Postprint of article to appear in *IEEE Transactions on Software Engineering*

Metamorphic Robustness Testing: Exposing Hidden Defects in Citation Statistics and Journal Impact Factors

Zhi Quan Zhou, T.H. Tse, and Matt Witheridge

Abstract—We propose a robustness testing approach for software systems that process large amounts of data. Our method uses metamorphic relations to check software output for erroneous input in the absence of a tangible test oracle. We use this technique to test two major citation database systems: Scopus and the Web of Science. We report a surprising finding that the inclusion of hyphens in paper titles impedes citation counts, and that this is a result of the lack of robustness of the citation database systems in handling hyphenated paper titles. Our results are valid for the entire literature as well as for individual fields such as chemistry. We further find a strong and significant negative correlation between the journal impact factor (JIF) of IEEE Transactions on Software Engineering (TSE) and the percentage of hyphenated paper titles published in TSE. Similar results are found for ACM Transactions on Software Engineering and Methodology. A software engineering field-wide study reveals that the higher JIF-ranked journals are publishing a lower percentage of papers with hyphenated titles. Our results challenge the common belief that citation counts and JIFs are reliable measures of the impact of papers and journals, as they can be distorted simply by the presence of hyphens in paper titles.

Index Terms—Metamorphic robustness testing, metamorphic testing, negative testing, fault-based testing, software robustness, oracle problem, citation count, journal impact factor, Scopus, Web of Science, Google Scholar, verification and validation.

1 Introduction

In software testing, an *oracle* is a mechanism against which testers can decide whether the outcomes of test case executions are correct. In many situations, an oracle is unavailable, or is theoretically available, but practically too expensive to be applied. This is known as the *oracle problem*, a fundamental challenge in software testing [1]–[3].

Among the various approaches to addressing the oracle problem, a growing body of research has examined the concept of metamorphic testing (MT) [4], [5], and proven it to be a highly effective testing methodology [6]–[14]. Compared with conventional testing methods, MT is focused on the examination of the *relations* among the inputs and outputs of *multiple* executions of the system under test (SUT). Such relations are called *metamorphic relations* (MRs) — they are necessary properties of the intended program's functionality. Even in the absence of an oracle for each individual output, a fault can still be detected if an MR is violated for certain test cases. As an example, consider the testing of search services [15]–[18]. It can be difficult to evaluate the accuracy and completeness of the search results; nevertheless, MT can be conducted by testing the SUT against a set of MRs prescribed by the tester. Such an MR, for instance, could be: $search(A \text{ and } B) \subseteq search(A)$, where A is a search criterion and B is an additional search criterion (such as a filter).

The unique perspective of MT (inspecting the relations among multiple executions — an area seldom explored by conventional testing methods) enabled the detection of previously unknown faults in a variety of real-world mature systems. Such examples include the detection of bugs in the GCC, LLVM, and other types of compilers and code obfuscators [19]–[22], in major search engines including Google, Bing, and Baidu [17], in the Web APIs of Spotify and YouTube [18], in the navigation system Google Maps [23] and, more recently, in self-driving cars' on-board computer software [14]. MT has also been applied to test NASA software [6], [24] and systematically adopted by Adobe Systems [10], [25]. Researchers from Accenture has recently applied MT to verify industrial-strength machine learning (ML) applications [26], and has reported on their patent titled "Verifying Machine Learning through Metamorphic Testing" [27, p. 12], in which they state that their methodology "needs only a few test cases (or even just one) to identify bugs in ML applications, thereby reducing the cost of testing significantly." In August 2018, Google acquired GraphicsFuzz, a spinout company from Imperial College London, to apply metamorphic testing to graphics drivers [22], [28]–[30].

