. *Scientometrics*, *97*(2), 383–395. https://doi.org/10.1007/s11192-013-0993-9
Mohan, V., & Shellard, T. (2014). Providing family planning services to remote communities in areas of high biodiversity through a population-health-environment programme in Madagascar. *Reproductive Health Matters*, *22*(43), 93–103. https://doi.org/10.1016/S0968-8080(14)43766-2
Petrovich, E. (2018). Accumulation of knowledge in para-scientific areas: The case of analytic philosophy. *Scientometrics*, *116*(2), 1123–1151. https://doi.org/10.1007/s11192-018-2796-5
Puccio, G. J., Mance, M., & Zacko-Smith, J. (2013). Creative leadership. Its meaning and value for science, technology, and innovation. In S. Hemlin, C. M. Allwood, B. Martin & M. D. Mumford (Eds.), *Creativity and leadership in science, technology, and innovation* (pp. 287–315). New York, NY: Taylor & Francis.
Rowlands, I. (2018). What are we measuring? Refocusing on some fundamentals in the age of desktop bibliometrics. *FEMS Microbiology Letters*, *365*(8). https://doi.org/10.1093/femsle/fny059
Siegel, R., Naishadham, D., & Jemal, A. (2013). Cancer statistics, 2013. *CA: A Cancer Journal for Clinicians*, *63*(1), 11–30. https://doi.org/10.3322/caac.21166
StataCorp. (2017). *Stata statistical software: release 15*. College Station, TX: Stata Corporation.
Stephan, P., Veugelers, R., & Wang, J. (2017). Blinkered by bibliometrics. *Nature*, *544*(7651), 411–412.
Tahamtan, I., & Bornmann, L. (2018). Creativity in science and the link to cited references: Is the creative potential of papers reflected in their cited references? *Journal of Informetrics*, *12*(3), 906–930. https://doi.org/10.1016/j.joi.2018.07.005
Tahamtan, I., & Bornmann, L. (2019). What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics*, *121*(3), 1635–1684. https://doi.org/10.1007/s11192-019-03243-4
Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472. https://doi.org/10.1126/science.1240474
Verardi, V., & McCathie, A. (2012). The S-estimator of multivariate location and scatter in Stata. *Stata Journal*, *12*(2), 299–307.
Wagner, C. S., Whetsell, T. A., & Mukherjee, S. (2019). International research collaboration: Novelty, conventionality, and atypicality in knowledge recombination. *Research Policy*, *48*(5), 1260–1270. https://doi.org/10.1016/j.respol.2019.01.002
Waltman, L., & Costas, R. (2014). F1000 recommendations as a new data source for research evaluation: A comparison with citations. *Journal of the Association for Information Science and Technology*, *65*(3), 433–445.
Wang, J., Lee, Y.-N., & Walsh, J. P. (2018). Funding model and creativity in science: Competitive versus block funding and status contingency effects. *Research Policy*, *47*(6), 1070–1083. https://doi.org/10.1016/j.respol.2018.03.014
Wang, J., Veugelers, R., & Stephan, P. E. (2017). Bias against novelty in science: A cautionary tale for users of bibliometric indicators. *Research Policy*, *46*(8), 1416–1436.
Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*, 378–382. https://doi.org/10.1038/s41586-019-0941-9
Wu, Q., & Yan, Z. (2019). Solo citations, duet citations, and prelude citations: New measures of the disruption of academic papers. Retrieved May 15, 2019, from https://arxiv.org/abs/1905.03461
Wu, S., & Wu, Q. (2019). A confusing definition of disruption. Retrieved May 15, 2019, from https://osf.io/preprints/socarxiv/d3wpk/