Using Citation Behavior to Rethink Academic Impact in Software Engineering

Simon Poulding*, Kai Petersen*, Robert Feldt* and Vahid Garousi†
*Blekinge Institute of Technology, Karlskrona, Sweden, Email: {simon.poulding,kai.petersen,robert.feldt}@bth.se
†Hacettepe University, Ankara, Turkey, Email: vahid.garousi@hacettepe.edu.tr

***Abstract*—Although citation counts are often considered a measure of academic impact, they are criticized for failing to evaluate impact as intended. In this paper we propose that software engineering citations may be classified according to how the citation is used by the author of the citing paper, and that through this classification of citation behaviour it is possible to achieve a more refined understanding of the cited paper's impact. Our objective in this work is to conduct an initial evaluation using the citation behaviour taxonomy proposed by Bornmann and Daniel. We independently classified citations to ten highly-cited papers published at the International Symposium on Empirical Software Engineering and Measurement (ESEM). The degree to which classifications were consistent between researchers was analyzed in order to assess the clarity of Bornmann and Daniel's taxonomy. We found poor to fair agreement between researchers even though the taxonomy was perceived as relatively easy to apply for the majority of citations. We were nevertheless able to identify clear differences in the profile of citation behaviors between the cited papers. We conclude that an improved taxonomy is required if classification is to be reliable, and that a degree of automation would improve reliability as well as reduce the time taken to make a classification.**

I. Introduction

Citations are the primary measure by which research impact is evaluated: citation counts are used for this purpose in the recruitment of researchers, the evaluation of funding applications, the assessment of journal impact, and the comparison of research institutions. Hence it is of utmost importance that citation counts are really measuring what we intend, i.e. the impact of the cited research.

David Parnas [1] highlights the limitations of counting publications and citations. As he points out there are a variety of reasons to cite: *"Some citations are negative. Others are included only to show that the topic is of interest to someone else or to prove that the author knows the literature."* Thus not all citations reflect the same degree of impact on the authors of the citing paper, but a metric based simply on citations counts will fail to make this distinction. Indeed there is a danger that researchers may seek to maximize the number of citations to their work at the expense of producing research of genuinely high scientific relevance and rigor. By defining a measure based purely on citation count and coupling it with a reward (e.g. career progression), actions will inevitably be taken to optimize this measure; see, for example, [2].

This calls for better methods of evaluating impact. Multiple attempts have been made to improve on the h-index, the quasi-gold standard for evaluation. For example, Wohlin [3] proposed to consider the overall citation distribution rather than a single point measurement, such as the h-index. But this would still not address the issue raised by Parnas.

We propose instead to consider the motivation for citing a particular paper as a first step to a more refined evaluation of impact. There is a significant body of research into the citation behavior of scientists that considers and classifies these motivations. Bornmann and Daniel [4] provide an extensive review, and they additionally propose their own taxonomy of citation behaviour synthesized from earlier taxonomies. It is this synthesized taxonomy that we use in this paper. In particular, we evaluate the ease and consistency, between multiple researchers, of classifying citations according Bornmann and Daniel's taxonomy. If we are able to unambiguously and efficiently classify citations with regard to the motivation for the citation, there is a possibility to improve on the existing citation measures.

The specific context of our evaluation is citation behavior in software engineering research, and for this purpose the evaluation considers citations to 10 highly-cited articles published at the International Symposium on Empirical Software Engineering and Measurement (ESEM).

The remainder of the paper is structured as follows: Section II presents the research method. Section III presents the results. Section IV concludes the paper.

II. Research method

Research questions and objectives: The objective of the research is formulated as follows:

- Analyze *Bormann and Daniel's citation behaviour taxonomy* for the purpose of *evaluation*,
- with respect to *the ease of applying the taxonomy*, *the agreement between multiple researchers making the classification*. and *the nature of the profile of citation behaviours for a cited papers*,
- from the point of view of the *researcher*,
- in the context of *software engineering researchers classifying citations to software engineering literature published at ESEM*.

The following research questions were stated:

RQ1: How easy was it to classify the papers as perceived by the researchers? This research question evaluates the taxonomy from the perspective of the ease of applying Bornmann and Daniel's taxonomy as perceived by the researcher conducting the classification.