MT was initially proposed as a verification technique [4], [5]. Xie et al. [31] studied MT at the algorithm selection level, for the purpose of testing and validating machine learning classifiers. From their findings, an MR violation could show that the target algorithm was not appropriate

- *Zhi Quan Zhou and Matt Witheridge are with the Institute of Cybersecurity and Cryptology, School of Computing and Information Technology, University of Wollongong, NSW 2522, Australia. E-mails: zhiquan@uow.edu.au, mw204@uowmail.edu.au.*
- *T.H. Tse is with the Department of Computer Science, The University of Hong Kong, Pokfulam, Hong Kong. E-mail: thtse@cs.hku.hk.*

(and not just that the implementation was defective). Zhou et al. [17] studied MT at the top level (that is, the system and service level) and conducted very large scale empirical evaluations by referring to the ISO/IEC software quality model standard [32]. They in turn developed MT into a paradigm that covers verification, validation, and other types of software quality assessment, and showed that MRs could be identified by users based on what they really cared about (and not just based on the system specifications or designs given by the developer).

Among the various types of software quality characteristics, there has been an increasing concern from both industry and the research community about software *robustness*: the ability of dealing with erroneous input or unexpected situations [33], [34]. To assess robustness, the SUT needs to be tested with invalid or erroneous input [34], and a major approach for this purpose is *fuzzing*, or *fuzz testing*, where random or semi-random input is used to test the SUT [35]. Although fuzzing can generate unexpected test cases, it may not necessarily cover all types of real-life erroneous input, and the tester may not be able to fuzz the *environment*. For example, when testing a Web search engine in a real-life operational environment, it is straightforward to apply a fuzzer (fuzz testing tool) to generate random query terms, but it is difficult to change the environment (which is the real-world Internet), unless the testing is conducted in a constrained environment with mock databases. Another limitation of fuzzing is that, due to the oracle problem, it is hard for a fuzzer to detect *logic errors* (which do not crash the SUT, but instead produce incorrect output values) [21].

The present research extends MT for robustness testing beyond fuzzing, in the context of testing big data applications. Our objective is to assess the SUT's robustness in terms of producing logically correct or reasonable output for erroneous input that does not crash the system. In this research, the subject software under consideration is *automatic indexing systems* [36], which provide fundamental IT infrastructure for the present-day knowledge society. On the one hand, *successful information access in the digital information age requires robust systems of indexing and abstracting* [37]; on the other hand, such systems are difficult to test and verify due to the sheer volume of data that they process — an oracle problem for many big data applications.

More specifically, this research is focused on the testing of two major citation database systems: *Scopus* [38] and the *Web of Science* [39]. Our method makes use of MRs and the statistics collected from large amounts of system input and output data to explore system behavior and discover underlying patterns and defects. Such patterns or defects can hardly be observed when the sample size is small. For the citation database systems under study, the ultimate goal of our research is to answer the following research question:

- **RQ1** Let P be a set of publications. When $|P| = 1$, it represents an individual publication. When $|P| > 1$, it represents a collection of publications such as articles published by a specific author, organization, journal, or field of research. Does the *citation count* of P (including any score derived from the citation count) generated by a computer system faithfully reflect the *actual impact* of P?

There is an oracle problem in RQ1: The "actual impact of P" is normally unknown, non-quantifiable, or assessed subjectively. To help provide an objective and quantifiable answer to RQ1, we further propose a second research question:

- **RQ2** Let P be a set of publications. Does the *citation count* of P (including any score derived from the citation count) generated by a computer system faithfully reflect the *actual citations* of P within the scope of the database in the computer system?

Observation 1. The difference between RQ1 and RQ2 is that the former considers the "actual impact," whereas the latter considers the "actual citations." For example, could the citation count generated by the computer system be wrong (that is, be different from the "actual citations" received by P within the scope of the computer system's database)? While overcounting can be relatively easy to detect by testers, undercounting is extremely difficult to detect due to the lack of an oracle.

