ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-201608.00040

# A New Parameter-Free Indicator for Identifying Sleeping Beauty Papers —Validation Based on Sleeping Beauty Papers in Science and Nature

**Authors:** Du Jian, Wuyi Mountains, Mount Wuyi

**Date:** 2016-08-16T00:00:00+00:00

## Abstract

[Purpose/Significance] Through optimization, we propose a simple, parameter-free indicator that can accurately identify Sleeping Beauty papers. [Method/Process] Drawing upon the basic framework of the Sleeping Beauty Index (B-index), we propose the Bcp index by optimizing the target variable from "annual citation count" to "cumulative percentage of annual citation count," and redefine awakening time, sleep depth, and awakening intensity. The index is validated through identifying Sleeping Beauty papers in Science and Nature. [Results/Conclusion] The Bcp index inherits the B-index' s advantage of being highly sensitive to extremely significant Sleeping Beauty papers. The awakening time calculated under the Bcp index framework better aligns with actual circumstances, whereas the awakening time under the B-index framework tends to be overestimated as occurring later. The Bcp index considers the complete citation curve of a paper during the observation period, circumventing the B-index' s limitation of being unable to reflect the citation curve after annual citations reach their maximum. The Bcp index' s constraining power on citation counts at the time of publication is significantly higher than that of the B-index, better conforming to the characteristic of Sleeping Beauty papers having zero or low citations in their early stage. The Bcp index is more sensitive to Sleeping Beauty papers that are older and have relatively lower total citation counts and annual citation counts, which is beneficial for uncovering the underlying patterns of Sleeping Beauty papers among non-highly-cited literature that are typically overlooked. The Bcp index avoids the shortcomings of subjective definition of artificial parameters and the interference of citation scale differences across different disciplinary fields on the unified definition of Sleeping Beauty papers.

ChinaRxiv 中X

## Full Text

# A New Parameter-Free Index for Identifying Sleeping Beauties in Science: Validation Based on Sleeping Beauty Articles in *Science* and *Nature*

**Jian Du[1],[2], Yishan Wu[3]*** [1]Institute of Medical Information, Chinese Academy of Medical Sciences, Beijing 100005 [2]School of Information Management, Nanjing University, Nanjing 210032 [3]Chinese Academy of Science and Technology for Development, Beijing 100038

### Abstract

[**Purpose/Significance**] This paper proposes a simplified parameter-free index for accurately identifying sleeping beauty papers. [**Method/Process**] Building upon the basic framework of the Beauty Coefficient (B-index), we introduce the Bcp-index by optimizing the measurement from "annual citation counts" to "annual cumulative percentage of citations," thereby redefining awakening time, sleeping depth, and awakening intensity. The index is validated through identification of sleeping beauty papers published in *Science* and *Nature*. [**Results/Conclusions**] The Bcp-index inherits the B-index' s advantage of high sensitivity to extremely prominent sleeping beauties. However, the awakening time calculated under the Bcp framework aligns better with actual circumstances, whereas the B-index tends to produce delayed estimates. The Bcp-index incorporates the complete citation curve throughout the observation period, overcoming the B-index' s limitation of failing to reflect the citation trajectory after annual citations reach their maximum. The Bcp-index demonstrates significantly stronger constraint on initial post-publication citations than the B-index, better conforming to the characteristic of sleeping beauties having zero or low citations in their early years. The Bcp-index is more sensitive to older sleeping beauties with relatively low total citations and lower annual citation peaks, facilitating the discovery of patterns among non-highly-cited sleeping beauties that are often overlooked. The Bcp-index avoids subjective parameter definitions and mitigates interference from disciplinary differences in citation scales on the unified definition of sleeping beauties.