RQ2: To what extent did independent researchers agree in their classifications? If there is little consistency between independent classifications of the same citation, then it is possible that Bornmann and Daniel's taxonomy lacks the necessary clarity in the context of software engineering papers.

RQ3: What is the citation profile of the cited papers? The purpose of this research question was to determine how the citation profiles—the relative proportions of the different citation behaviours identified by Bornmann and Daniel's taxonomy—differed between the cited papers. Our argument is that the citation profile can provide a more detailed understanding as to the impact of the cited paper since, for example, a citation indicating a use of methodology proposed in the cited paper is likely to indicate more academic impact than a perfunctory citation to the cited paper.

Data collection: The starting pool of *cited* papers was full papers published between 2005 and 2014 in the International Symposium on Empirical Software Engineering and Measurement (ESEM)—and prior to 2007, in the International Symposium on Empirical Software Engineering (ISESE)—with 20 citations or more (as determined by Scopus). From this pool we randomly selected 10 *cited* papers which were given ids of D01 to D10.

For each cited paper, we considered a pool of *citing* papers as being journal and conference papers written in English that included one or more citations to the cited paper (again as determined by Scopus). For each cited paper, we randomly selected 10 *citing* papers.

The data collection was conducted in two phases. In the first phase, two of the cited papers, D01 and D09, were selected to pilot the data collection process as they were of different types. All four authors independently collected the following data from the citing papers:

- Identify the citation(s) to the cited paper.
- Record the text surrounding the citation.
- Classify each citation according to Bornmann and Daniel's taxonomy, using their descriptions of each category reproduced here in Table I.
- Assess how easy it was to classify the citation, using a 5-point Likert scale.

Based on the experience of this pilot, the process for recording multiple citations was clarified: if, for example, two citations were part of the same 'semantic' citation, such as repeated usage in a paragraph discussing a single idea, then these were treated and classified as if they were one citation. The pilot was also useful in improving familarity with Bornmann and Daniel's taxonomy, and since it was a learning exercise, the data collected from the pilot is not used in the analysis below.

In the second phase, each of the remaining eight *cited* papers were assigned two researchers drawn from the authors of this paper. The same data collection was performed as in the first phase, and the researchers performed the collection and classification independently.

Data analysis: Prior to analysis, the citations in each citing paper found by each of the two researchers were matched based on the surrounding text. In some cases, one researcher identified a citation to the cited paper that the other researcher did not. Where one researcher had merged multiple citations into a single 'semantic' instance, but the other had not, the single classification made by the first researcher was duplicated in order to match the multiple citations of the second researcher.

TABLE I. BORNMANN AND DANIEL'S TAXONOMY OF CITATION BEHAVIOUR FROM [4]

Category	Description
Assumptive	Citing work refers to assumed knowledge that is general/specific background; citing work refers to assumed knowledge in an historical account; citing work acknowledges cited work pioneers.
Perfunctory	Citing work makes a perfunctory reference to cited work; cited work is cited without additional comment; citing work makes a redundant reference to cited work; cited work is not apparently strictly relevant to the author's immediate concerns
Persuasive	Cited work is cited in a "ceremonial fashion"; the cited work is authored by a recognized authority in the field
Conceptual	Use of definitions, concepts, or theories of cited work
Methodological	Use of materials, equipment, practical techniques, or tools of cited work; use of analysis methods, procedures, and design of cited work
Affirmational	Citing work confirms cited work; citing work is supported by cited work; citing work depends on cited work; citing work agrees with ideas or findings of cited work; citing work is strongly influenced by cited work
Contrastive	Citing work contrasts between the current work and cited work; citing work contrasts other works with each other; citing work is an alternative to cited work
Negational	Citing work disputes some aspects of cited work; citing work corrects/questions cited work; citing work negatively evaluates cited work

After matching the citation data, the following analysis was performed:

- For each cited paper, the agreement between the two researchers on the classification of each citation was quantified using Cohen's Kappa. The Kappa was also calculated across all cited papers.
- For each cited paper, the frequency was calculated of each category in Bornmann and Daniel's taxonomy among the citations in the citing papers.
- Across all citations, the frequency of each point of the ease of classification Likert scale.

