


# Reply to Bakker et al.: Assessing the Accuracy of the Scite Citation Classification System Requires the Same Definitions to be Used for Training as for Testing

Sean C. Rife, PhD<sup>a</sup>, Joshua M. Nicholson, PhD<sup>b</sup>, Ashish Uppala, MBA<sup>c</sup>, Domenic Rosati, MS<sup>d</sup>

<sup>a</sup>Head of Academic Relations at Research Solutions, Co-founder of Scite, and Associate Professor of Psychology, Murray State University, Murray, KY, <https://www.orcid.org/0000-0002-6748-0841> , [srife@researchsolutions.com](mailto:srife@researchsolutions.com)

<sup>b</sup>Chief Strategy Officer at Research Solutions and Co-founder of Scite, Brooklyn, NY, <https://www.orcid.org/0000-0002-1111-1828> , [jnicholson@researchsolutions.com](mailto:jnicholson@researchsolutions.com)

<sup>c</sup>Head of Engineering and Product, Portal Innovations, LLC, Chicago, IL, <https://www.orcid.org/0000-0001-8748-1465> , [ashish.uppala.93@gmail.com](mailto:ashish.uppala.93@gmail.com)

<sup>d</sup>Head of AI, Research Solutions and Doctoral Candidate, Dalhousie University, Brooklyn, NY, <https://orcid.org/0000-0003-2666-7615> , [drosati@researchsolutions.com](mailto:drosati@researchsolutions.com), Social Media

**Cite as:** Rife SC, Nicholson JM, Uppala A, Rosati D. Reply to Bakker et al.: Assessing the Accuracy of the Scite Citation Classification System Requires the Same Definitions to be Used for Training as for Testing. *Hypothesis*. 2025;37(1). doi:10.18060/28018



Rife, Nicholson, Uppala, Rosati. All works in *Hypothesis* are licensed under a [CC BY-NC 4.0 DEED Attribution-NonCommercial 4.0 International](https://creativecommons.org/licenses/by-nc/4.0/). Authors own copyright of their articles appearing in *Hypothesis*. Readers may copy articles without permission of the copyright owner(s), as long as the author(s) are acknowledged in the copy, and the copy is used for educational, not-for-profit purposes. For any other use of articles, please contact the copyright owner(s).

## Abstract

Bakker, Theis-Mahon, and Brown<sup>1</sup> recently presented an analysis of citations in scite.ai, a citation analysis platform. They concluded that the Scite algorithm inaccurately classified citations, particularly those they regarded as supporting previous work. While we strongly believe that independent assessments of Scite's classifications are valuable, we argue that Bakker et al.'s assessment is incorrect, and that their conclusions are due to their definition of what constitutes supporting or contrasting citations. Additionally, Bakker, et al. restricted their analyses to citations of retracted works in systematic literature reviews, which artificially limits the types of statements that could be considered supporting or contrasting a specific claim. In our reply, we document the rationale for Scite's classification scheme. We also provide examples of how Scite classifies different types of citations, as well as how these classifications differ from those presented in Bakker et al.

In a recent issue of *Hypothesis*, Bakker, Theis-Mahon, and Brown<sup>1</sup> present an analysis of citation classifications by scite.ai, a platform that facilitates citation analyses using machine learning. They conclude that the accuracy of Scite's classifications of citations is low, based on their own coding of citations. While there is value in this type of analysis, we believe there is a fundamental difference in how citation classifications are being coded and assessed; hence, the discrepancy from their analysis and our own.

First, we wish to highlight some positive aspects of Bakker et al.'s<sup>1</sup> work. As co-founders of Scite and practicing researchers, we appreciate any attempt to assess the accuracy of our citation classifications and the overall utility of the product suite Scite provides. We have published our own assessment and in-depth treatment of how Scite works elsewhere<sup>2</sup>, yet outside investigations are free from the biased attachment that comes with developing a new tool and are useful in the continual development and improvement of Scite. Additionally, we think the study of retracted publications is an interesting (and understudied) bibliometric topic; in fact, we have done our own investigation of citations of retracted publications<sup>3</sup> and uncovered a number of concerning phenomena (e.g., authors citing retracted works while apparently unaware that they have been retracted). This is not mere dicta on our part - we genuinely believe efforts like those of Bakker et al.<sup>1</sup> are valuable, and wish to see more of them in the future.