**Keywords:** sleeping beauty; identification methods; parameter-free index; yearly cumulative percentage of citations

## 1. Introduction

In 2004, Dutch scientometrician van Raan proposed the concept of "Sleeping Beauties in Science" —papers that remain uncited or low-cited for a considerable period after publication, as if in slumber, before suddenly receiving high

citations, as if awakened [**?**]. Sleeping beauties quantitatively describe, from a scientometric perspective, the phenomenon of delayed recognition in the sociology of science by dynamically reflecting the temporal characteristics and historical trajectory of citations. Compared to papers with flash-in-the-pan or normal citation patterns, sleeping beauties' distinctive feature of "long-term initial neglect followed by sudden high citation" provides new avenues for analyzing scientific research fronts and revealing patterns of scientific development.

Philosophically, we must distinguish between phenomenon and essence. Delayed recognition and citation sleep are phenomena, while their essence lies in pioneering or transformative research. Pioneering studies often exceed existing cognitive domains, causing the scientific community to be unaware of their existence or potential value, leading to neglect. Transformative research, by overturning existing paradigms, creates psychological distance within the scientific community, causing underestimation of its value and resulting in resistance [**?**, **?**]. Delayed recognition caused by neglect and resistance is always associated with major scientific discoveries, and sleeping beauties in their publication form are not as rare as previously thought [**?**, **?**]. Citation analysis and peer review represent two mainstream methods for research evaluation and two approaches for assessing the potential value of zero or low-cited papers. The former involves retrospective analysis where zero-cited papers become highly cited after a sleep period, allowing post-hoc analysis of their characteristics. The latter involves expert evaluation of a paper's knowledge value while it remains zero or low-cited, assessing its potential to become a sleeping beauty. If early indicators of sleeping beauties can be identified through these dual perspectives to prompt timely scientific attention, it would be significant for encouraging scientists to pursue innovative work and avoid delayed recognition. Practically, it would also help funding agencies and science policymakers discover pioneering or transformative innovations and deploy relevant frontier research and planning in advance. Achieving this goal requires systematically revealing the typical characteristics that distinguish sleeping beauties and their "prince" papers from others, making identification of sleeping beauties from massive literature corpora the foundational first step.

This paper attempts to propose a new and simple identification method by building upon existing methods while avoiding their shortcomings.

## 2. Overview of Sleeping Beauty Identification Methods

Current identification methods can be summarized into three categories. First, **curve fitting** uses mathematical expressions or appropriate curve types to fit the annual citation distribution of individual papers [**?**, **?**]. However, for large samples, manual observation and classification are required, resulting in low efficiency. Second, **manual parameter setting** includes "mean standards" and "quartile standards." Similar to van Raan's definition criteria, most scholars define the "initial publication period" as 3-5 years, "initial low citations" as 1-2 citations per year, and use an average or cumulative number to define the degree

of "sudden high citation" [?, ?, ?, ?]. The quartile standard defines delayed recognition papers as those that only receive 50% of their total citations after 75% of papers in the same field have already received over 50% of theirs [?]. Since such papers constitute a very small proportion, they do not significantly impact individual or team academic performance [?, ?]. Some studies combine mean standards, quartile standards, and boost factors to better identify exponentially growing sleeping beauties [?], or use indicators incorporating paper age and citation curve shape parameters [?], but such combined indicators lack simplicity and transparency. Overall, manual parameter definitions are subjective and strict, without considering disciplinary differences. The number of identified sleeping beauties largely depends on the set thresholds—the higher the threshold, the fewer sleeping beauties identified.

Third, **parameter-free objective indices** recognize that citation counts represent a cumulative process from zero within the observation window. "Citation Speed" examines average annual citations to reflect the accumulation rate of total citations [?], with smaller values indicating slower overall accumulation and higher later-year citations relative to earlier years, enabling preliminary screening of sleeping beauties [?]. The "Beauty Coefficient" (B-index) calculates the difference between the citation curve and a reference line determined by publication year citations and maximum annual citations [?]. Longer sleep duration, deeper sleep, and greater post-awakening citations yield higher B-values. This method identifies far more sleeping beauties than threshold-based approaches, suggesting they are not isolated incidents [?]. Empirical studies show that while Citation Speed identifies papers with long citation lifespans and sustained high-frequency citations (e.g., linear growth curves), it cannot precisely screen sleeping beauties [?]. The B-index enables rapid identification but fails to reflect citation patterns after the annual citation peak, though it remains the best current parameter-free indicator. In summary, existing methods each have advantages and disadvantages, necessitating continued research on rapid, precise, parameter-free quantitative methods.

