3-22-2021

# scite: The next generation of citations

Sean Rife
*Murray State University*, srife1@murraystate.edu

Domenic Rosati
*scite.ai*

Joshua M. Nicholson
*scite.ai*

# scite: The Next Generation of Citations

Sean C. Rife[1,2], Domenic Rosati[1], and Joshua M. Nicholson[1]

1. scite, Inc., Brooklyn, NY, USA

2. Murray State University, Murray, KY, USA

Address correspondence to: Sean C. Rife, PhD, and Joshua M. Nicholson, PhD, scite Inc., 334 Leonard St, #6, Brooklyn, NY 11249, USA
Email: josh@scite.ai; sean@scite.ai

## Key Points

- While the importance of citation context has long been recognized, simple citation counts remain as a crude measure of importance.
- Providing citation context should support the publication of careful science instead of headline-grabbing and salami-sliced non-replicable studies.
- Machine learning has enabled the extraction of citation context for the first time, and made the classification of citation types at scale possible.

**The Problem: Citation counts are not enough**

Bibliographic metrics are a pervasive - and necessary - aspect of science and other academic research. Since the 1960s, these metrics have driven decisions on journal subscriptions, grant allocations, tenure and promotion, and individual assessments of scholarly impact. Such metrics, while pervasive, all rely upon crude citation counts, ignoring the context of and meaning of the citations - that is, what the authors of a paper say while making a citation (hereafter referred to as “citation context”).

In any other system of assessment or review, this practice would surely seem odd. Imagine if Rotten Tomatoes simply indicated how many times a movie had been reviewed but withheld both the content of the review and the actual ratings that were given with the review; or if Amazon simply indicated how many times a product had been reviewed without revealing the average star rating. The fact that in so many other areas, the content of a reference is evaluated rather than the simple fact that a reference exists, should give pause to bibliometric researchers.

The practice of relying solely on citation counts and ignoring context has a number of downstream effects. Researchers are incentivized to publish as much as possible, possibly without regard to whether or not their work makes a substantial contribution to the scientific literature, or whether or not the work is of high quality. At the same time, some of the most important insights have come from a relatively small number of papers. The physicist Peter Higgs, who predicted the discovery of the particle that bears his name, has famously remarked that the current climate - with its emphasis on churning out a constant stream of papers - makes groundbreaking discoveries nearly impossible (Aitkenhead, 2013). Similarly, simply examining the number of citations to a given paper does not necessarily indicate the findings it presents are valuable insights; indeed, some of the most highly-cited papers in the field of social psychology (a discipline recently revealed to be filled with irreproducible findings) are to papers that present results that subsequently fail to replicate (see, e.g., Baumeister, Bratslavsky,

Muraven, & Tice, 1998; Hagger et al., 2016; Bargh, Chen, & Burrows, 1996; Doyen, Klein, Pichon, & Cleeremans, 2012). Clearly, quantity and quality are not neatly coupled with respect to scientific insights. In recent years, a growing number of commentators have called for a move away from rewarding a larger quantity of publications, focusing instead on rigor and replicability among a smaller number of papers (e.g., Kiai, 2019).

Why has the field of bibliometrics relied so heavily on simple citation markers and ignored citation contexts? The importance of citation contexts has long been recognized (see, e.g., Garfield, 1959). The underlying problem thus far is that unlike other types of metrics (e.g., movie and product reviews), citations in scientific literature do not contain easily-quantifiable markers (e.g., stars, rankings, etc). The classification of a given citation requires that the author indicate the intent of their citation explicitly, as has been suggested and trialed in certain journals (Willighagen, 2020), or that the citation text be extracted for analysis by machines and/or the scientific community. The amount of scientific literature makes the prospect of manually classifying citations virtually impossible and the technical challenges in automating citation context extraction on scholarly documents have been previously insurmountable. Those challenges would have prevented previous generations of bibliometrics from incorporating this information However, recent advances in machine learning have enabled automatic classification of scientific citations at scale possible.

This type of automated citation classification is the focus of scite - a new platform that extracts and analyzes the content of scientific citations, and in doing so provides a novel type of bibliometric indicator, applicable at the level of the article, author, journal, funder, institution, and even publisher.