However, Bakker et al.'s<sup>1</sup> work has a critical flaw, in that it does not appear to assess what it claims it is assessing. Bakker et al. report that a citation was classified as supporting if it was included in a review "without indication of its retracted status or significant critique of the research quality," and classified as contrasting if it was "described as retracted, or if concerns about the publication or underlying research were discussed" (p. 3). While these are perfectly valid attributes one may attach to a dataset of citation statements, they differ markedly from the annotations in Scite's training dataset. In contrast, Scite's classification schema is very different. As we note<sup>2</sup>:

scite focuses on the authors' reasons for citing a paper. (...) We consider that for capturing the reliability of a claim, a classification decision into supporting or contrasting must be backed by scientific arguments. The evidence involved in our assessment of citation intent is directed to the factual information presented in the

citation context, usually statements about experimental facts and reproducibility results or presentation of a theoretical argument against or agreeing with the cited paper. (p. 888)

This approach, which was part of the training given to annotators who created the labels in Scite's training data differs markedly from that of Bakker et al.<sup>1</sup> in two critical ways: first, noting or failing to note that a cited paper has been retracted does not indicate that the citing paper is presenting supporting or contrasting evidence. It simply looks at whether the authors are aware of the retraction or not, or whether or not they meaningfully engage with the cited work. Indeed, most debates or direct refutations in science have little to do with retractions; for example, there are over 64,000 retractions in the scholarly literature at the time of this writing<sup>3</sup>, while there are over 6.4 million contrasting citations in the Scite database. Second, while important, retractions are distinct from disagreements and distinct from looking at supporting or contrasting evidence. For example, in another independent study looking at disagreements in the literature, there was no mention of retractions or use of retractions in the definition of what is a scientific disagreement<sup>4</sup>.

To provide concrete examples, Table 1 (reprinted from Nicholson et al.<sup>2</sup>) shows a sample of citations and how they are classified, as well as an explanation of each classification. We should note that Scite's approach to citation classification is more precise than sentiment analysis (the automatic classification of statements as generally positive or generally negative), in that it focuses specifically on whether or not new evidence supports or contrasts a claim. As such, statements that refer to another paper with a positive or negative valence are deliberately classified by Scite as mentioning (see, for example, the fifth example in Table 1).

Table 1. Examples of Citation Classifications

<b>Citation statement</b>	<b>Classification</b>	<b>Explanation</b>
“In agreement with previous work (Nicholson et al., 2015), the trisomic clones showed similar aberrations, albeit to a lesser extent (Supplemental Figure S2B).”	Supporting	“In agreement with previous work” indicates support, while “the trisomic clones showed similar aberrations, albeit to a lesser degree (Supplemental Figure S2B)” provides evidence for this supporting statement.
“In contrast to several studies in anxious adults that examined amygdala activation to angry faces when awareness was not restricted (Phan, Fitzgerald, Nathan, & Tancer, 2006; Stein, Goldin, Sareen, Zorrilla, & Brown, 2002; Stein, Simmons, Feinstein, & Paulus, 2007), we found no group differences in amygdala activation.”	Contrasting	“In contrast to several studies” indicates a contrast between the study and studies cited, while “we found no group differences in amygdala activation” indicates a difference in findings.
“The amygdala is a key structure within a complex circuit devoted to emotional interpretation, evaluation and response (Stein et al., 2002; Phan et al., 2006).”	Mentioning	This citation statement refers to Phan et al. (2006) without providing evidence that supports or contrasts the claims made in the cited study.
“In social cognition, the amygdala plays a central role in social reward anticipation and processing of ambiguity [87]. Consistent with these findings, amygdala involvement has been outlined as central in the pathophysiology of social anxiety disorders [27], [88].”	Mentioning	Here, the statement “consistent with these findings” sounds supportive, but, in fact, cites two previous studies: [87] and [27] without providing evidence for either. Such cites can be valuable, as they establish connections between observations made by others, but they do not provide primary evidence to support or contrast the cited studies. Hence, this citation statement is classified as mentioning.
“For example, a now-discredited article purporting a link between vaccination and autism (Wakefield et al., 1998) helped to dissuade many parents from obtaining vaccination for their children.”	Mentioning	This citation statement describes the cited paper critically and with negative sentiment but there is no indication that it presents primary contrasting evidence, thus this statement is classified as mentioning.

In contrast, Bakker et al.'s methodology classifies citations based on methodological discussion and acknowledgement of a target article's retraction. This results in many citations that Scite (justifiably, we believe, given the criteria outlined in Nicholson et al.) classifies as mentioning being classified as supporting. (A previous study<sup>5</sup>, which describes the data used in Bakker et al. in greater detail, does not provide access to the specific examples used in Bakker et al. As such, we present randomly-selected examples of citations to retracted works in systematic reviews.) Consider the following citation<sup>6</sup> in Gatto (2020; emphasis added)<sup>7</sup>, which Scite classifies as mentioning:

“Interestingly, recent studies have demonstrated that magnesium treatment protects cognitive function and synaptic plasticity by inhibiting GSK-3 $\beta$  in sporadic Alzheimer's disease (AD) model rats (Xu et al, 2014). **As such, Mg 2+ ions are a critical factor in controlling synapse density and plasticity, showing a reduction in A $\beta$ -plaques and cognitive deficits in APPswe/PS1dE9 mice, a transgenic mouse model of AD (Li et al, 2014).** However, the effects of dietary Mg 2+ deficiency on learning and memory are not followed by changes in the spine density and morphology of hippocampal neurons of rodent models (Serita et al, 2019), and further examination of this topic is undoubtedly needed. (p. 573)”

Since this citation does not mention that Li et al. is retracted, nor does it contain a significant methodological critique, Bakker et al. would classify it as supporting. Another example can be found in a review by Zhang et al. (2021<sup>8</sup>; emphasis added), citing Kaur (2014)<sup>9</sup>, which Scite also classifies as mentioning:

“Some cross-sectional studies have shown a close association between thyroid disease and metabolic disorders (i.e., metabolic syndrome [MetS] and its components) (11–13). **MetS is characterized by a cluster of abnormal metabolic parameters consisting of insulin resistance, central obesity, type 2 diabetes, impaired glucose tolerance, hyperinsulinemia, and dyslipidemia (14).** The global prevalence of MetS is between 11.6% and 62.5% (15). (p. 2)”

Again, because Kaur does not mention the fact that Zhang et al. was retracted, Bakker et al. would classify this citation as supporting.