## 3. The B-Index and Its Improvements

**3.1 The B-Index** Ke et al. (2015) proposed the Beauty Coefficient (B-index), determined by several parameters: $c_t$ represents citations in year $t$ after publication, where $t$ denotes paper age. Citations in the publication year are $c_0$, the year when annual citations reach maximum is $t_m$, with citations $c_{t_m}$. A reference line (denoted as $\ell$) is drawn from point $(0, c_0)$ to $(t_m, c_{t_m})$. The B-value is obtained by comparing the citation curve with this reference line. The slope of the reference line is $(c_{t_m} - c_0)/t_m$. For any $t < t_m$, the ratio of $\ell \cdot t + c_0$ to $\max\{1, c_t\}$ is calculated. Summing these ratios from $t = 0$ to $t = t_m$ yields the B-value.

According to this definition, B equals 0 when citations peak in the publication year or when the annual citation curve is linear. B is non-positive when the citation curve is a concave function of paper age. The B-index has several

characteristics: (1) It can be calculated for any non-zero cited paper, overcoming subjective definitions based on sleep duration and awakening intensity, enabling study of sleeping beauties as a universal phenomenon rather than extreme cases; (2) Longer sleep duration and greater awakening intensity produce larger B-values; (3) B only considers the citation history from publication to peak annual citations, not the complete trajectory; (4) The denominator uses $\max\{1, c_t\}$, substituting 1 when annual citations are zero and the actual value otherwise. Therefore, with equivalent total citations, B-values are larger when more citations accumulate in later periods.

**3.2 The SBc-Index** Peruzzo (2015) argued that B is sensitive to extremely prominent sleeping beauties with high post-awakening citations but less discriminative for papers with lower total citations (e.g., <50). He modified the vertical axis from annual citations to cumulative citations, proposing the SBc-index [**?**]. The reference line $\ell$ changes to connect $(0, c_0)$ and $(t_m, c_{t_m})$, where $c_{t_m}$ represents total citations at first reaching maximum (i.e., total citations in the observation period). The slope remains $(c_{t_m} - c_0)/t_m$. For any $t < t_m$, the difference between $\ell$ and cumulative citations is calculated and summed from $t = 0$ to $t = t_m$, yielding SBc—the difference between green and red areas in Figure 2 [Figure 2: see original paper]. Peruzzo (2015) did not empirically validate SBc. We hypothesize that since the reference line connects the origin to total citations, SBc depends heavily on total citation counts—higher totals yield larger SBc values. We will test this hypothesis empirically.

**3.3 Improving the B-Index and SBc-Index** The B-index has several drawbacks: (1) It only considers citations from publication to peak annual citations, not the complete curve; (2) According to Li & Ye (2016), the denominator $\max\{1, c_t\}$ fails to effectively constrain initial citations [**?**], and Peruzzo suggested it merely ensures non-zero denominators [**?**]. Using annual citations as denominator loses original information, especially when $c_t > 1$—for example, ratios with numerator and denominator both equal to 2 versus both equal to 50 yield identical values, though the differences from the reference line $\ell$ differ significantly (2 vs. 50); (3) Both B-index and SBc-index depend heavily on total citations, particularly SBc.

We therefore propose improvements. To avoid dependence on citation scale, we modify the vertical axis from "annual cumulative citation counts" to "annual cumulative percentage of citations" based on the SBc framework, defining this as the Bcp-index. The Bcp-index enables comparison of delayed recognition degrees across papers with different citation counts.