Validity: The key threat to validity is the external validity. As we are piloting the Bornemann and Daniel classification for software engineering literature, we only chose a specific forum with a well defined scope (ESEM). For different forums the classification consistency may have been different. Also, the ability to classify may depend on the specific set of reviewers and their background. Another threat to validity is the learning effect, in this case reflected in the time needed to learn and understand the classification. For this purpose, a pilot study has been conducted to reduce this threat.

III. RESULTS

The results are structured based on the three research questions stated in Section II.

Ease of use (RQ1): Figure 1 summarizes the perceived ease of applying Bornmann and Daniel's taxonomy. 51% of the citations were found to be "very easy" or "easy" to classify, while comparatively few (18%) were perceived as "hard" or "very hard".

Level of agreement between reviewers (RQ2): Table II reports the Kappa value analysis. Values close to one indicate near-complete agreement, while values close to zero indicate agreement little different from random.

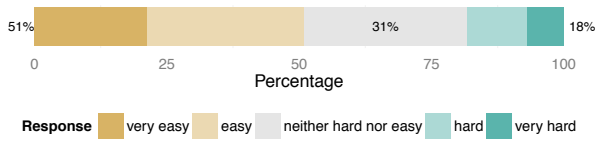

Fig. 1. Ease of applying the Bornmann and Daniel taxonomy

In addition to reporting the Kappa values among the eight categories of Bornmann and Daniel's taxonomy, the columns labelled "(grouped)" report the Kappa values when these eight categories are summarized into three logical groupings that we propose: citations relating to background work that does not directly contribute to the citing paper (Assumptive, Perfunctory, and Persuasive); actual usage of the cited work as input to the citing paper (Conceptual and Methodological); and direct comparison made between the citing and cited paper (Affirmational, Contrastive, and Negational). For example, if one researcher classifies a citation as Assumptive, and one as Perfunctory, this would count as a disagreement for the individual Kappa calculation, while it is considered an agreement in the grouped analysis.

If one researcher does not identify a citation in the citing paper that the other does, this is considered a disagreement for the Kappa values reported values in the set of columns headed "All Citations". In the set of columns headed "Excluding Missed Citations", the Kappa values are recalculated with these cases excluded.

The table shows that for most of the cited papers the Kappa value improves when checking consistency on a higher level of abstraction using these logical groups. This provides some support to the notion that when disagreements occur, they are more often between categories within a group rather than between categories in different groups. When looking at the data excluding missed citations, the agreeement improves slightly in most cases.

Landis and Koch [5], and Fleiss [6], propose ranges of values that may be used to interpret the results of the Kappa analysis. According to Landis and Koch, "substantial" agreement was only achieved for one paper (D08) when grouping the categories. Generally, a "slight" to "fair" agreement has been reached. When applying the classification by Fleiss, the majority of agreements would be classified as "poor". Considering all citations in aggregation, the agreement was "fair" according to Landis and Koch and "poor" according to Fleiss.

Figure 2 shows the relative frequency of the each possible

TABLE II. KAPPA ANALYSIS OF CLASSIFICATION AGREEMENT

	All Citations		Excluding Missed Citations	
Paper ID	Kappa	Kappa (grouped)	Kappa	Kappa (grouped)
D02	0.249	0.478	0.320	0.651
D03	0.008	-0.078	0.012	-0.091
D04	0.250	0.379	0.181	0.330
D05	0.123	0.272	0.123	0.273
D06	0.306	0.349	0.317	0.364
D07	0.116	0.037	0.125	0.028
D08	0.416	0.756	0.447	0.859
D10	0.333	0.172	0.509	0.308
All	0.272	0.371	0.297	0.377