There are also instances where Bakker et al. would presumably classify a citation as contrasting while Scite would classify it as mentioning. For example, one paper<sup>10</sup> cites Tsukumo et al. (2007; emphasis added)<sup>11</sup>, stating:

“A total of 37% of the retrieved articles had a positive-citation pattern; meanwhile, 63% had a negative-citation pattern. **The most cited article with a negative-citation-pattern was published in 2007 and was retracted in 2016 [24].** Thus far, it has received a total of 490 citations and of these, 58 were from after the retraction of the article. (p. 9)”

Because the citing paper does not present findings contrary to Tsukumo et al., Scite classifies this citation as mentioning. However, because it acknowledges the retraction of the target article, Bakker et al. would classify this citation as contrasting.

In summary, we think that while Bakker et al.'s efforts are laudable, their coding system bears no resemblance to that of the Scite training dataset. We therefore strongly disagree with

Bakker et al.'s conclusion that the accuracy of Scite's classifications is low. While we would all benefit from outside assessments of Scite's accuracy, Scite included, this will have to come from future works that rely on a classification scheme that is the same, or at least similar, to that of Scite. Until such a paper is published, we recommend that readers consult our previous work<sup>2</sup> for information on the accuracy of Scite's classifications.

## Disclosure of Conflicts

All authors are employees/contractors of and shareholders in Research Solutions, Inc., the company that owns [Scite](#).

## References

1. Bakker C, Theis-Mahon N, Brown SJ. Evaluating the accuracy of scite, a smart citation index. *Hypothesis: Research Journal for Health Information Professionals*. 2023 Sep 13;35(2). doi:10.18060/26528
2. Nicholson JM, Mordaunt M, Lopez P, Uppala A, Rosati D, Rodrigues NP, Grabitz P, Rife SC. Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning. *Quant Sci Stud*. 2021 Nov 5;2(3):882-98. doi:10.1162/qss\_a\_00146
3. scite.ai: scite Search; [reviewed 2024 May 21; cited 024 May 21]. Available from: <https://scite.ai/search?hasRetraction=true&mode=all>
4. Lamers WS, Boyack K, Larivière V, Sugimoto CR, van Eck NJ, Waltman L, Murray D. Meta-Research: Investigating disagreement in the scientific literature. *Elife*. 2021 Dec 24;10:e72737. doi:10.7554/eLife.72737
5. Brown SJ, Bakker CJ, Theis-Mahon NR. Retracted publications in pharmacy systematic reviews. *J Med Libr Assoc*. 2022 Jan 1;110(1):47. doi:10.5195/jmla.2022.1280
6. Li W, Yu J, Liu Y, Huang X, Abumaria N, Zhu Y, Huang X, Xiong W, Ren C, Liu XG, Chui D. Elevation of brain magnesium prevents synaptic loss and reverses cognitive deficits in Alzheimer's disease mouse model. *Mol Brain*. 2014 Dec;7(1):1-20. doi:10.1186/s13041-014-0065-y
7. Gatto RG. Molecular and microstructural biomarkers of neuroplasticity in neurodegenerative disorders through preclinical and diffusion magnetic resonance imaging studies. *J Integr Neurosci*. 2020 Sep 30;19(3):571-92. doi:10.31083/j.jin.2020.03.165
8. Zhang C, Gao X, Han Y, Teng W, Shan Z. Correlation between thyroid nodules and metabolic syndrome: a systematic review and meta-analysis. *Front Endocrinol*. 2021 Sep 16;12:730279. doi:10.3389/fendo.2021.730279
9. Kaur J. A comprehensive review on metabolic syndrome. *Cardiol Res Pract*. 2014 Oct;2014. doi:10.1155/2014/943162

10. Stavale R, Ferreira GI, Galvão JA, Zicker F, Novaes MR, Oliveira CM, Guilhem D. Research misconduct in health and life sciences research: A systematic review of retracted literature from Brazilian institutions. *PLoS One*. 2019 Apr 15;14(4):e0214272. doi: 10.1371/journal.pone.0214272

11. Tsukumo DM, Carvalho-Filho MA, Carvalheira JB, Prada PO, Hirabara SM, Schenka AA, Araujo EP, Vassallo J, Curi R, Velloso LA, Saad MJ. Loss-of-function mutation in Toll-like receptor 4 prevents diet-induced obesity and insulin resistance. *Diabetes*. 2007 Aug 1;56(8):1986-98. doi:10.2337/db06-1595