Observation 2. A careful consideration of RQ2 leads to two further questions, which can be considered subquestions of RQ2, explained as follows: We call x a *cited* publication, and y a *citing* publication of x, if y cites x. Let p be a publication, and $Q = \{q_1, q_2, \ldots, q_n\}$ $(n \geq 0)$ be the set of all citing publications of p.[1] Q can be divided into two disjoint sets $Q_{correct}$ and $Q_{erroneous}$ ($|Q_{correct}| \geq 0$ and $|Q_{erroneous}| \geq 0$): $Q_{correct} \cup Q_{erroneous} = Q$, and $Q_{correct} \cap Q_{erroneous} = \emptyset$, where $Q_{correct} = \{q_{k_1}, q_{k_2}, \ldots, q_{k_m}\}$ $(0 \leq m \leq n)$ is the set of all publications that have correctly cited p (that is, the reference list of q_{k_i} $(1 \leq i \leq m)$ has included complete and correct bibliographic data of p), and $Q_{erroneous} = \{q_{k_{m+1}}, q_{k_{m+2}}, \ldots, q_{k_n}\}$ is the set of all publications that have incorrectly cited p (that is, the reference list of q_{k_j} $(m+1 \leq j \leq n)$ has included incomplete or incorrect bibliographic data of p, such as a typo in p's author names, title, page numbers, etc). Then, the following two questions can be derived from RQ2:

- **RQ3** Can a citation database system *accurately* identify the citing publications in $Q_{correct}$ when calculating p's citation count?
- **RQ4** Can a citation database system *properly* identify the citing publications in $Q_{erroneous}$ when calculating p's citation count?

RQ3 is related to the functional correctness of the citation database system and, in theory, can be measured using the conventional evaluation metrics for information

1. In this paper, our discussions are always within the scope of the citation database system under consideration. The discussion here is under the assumption that both the cited publication p and all citing publications in Q are covered by the citation database system. While there can be citing publications not covered by the citation database, such publications are not considered in this research.

retrieval: *precision* and *recall* [16]. Given a query, let A be the set of all items retrieved by the software, $R \subseteq A$ be the set of retrieved items that are indeed relevant to the query, and R' be the set of relevant items in the database but not retrieved by the software. *Precision* is calculated as $|R| \div |A|$, and *recall* is calculated as $|R| \div (|R| + |R'|)$. In practice, however, for the citation database systems under study, it is difficult for a tester (especially an end-user tester) to measure recall because it requires the knowledge of not only retrieved records but also the records in the database not retrieved. Readers who are interested in alternative testing methods beyond precision and recall are referred to our previous work [16], where we applied MT to investigate the functional correctness of Web search engines such as Google. It is to be noted that RQ3 (which is more relevant to program correctness than robustness) is *not* the focus of the present research; we pose RQ3 in order to derive RQ4, which is more relevant to robustness.

RQ4 on its own is a significant research question in system validation that, to the best of our knowledge, has never been investigated before. RQ4 uses the phrase "*properly* identify" rather than "*accurately* identify" (as in RQ3) because, if the fault in the citing publication's reference list is minor and does not affect the identification of the cited publication, the citation should be counted towards the cited publication; if, however, the fault is serious, then it is possible that neither a computer system nor a human operator could identify the cited publication.

Observation 3. Given the sheer volume of records in modern citation databases, it is obvious that there is an oracle problem for all four research questions RQ1, RQ2, RQ3, and RQ4.

Observation 4. A negative answer to RQ4 may imply a negative answer to RQ2 and, hence, a negative answer to RQ1. Any of these negative answers would mean that we should *not* use citation counts as a proxy for research impact and that we need to avoid such practices in research assessment (for more discussions on this topic, readers are referred to the Declaration on Research Assessment [40]).

Observation 5. Answers to our research questions could help the users and stakeholders of citation database systems to better understand such systems, thereby making better use of them (including improving the citation-related scores of their own work).