**The Platform: scite**

scite ingests scientific papers from a variety of sources, including publishers, open access articles, and other sources. It then identifies citations within the text of each paper,

extracts the context (the text surrounding the citation marker), and uses a deep learning model to classify the citation as supporting, disputing, or simply mentioning the article being cited. Importantly, this model is trained on a dataset of text from scientific publications, and goes beyond simple sentiment analysis; that is, the individuals annotating the text used to train the model were instructed to assess whether or not a citing statement provided supporting or disputing evidence to the paper being cited resulting in a training set that captures citation intent. The resulting qualified citations capturing a classification and context are called “Smart Citations” and are provided to researchers and research organizations in the form of various research tools.

At the time of this writing, scite has ingested papers from over 23 million articles, and the scite database contains over 800 million citation statements, all of which have been classified, and are available to researchers primarily in the form of interactive reports. A scite report displays all citations of a given paper, and the report can be filtered by publication year and type, as well as the type of citation - supporting, mentioning, or disputing - as classified by scite’s machine learning algorithm. Researchers are also able to upload a manuscript of their own and check how the references they are citing have been received by other papers including whether those references have been highly disputed, retracted, or have been previously supported. Finally, researchers are able to see aggregate citation information, such as number of supporting citations, for journals, funders, and a custom set of publications.

**Implications**

Using Smart Citations, scholars can assess citations on their merits rather than relying on simple citation counts as a measure of importance. This will make the process of conducting a literature review both faster and more comprehensive: authors will be able to select references that are well-supported (or, at the very least, not plagued by published replication failures, for example). Reviews will be written with a more complete understanding of the

context in which the papers it cites reside. Researchers could choose their methods with a fuller understanding of what has worked in the past. Overall, authors will be more informed about the papers they choose to read or cite. Finally, if a metric based on citation contexts gains widespread acceptance, this could encourage additional replication efforts, as researchers would be incentivized to affect the reputation of a given paper, finding, or methodology.

Taken together, normalizing the examination of context-aware citation metrics (as well as summaries of how an article has been cited) rather than relying exclusively on existing metrics (e.g., Impact Factor, H-index) that do not take context into account could incentivize quality over quantity. That is, if researchers know that their work will be evaluated based on *how* it is cited as opposed to simply *how often* it is cited, they might be more likely to prioritize research that is more likely to replicate. The pursuit of novel, headline-grabbing findings at the expense of slow, careful, iterative science that has plagued the social and life sciences might become a thing of the past.

## Conflicts of Interest

The authors are shareholders and/or consultants or employees of Scite Inc.

## References


Aitkenhead, D. (2013). Peter Higgs: I wouldn't be productive enough for today's academic system. *The Guardian*. Retrieved from https://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244.

Baumeister, R. F., Bratslavsky, E., Muraven, M., & Tice, D. M. (1998). Ego depletion: Is the active self a limited resource?. *Journal of Personality and Social Psychology, 74*(5), 1252-1265.

Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind?. *PLoS ONE, 7*(1), e29081.

Garfield, E. (1965). Can citation indexing be automated? In Statistical association methods for mechanized documentation, symposium proceedings (Vol. 269, pp. 189-192).

Hagger, M. S., Chatzisarantis, N. L., Alberts, H., Anggono, C. O., Batailler, C., Birt, A. R., ... & Zwienenberg, M. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science, 11*(4), 546-573.

Kiai, A. (2019). To protect credibility in science, banish “publish or perish”. *Nature Human Behavior, 3*(10), 1017-1018.

Willighagen, E. (2020). Adoption of the Citation Typing Ontology by the Journal of Cheminformatics. *Journal of Cheminformatics, 12*(47), 1-3.

## Appendices

**SUPPORTING**

"When energy is depleted, job strain is likely to occur. This is in line with Baumeister, Bratslavsky, Muraven, and Tice's (1998) theory on ego depletion, which suggests that volitional acts draw on a limited portion of energy resources. Subsequent acts that require self-control use up this limited energy, and exhaust it…."

**DISPUTING**

"…there was no significant main effect of ego-depletion condition on task performance in Study 3, which is in contrast to previously found ego-depleting effects of the e-erasing task on secondary self-control tasks (e.g., Baumeister et al, 1998; Tice et al, 2007)."

**MENTIONING**

"The mean time spent on the persistence task in our study is lower than those reported by Baumeister et al (1998). We believe this is because of a critical difference in the procedure…."

**Figure 1.** *Example citations for each category.*

*Note: Classification of citations is based on information from discourse markers such as the highlighted text.*