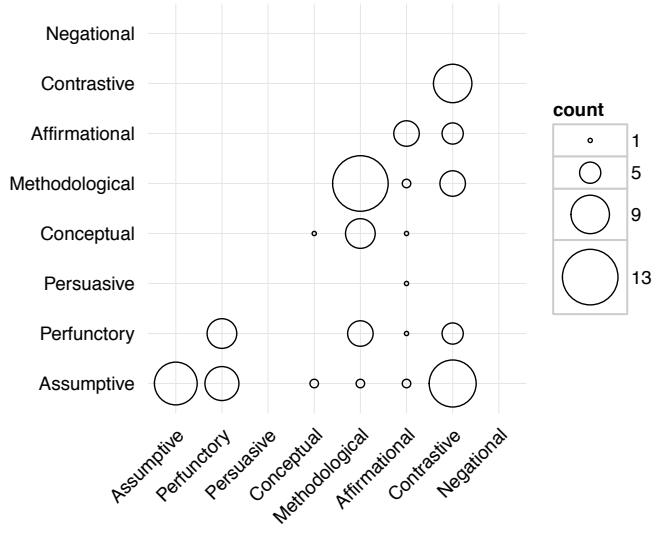

Fig. 2. Balloon plot of classification pairs

pair of classification, i.e. the two classifications to a single citation made by the two indepedent researchers. It demonstrates that (a) which categories are scored most consistently, and (b) the categories that reviewers mostly disagree on. The most commonly agreed on categories were Methodological, Assumptive, and Contrastive. The most commonly disagreed on pairs of categories were: (Assumptive; Contrastive), (Assumptive; Perfunctory), (Conceptual; Methodological), (Contrastive; Affirmational), and (Perfunctory; Methodological). It is interesting that only very few citations were classified as either Persuasive or Negational.

Consistency of classification per cited paper (RQ3): Figure 3 shows the number of citations to the cited paper in each of the categories proposed by Bornmann and Daniel. The counts have been normalized so that if a paper had four citations, then each citation contributed 0.25 to the count for that category. Only two papers had a clear, dominating category (D03 and D06), while in the remaining papers two or three citation categories stood out. We note that D02, D03, D04, D05, and D07 are empirical studies; while D06, D08 and D10 are methodology papers. The clear differences in the profiles for, say D03 and D06, are not unexpected given the different paper types.

IV. CONCLUSION

Lesson 1: Need for an improvement of the classification: The ease of applying Bornmann and Daniel's taxonomy (RQ1) was perceived as largely positive, with only relatively few citations perceived as either hard or very hard to classify. But the agreement between classifications by independent researchers (RQ2) was relatively poor. We therefore conclude that there is a need to disambiguate the taxonomy. On a more detailed level, it is interesting to look at examples of the most frequent disagreements, which need to be further investigated, such as: Perfunctory and Assumptive; Methodological and Conceptual; Contrastive and Perfunctory. In an improved classification scheme we suggest that a few, well-defined categories are used—such as background, conceptual/methodological usage,