The contributions of this research are summarized as follows:

- We present a *metamorphic robustness testing* approach, which tests the robustness of software systems for erroneous inputs in the absence of an oracle. We identify three MRs to test citation database systems.
- We report a surprising finding that the inclusion of hyphens in paper titles impedes citation counts, and that this is a result of the lack of robustness of the Scopus and Web of Science citation database systems in handling hyphenated paper titles — this finding is obtained through large-scale empirical studies using metamorphic robustness testing. We show that our results are valid for the entire literature as well as for individual fields such as chemistry.
- We go on to investigate the impact of hyphens in paper titles at the journal level, and report a further surprising finding that there is a strong and significant negative correlation between the *journal impact factor* (JIF) of *IEEE Transactions on Software Engineering* (TSE) and the percentage of hyphenated paper titles published in TSE (Pearson's $r = -0.688$, $p = 0.028$; Spearman's rho $= -0.636$, $p = 0.048$; 2-tailed). A similar (and more significant) finding is made for *ACM Transactions on Software Engineering and Methodology* (Pearson's $r = -0.702$, $p = 0.024$; Spearman's rho $= -0.855$, $p = 0.002$; 2-tailed). A software engineering field-wide study reveals that the higher JIF-ranked journals are publishing a lower percentage of papers with hyphenated titles.
- We provide a careful analysis of the validity of this research to avoid falling into the trap of equating correlation with causation.
- Our results challenge the common belief that citation counts are a reliable measure of the impact of papers, as they can be distorted simply by the presence of hyphens in paper titles, which is unrelated to the quality of the papers in question. Similarly, our results also challenge the validity of citation-based journal-level metrics, including the *journal impact factors*.

The rest of this paper is organized as follows: Section 2 presents some real-world examples of citation errors that motivate this research. Section 3 introduces our metamorphic relations. Section 4 provides an overview of our empirical studies. Sections 5 and 6 conduct empirical studies at the article and discipline levels, respectively. Section 7 conducts an empirical study at the journal level by looking at the *journal impact factors* in software engineering. Section 8 discusses several topics related to the validity of this research. Section 9 further presents some related work and shows that our research is fundamentally different from the field of *citation analysis*. Section 10 discusses the limitations of this work, and Section 11 concludes the paper.

2 Motivating Examples

Consider RQ4, which is an essential question about the robustness of the SUT. A common approach for assessing software robustness is to conduct fuzz testing, where synthetic (random or semi-random) test cases are generated and executed. This technique is *not* suitable for the present research because, first, such random or fuzz test cases cannot represent $Q_{erroneous}$ of RQ4. In other words, random or semi-random strings generated by a fuzzer are not representative of erroneous bibliographic data that humans can commonly create. Rather than to crash the

SUT to detect security vulnerabilities, the objective of this research is to examine the SUT's capability in handling real-life erroneous bibliographic data, most of which are unintentionally created by the citing authors. Fuzzing, therefore, is obviously not a choice. Furthermore, the SUT also does not allow the users to perform fuzz testing on its indexing/crawling sub-systems: Users can only search the citation database (in this research, Scopus or Web of Science) and cannot write it or direct it to any external file for crawling or indexing purposes. Of course, users can still perform fuzz testing on the graphical user interface (GUI) or the application programming interface (API) of the SUT, but detecting GUI or API vulnerabilities is not an objective of the present research, as our main research interest is on the robustness of citation indexing.

Therefore, the only type of testing we could perform is to issue queries to the SUT, and then collect and analyze its output. To address RQ4, we must consider what kind of error a real-world citing author or indexing software could possibly make when creating or processing bibliographic data. Let us consider the following examples:

In the first example, the *cited article* is [5] and the citing article is [41]. These two articles have been indexed by both Scopus and Web of Science. Fig. 1a shows an excerpt of the reference list of [41], where the bibliographic data "Z. Z" and "44(15):923–931, 2002" were wrong. Both Scopus and Web of Science have failed to match this citation to the cited article; in other words, at the time of writing, this cite had been omitted when Scopus and Web of Science calculate the citation count of [5]. Arguably, this observation may suggest that both these two citation databases are not robust enough in handling citation errors in bibliographies — such errors can be common in the real world [42], [43]. However, it could also be argued that it is reasonable for the computer system to omit the erroneous citation because the error shown in Fig. 1a is quite serious. In this research, therefore, we consider a type of less serious error, as shown in Fig. 1b.