For any non-zero cited paper, the annual cumulative percentage curve is monotonically increasing with all vertical coordinates >0, eliminating the need for the B-index denominator. The reference line slope remains $(c_{t_m} - c_0)/t_m$. For any $t < t_m$, the difference between $\ell$ and the cumulative percentage is calculated and summed from $t = 0$ to $t = t_m$, yielding the Bcp-index (Figure 3 [Figure 3:

see original paper]).

Adapting B-index' s awakening time definition, we draw perpendiculars from points on the cumulative percentage curve to $\ell$, obtaining distances $d(t)$. The time $\Delta t$ when $d(t)$ is maximized defines the awakening time, denoted as $t_{aw}$ (Figure 3 [Figure 3: see original paper]).

Under the Bcp framework, a paper is awakened in its 32nd year, rising from 4 citations in year 32 to 12 in year 33 (Figure 3 [Figure 3: see original paper]). Under the B-index framework, awakening occurs in year 36, rising from 3 citations in year 36 to 13 in year 37 (Figure 4 [Figure 4: see original paper]). However, the annual citation curve suggests year 32 is more appropriate. Under the B-index, awakening years often correspond to later points with lowest citations because these points have maximum distance from the reference line, demonstrating that B-index awakening times tend to be delayed estimates.

Li & Ye (2016) proposed four principles for identifying sleeping beauties: (1) early citations should be constrained, (2) the complete citation curve should be considered, (3) awakening time should be fixed and not change over time (a B-index limitation), and (4) subjective parameter definitions should be avoided [**?**]. Their example (Figure 5 [Figure 5: see original paper]) shows Papers P1 and P2 with identical publication years and total citations. While P1 shows higher delayed recognition visually, its B-value is lower than P2' s ($164.75 < 177.95$). Under our Bcp framework, P1' s Bcp-value is significantly higher than P2' s ($5.075 > 1.075$) (Figure 6 [Figure 6: see original paper]).

These theoretical analyses and 典型案例 suggest Bcp-index superiority. We now proceed to empirical validation and systematic comparison of B-index, SBc-index, and Bcp-index.

## 4. Empirical Validation

Using research articles published in *Science* and *Nature* since 1970 as the dataset, with citation counts observed on March 16, 2016, and data compiled through December 2015. To ensure at least a 10-year citation window, we included 78,403 papers published from 1970-2005. Since sleeping beauties are first and foremost highly cited papers, we selected the 20,000 papers with 200 total citations by 2015 (averaging at least 20 citations per year). We calculated B-index, SBc-index, and Bcp-index values, along with awakening times under the Bcp framework, comparing their fundamental characteristics.

### 4.1 SBc-Index Shows Highest Correlation with Total Citations

Parameter-free indices lack strict thresholds; we adopt the top 1% as the criterion, selecting 200 papers for each index from the 20,000-paper pool. Correlation tests with total citations reveal that SBc-index is significantly positively correlated with total citations (coefficient = 0.6). B-index and Bcp-index show no significant positive correlation with total citations (Spearman test). However, Bcp-index' s correlation coefficient is lower than B-index' s,

though not statistically significant (Table 1 ). The top 10 SBc-index papers show high consistency with total citation rankings (Table 2 ), confirming our hypothesis of SBc' s high dependence on total citations. Therefore, SBc-index is unsuitable for identifying sleeping beauties. Bcp-index depends less on total citations than B-index. We next compare B-index and Bcp-index directly.

### 4.2 Comparing B-Index and Bcp-Index (1) Top 1% Papers (200 papers each)

Among the top 1% papers for each index, 133 overlap, with 67 unique to each ranking. We conducted t-tests on the 134 non-overlapping papers to analyze differences.