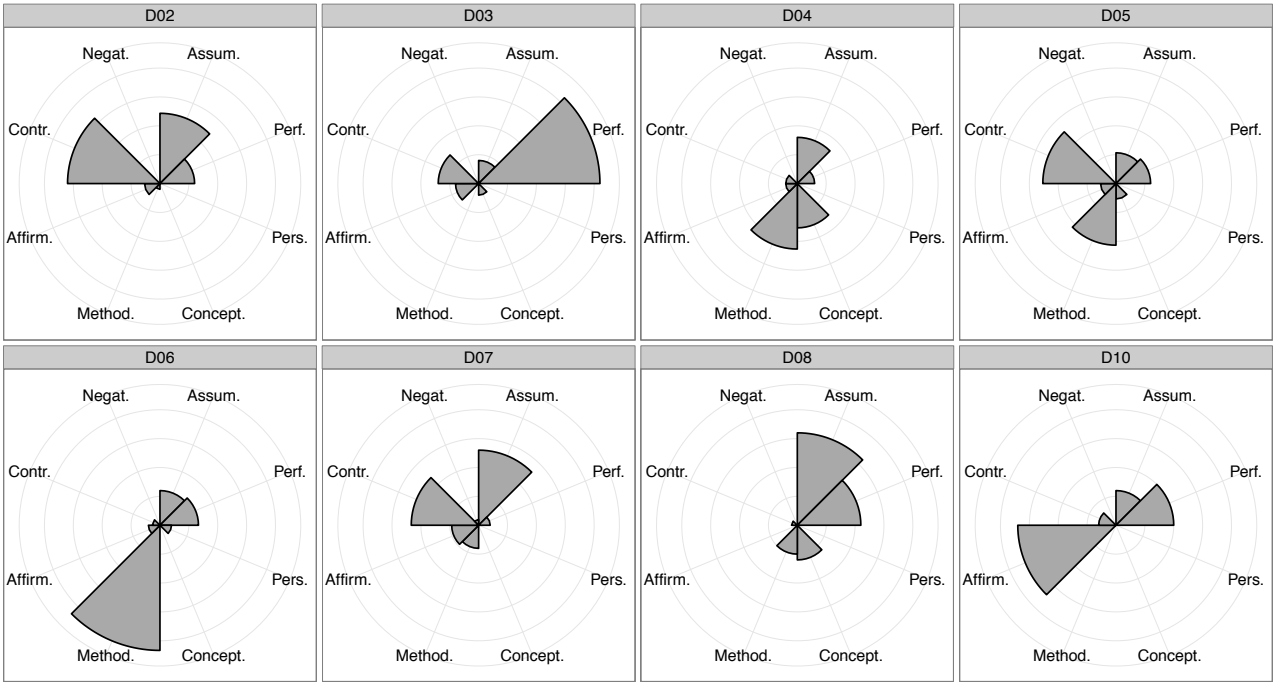

Fig. 3. Citation behavior profiles of the cited papers

and comparison)—since grouping Bornmann and Daniel's categories in this manner increased the agreement.

Lesson 2: Citation profile patterns are obtainable, though not reliable: There is value in profiling a cited paper based on the citation behaviour in its citing papers in order to judge its 'true' impact; the profile more clearly shows the impact of the research than a simple count of citations. As an example, Figure 3 shows that impact of D06 is largely methodological, while impact of D03 is largely perfunctory. We argue that the impact of D06 is more significant that D03 since perfunctory citations have little or no direct impact on the research described in the citing paper. An interesting observation was that only very few citations fall into the Persuasive and Negational categories. We speculate that in the case of Persuasive, knowledge of the specific research field of the cited and citing papers is required to classification on the basis of "cited work is authored by a recognized authority in the field". The low number of Negational citations may be due to publication bias, or may be specific to the community studied. Given Lesson 1, the major confounding factor in the use of citation profiles is the reliability of the classification. We also recommend to use radar charts to illustrate the citation profile, as it clearly highlights the dominating categories (see Figure 3).

Lesson 3: Need for semi-automation: Classifying the citations in 10 citing papers took a single researcher approximately 45 minutes. Analyzing a set of papers with a few thousand citations would therefore not be feasible using a manual process. Hence, there is a need to automation, which requires a clear taxonomy and understanding of the classes. A potential solution would require to combine different approaches, such as natural language processing, topic modeling, and machine learning. Additional information (e.g. the section of the citing paper in which the citation occurs) may be useful to support semi-automated classification.

Vision and future work: The ability to assess different types of citation behavior opens the door to rich analyses. We might, for example, investigate how the quality of a paper affects the different dimensions of its impact as measured by the citation profile: do scientific rigor and practicality relate to the citation of the work in a methodological, affirmational, contrastive, and negational way? The analysis of a complex network of citation relationships may yield more information when edges in the network are filtered by citation behavior. The next steps to support this vision are to: (1) develop a taxonomy of citation behavior that faciliates more consistent application, and (2) semi-automate the classification process. In addition, we see the need for qualitative studies capturing the rationale and motivation of authors through interviews, as well as replications of the study presented in this paper.

REFERENCES

[1] D. L. Parnas, "Stop the numbers game," *Commun. ACM*, vol. 50, no. 11, pp. 19–21, 2007. [Online]. Available: http://doi.acm.org/10.1145/1297797.1297815

[2] P. Weingart, "Impact of bibliometrics upon the science system: Inadvertent consequences?" *Scientometrics*, vol. 62, no. 1, pp. 117–131, 2005.

[3] C. Wohlin, "A new index for the citation curve of researchers," *Scientometrics*, vol. 81, no. 2, pp. 521–533, 2009. [Online]. Available: http://dx.doi.org/10.1007/s11192-008-2155-z

[4] L. Bornmann and H.-D. Daniel, "What do citation counts measure? a review of studies on citing behavior," *Journal of Documentation*, vol. 64, no. 1, pp. 45–80, 2008.

[5] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *biometrics*, pp. 159–174, 1977.

[6] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical methods for rates and proportions*. John Wiley & Sons, 2013.