Fig. 1b shows that the citing article is [44] and the cited article is [45], both of which have been indexed by Scopus and Web of Science. There are two minor problems in the data entry: The author name "T. Tse" should be "T.H. Tse," and the phrase "Fault based" should be "Fault-based." It is reasonable to expect that a robust citation database should be able to link this citation to the cited article because the typos are really minor. The Web of Science has successfully built the link; however, Scopus failed. After we wrote to Scopus to report the missing citation, they confirmed the error and corrected the citation index. This example shows that even a minor typo could cause serious citation indexing failures due to lack of robustness of the software system. Compared with the issues associated with author names (such as typing "T.H. Tse" as "T. Tse"), in this research we are more interested in the missing-hyphen error such as mistyping "Fault-based" as "Fault based." This is because the latter is a real typing error and may occur very frequently. Therefore, we decide to conduct a systematic

[7] T. Y. Chen, T. H. Tse, and Z. Z. Fault-based testing without the need of oracles. *Information and Software Technology*, 44(15):923–931, 2002.

(a) Erroneous bibliographic data in the reference list of [41]: The author name "Z. Z" should be "Z.Q. Zhou," and "44(15):923–931, 2002" should be "45(1): 1–9, 2003." Both the *Scopus* and the *Web of Science* databases have failed to match this citation to the cited paper [5].

[9] T. Y. Chen, T. Tse, and Z. Zhou. Fault based testing in the absence of an oracle. In *Proceedings of the 25th IEEE Annual International Computer Software and Applications Conference (COMPSAC 2001)*, pages 172–178. IEEE Computer Society, 2001.

(b) A minor error in the reference list of [44]: The author name "T. Tse" should be "T.H. Tse," and "Fault based" should be "Fault-based." The *Web of Science* has successfully matched this citation to the cited paper [45], but *Scopus* failed to do so. We reported the missing citation to Scopus, who then confirmed the error and corrected the citation index.

Fig. 1: Examples of citation errors in bibliographies that reveal the lack of robustness of the citation databases.

investigation into the impact of hyphens in paper titles on citation statistics. If the citation database system is not robust when dealing with missing-hyphen errors in bibliographies, then it may fail to link a citation (with the missing-hyphen error) to the cited article, which would mean that the inclusion of hyphens in paper titles could have a negative impact on the papers' citation counts generated by the system. Fortunately, compared with personal names and affiliations, researchers normally have much more freedom to decide their paper titles. This is also a reason why we decide to study the impact of paper titles instead of author names — so our research results may provide practical hints for authors to select "robust" publication titles that would avoid potential citation errors, hence improving their citation scores calculated by not-so-robust citation database systems.

3 MRs for Citation Database Systems

To provide a solution to our research questions, we must address the oracle problem. Therefore, we propose to use MT. We first specify our MRs, and then give further elaboration.

3.1 The Identified Metamorphic Relations (MRs)

To perform MT, we define the following MRs for an *ideal* citation database system:

- **$\mathbf{MR_{similar}}$.** Let P_x and P_y be two **large** sets of publications, and $cite(P_x)$ and $cite(P_y)$ be the mean citation counts per publication of P_x and P_y, respectively. Generally speaking, if P_x and P_y do not have any

systematic difference in factors related to potential impact or likely citations, then *cite*(P_x) and *cite*(P_y) should have little systematic difference.