Following van Raan (2004), we define: (1) **Sleep duration**: years between publication and awakening; (2) **Sleep depth**: cumulative citation percentage during sleep period (citations before awakening divided by total citations); (3) **Awakening intensity**: cumulative citation percentage within 5 years post-awakening (or until 2015 if insufficient); (4) **Citation boost rate**: awakening intensity minus sleep depth.

Table 3 shows Bcp-index identifies sleeping beauties with shorter sleep duration than B-index. Sleep depth (cumulative percentage during sleep) is significantly smaller under Bcp-index, demonstrating its stronger constraint on pre-awakening citations. Awakening intensity (5-year cumulative percentage) is significantly higher under B-index because its sleep depth is higher. Boost rates show no significant difference between indices.

While paper age, total citations, and maximum annual citations show no significant differences, Bcp-index is more sensitive to older sleeping beauties and better identifies those with lower total and annual citations, confirming its lower dependence on total citations.

### (2) Top 0.1% Papers (20 papers each)

Comparing top 0.1% papers (20 each), 8 overlap in the top 10, showing Bcp-index similarly identifies extreme sleeping beauties like B-index. However, the 10th-ranked B-index paper shows insignificant sleeping beauty characteristics. Among the top 20, 12 overlap. Comparing the 8 unique papers from each list (Table 4 italics, Figures 7 [Figure 7: see original paper] and 8 [Figure 8: see original paper]) reveals B-index inadequately constrains early citations—5 of its unique papers have high early citations, with 2 showing significant later declines. Bcp-index's unique top-20 papers clearly exhibit sleeping beauty characteristics, with overall significantly increasing annual citation trends due to incorporating the complete citation curve. Thus, Bcp-index is more precise than B-index.

## 5. Conclusion

Building upon the Beauty Coefficient (B-index), this paper proposes a new simple parameter-free index—the Bcp-index—by optimizing the measurement from

"annual citation counts" to "annual cumulative percentage of citations," and redefines awakening time within this framework. Theoretical analysis and empirical findings demonstrate that Peruzzo' s (2015) SBc-index, based on annual cumulative citations, is unsuitable for identifying sleeping beauties due to its high dependence on total citations. Compared to the B-index, Bcp-index offers several advantages:

1. **Inherits B-index sensitivity** to extremely prominent sleeping beauties: 80% of top-10 papers overlap between indices.
2. **More realistic awakening times**: Bcp-framework awakening times align better with reality, while B-index times tend to be delayed.
3. **Incorporates complete citation curves**: Bcp-index overcomes B-index' s limitation of only considering citations up to the annual maximum.
4. **Stronger constraint on early citations**: Bcp-index better matches sleeping beauties' characteristic zero/low initial citations.
5. **Lower dependence on citation magnitudes**: B-index and SBc-index' s high dependence on total citations focuses attention on highest-cited papers, potentially overlooking meaningful sleeping beauties among moderately-cited literature. Bcp-index is more sensitive to older papers and those with lower total and annual citations.
6. **Parameter-free metrics**: Using cumulative percentages allows redefining sleep depth (cumulative percentage before awakening) and awakening intensity (cumulative percentage within 5 years post-awakening), calculating their difference to reflect relative citation boosts. These metrics avoid subjective thresholds (e.g., 1-2 citations/year during sleep, 5 citations/year during awakening) and disciplinary citation pattern differences.

In summary, Bcp-index measures citation impact delay more precisely than B-index. Future research will analyze the disciplinary distribution, research types, and bibliometric characteristics (authors, institutions, countries, journals, references) of sleeping beauties identified from *Science* and *Nature*, classifying them by discovery type and delayed recognition causes. Comparative studies with control papers (highly-cited papers from same journal/period, same-topic highly-cited papers from other journals) will reveal typical sleeping beauty characteristics.

---

## References

[1] van Raan, A. F. J. (2004). Sleeping Beauties in science. *Scientometrics*, 59(3), 467-472. [2] Campanario, J. M. (2009). Rejecting and resisting Nobel class discoveries: Accounts by Nobel Laureates. *Scientometrics*, 81(2), 549-565.

[3] Fang, H. (2014). An explanation of resisted discoveries based on construal-level theory. *Science and Engineering Ethics*, 21(1), 41-50. [4] Wang, J., Ma, F., Chen, M., & Rao, Y. (2012). Why and how can 'sleeping beauties' be awakened? *The Electronic Library*, 30, 5-18. [5] Ke, Q., Ferrara, E., Radicchi, F., & Flammini, A. (2015). Defining and identifying Sleeping Beauties in science. *Proceedings of the National Academy of Sciences of the United States of America*, 112(24), 7426-7431. [6] 李江, 姜明利, 李玥婷. 引文曲线的分析框架研究——以诺贝尔奖得主的引文曲线为例. 中国图书馆学报, 2014, 40(2): 41-49. [7] Baumgartner SE, Leydesdorff L. Group-Based Trajectory Modeling (GBTM) of Citations in Scholarly Literature: Dynamic Qualities of "Transient" and "Sticky Knowledge Claims" . *Journal of the American Society for Information Science and Technology*, 2013, 65(4): 797–811. [8] Glänzel, W., & Garfield, E. (2004). The myth of delayed recognition. *Scientist*, 18(11), 8. [9] Glänzel, W., Schlemmer, B., & Thijs, B. (2003). Better late than never? On the chance to become highly cited only beyond the standard bibliometric time horizon. *Scientometrics*, 58(3), 571-586. [10] Kozak, M. (2013). Current science has its "sleeping beauties" . *Current Science*, 104(9), 1129. [11] Ohba, N., & Nakao, K. (2012). Sleeping beauties in ophthalmology. *Scientometrics*, 93(2), 253-264. [12] van Raan, A. F. J. (2015). Dormitory of Physical and Engineering Sciences: Sleeping Beauties May Be Sleeping Innovations. *PLoS ONE*, 10(10), e0139786. [13] Costas, R., Van Leeuwen, T. N., & Van Raan, A. F. J. (2010). Is scientific literature subject to a "sell-by-date" ? A general methodology to analyze the "durability" of scientific documents. *Journal of the American Society for Information Science and Technology*, 61(2), 329-339. [14] Costas, R., van Leeuwen, T. N., & van Raan, A. F. J. (2011). The "Mendel syndrome" in science: Durability of scientific literature and its effects on bibliometric analysis of individual scientists. *Scientometrics*, 89(1), 177-191. [15] Costas, R., Van Leeuwen, T. N., & Van Raan, A. F. J. (2013). Effects of the durability of scientific literature at the group level: Case study of chemistry research groups in the Netherlands. *Research Policy*, 42(4), 886-894. [16] Li, J, & Shi, D. B. (2016). Sleeping beauties in genius work: When were they awakened? *Journal of the Association for Information Science and Technology*, 67(2), 432-440. [17] Sun, J. J., Min, C., & Li, J. (2016). A Vector for Measuring Obsolescence of Scientific Articles. *Scientometrics*, 1-13. doi: 10.1007/s11192-016-1884-7. [18] Wang, J. (2013). Citation time window choice for research impact evaluation. *Scientometrics*, 94(3), 851-872. [19] 杜建, 武夷山. 文献引文轨迹：分类及测度. 情报理论与实践, 2015, 38(7): 52-58. [20] 杜建, 武夷山. 睡美人与王子文献的识别方法研究. 图书情报工作, 2015, 59(19): 84-92. [21] Peruzzo F. Sleeping beauties and the citation dynamics in the network of scientific papers. 2015. http://tesi.cab.unipd.it/50039/1/Peruzzo_Fabio.pdf. [22] Li J, Ye F Y. (2016). Distinguishing sleeping beauties in science. *Scientometrics*, 108(2), 1-8.

*Note: Figure translations are in progress. See original paper for figures.*

*Source: ChinaXiv —Machine translation. Verify with original.*