- **MR$_{\text{older}}$.** Generally speaking, older publications should have higher citation counts than newer publications.
- **MR$_{\text{aging}}$.** Let P_x and P_y be two **large** sets of publications without systematic differences in factors related to potential impacts or likely citations. When the publications in P_x and P_y become older, their mean citation counts *cite*(P_x) and *cite*(P_y) should increase at a similar rate. The subscript "aging" is used in the sense of maturing or ripening.

3.2 An Example

Suppose we define P_1 as the set of all papers that satisfy all three criteria stated as follows:

- they are indexed in the citation database under consideration,
- they are in the field of software engineering, and
- their paper titles include a hyphen (for example, see the paper title of [46]).

Next, suppose we define P_2 as the set of all papers that satisfy all three criteria stated as follows:

- they are indexed in the citation database under consideration,
- they are in the field of software engineering, and
- their paper titles do not include any hyphen.

For software engineering papers, it is reasonable to believe that the inclusion of a hyphen in paper titles is not a factor related to potential impact or likely citations of the paper — a hyphen is irrelevant to the paper quality, significance, innovation, readability, or accessibility. According to MR$_{\text{similar}}$, therefore, the mean citation counts of P_1 and P_2 should have little systematic difference.

Furthermore, according to MR$_{\text{older}}$, the older publications in P_1 and P_2 should generally have higher citation counts than the newer publications in these two sets and, according to MR$_{\text{aging}}$, the mean citation counts of older publications in P_1 and P_2, as compared with those of newer publications, should increase at a similar rate.

3.3 Validity

It should be noted that all the MRs and discussions presented in Section 3 are under the assumption that the citation database system is *ideal* and that P_1 and P_2 are *large enough*. Further discussions on the validity of this research are presented in Section 6 and Section 8.

4 Overview of Empirical Studies

In general, citation statistics can be divided into three tiers: the article level, journal level, and author level [47]. Some researchers may refer to articles as papers, refer to journals as sources, or refer to authors as individual scientists, but the classification levels are largely equivalent [48]. Citation counts are generally recognized to be a reliable metric for the evaluation of individual papers [49], [50]. They are also used to compute the *journal impact factor* [51], which is the evaluation metric at the journal level, and the h index [52], which is the evaluation metric at the author level.

Our empirical studies are conducted at the article and journal levels. Furthermore, to enhance the validity of this research, we also conduct empirical studies at the discipline level (looking at groups of journals within individual research fields).

5 Empirical Study at Article Level

Letchford et al. [53] conducted a large scale study on the $20,000$ most cited papers indexed in the Scopus citation database system every year from 2007 to 2013, giving a total of $140,000$ articles. They found that papers with shorter titles tended to be cited more than those with longer titles. Their results were also reported in Science [54] and Nature [55].

In our empirical study, we find that it is actually *the number of hyphens in the title, not the title length*, that serves as the more dominating factor for citation counts. This impact of hyphens in paper titles on citation statistics is discovered by *metamorphic robustness testing* in combination with a *fault-based testing* strategy, targeting the system robustness problems in handling missing-hyphen citation errors. More specifically, we conduct metamorphic testing of Scopus and Web of Science against the metamorphic relations MR$_{\text{similar}}$, MR$_{\text{older}}$, and MR$_{\text{aging}}$.

5.1 Setup of Empirical Study

For ease of comparison, we use the same Scopus dataset as Letchford et al. [53], downloaded from Dryad Digital Repository [56]. Scopus [38], owned by Elsevier, is the largest citation database of peer-reviewed literature. According to the latest statistics, it covered over 69 million core records and 1.4 billion cited references. It covered more than 5,000 international publishers, 21,950 journals, 100,000 conferences, and 150,000 books. Approximately 3 million new items are added to its database each year. The Scopus database provides the data behind the Times Higher Education World University Rankings [57] and the QS World University Rankings [58]. We supplement the first dataset with further citation statistics from the Web of Science because of its world recognition. The Web of Science [39] is the citation database system with the longest history, focusing on depth and quality [59]. It was formerly known as the ISI Web of Knowledge and is currently under Clarivate Analytics. It is the provider of *Journal Citation Reports*, in which *journal impact factors* are computed. It includes a core collection and various external supplementary citation databases. We have captured the 5,000 most cited papers indexed in all of its databases every year between 2007 and 2013 (to be consistent with the years of publications collected

by Letchford et al. [53]), which is achieved through the Web of Science platform by selecting "All Databases" for "Advanced Search" with a query "PY = *yyyy*," where *yyyy* denotes the publication year such as 2013. We have deleted duplications and faulty entries due to missing dates or erroneous dates. The final version of the second dataset comprises a total of 34,982 papers. Table 1 shows the descriptive statistics of the two datasets for the empirical study reported in Section 5 (the last row of Table 1 is explained in Section 5.2.2).

We would like to add that the Web of Science dataset should be used with caution. We find (occasional) inaccuracies in the paper titles downloaded from the Web of Science website, but it is infeasible for us to manually verify and correct each of the 34,982 records. These inaccuracies may create some noise in data analysis. The results derived from the Web of Science dataset should, therefore, be used only as a supplement for the purpose of confirming the finding from the main data source, namely, the Scopus dataset.

5.2 Violation of Metamorphic Relation $MR_{similar}$

In this section, we inspect the test results against $MR_{similar}$, and perform in-depth analyses on the resulting datasets.

5.2.1 Hyphens in paper titles "impede" citations

Fig. 2a shows how the mean citation count per article in the Scopus dataset varies according to the number of hyphens in the paper title. Let P_0, P_1, ..., P_7, and $P_{>7}$ be the sets of publications with 0, 1, ..., 7, and > 7 hyphens in each paper title, respectively. They are represented by bars from left to right in Fig. 2a with colors gradually changing from dark to light. They are generally large sets that do not have any systematic difference in factors related to potential impacts or likely citations. The only difference is the number of hyphens per paper title. According to metamorphic relation $MR_{similar}$, we expect their mean citation counts per article $cite(P_0)$, $cite(P_1)$, ..., $cite(P_7)$, and $cite(P_{>7})$ to have little systematic difference. However, Fig. 2a clearly shows a systematic pattern where the mean citation decreases when the number of hyphens increases. This observation means that $MR_{similar}$ is violated. In other words, the mean citation count of an article calculated by the SUT is adversely affected by the number of hyphens in the title. Fig. 2b also shows a violation of $MR_{similar}$ in the Web of Science dataset. These findings are alarming because the citation counts are affected by a minor symbol totally unrelated to the quality of the papers in question.

The above violations of $MR_{similar}$ may suggest that the systems do not consider robustness, which refers to the capability to cope with erroneous inputs. A plausible reason for the erroneous inputs is that when authors cite a paper with hyphens in the title, they may overlook some of the hyphens. As a result, citation databases may not be able to match it with the original paper and, hence, the original's citation count is not increased. Such an example is given in Fig. 1b. The citation database systems should be sufficiently robust to cater for such common mistakes.

These mistakes are reinforced by the previous findings by Simkin and Roychowdhury, who suggested that many authors do not verify the articles that they cite when copying and pasting entries in an existing reference list to their own [60]. They also reported that many authors not only cite a paper but also copy a few references from that paper [61]. Our results are consistent with their findings because hyphens in paper titles may lead to mistakes in both first- and second-hand citations. In addition, first-hand citation mistakes will also be propagated to second-hand citations. Furthermore, if an author makes a citation mistake, he/she is likely to repeat the same mistake for several years across multiple articles, as a result of reusing the same bibliography file (such as a BibTeX file) where bibliographic references are stored.

A real-world example showing the propagation of a citation mistake across multiple papers can be produced by searching at http://www.google.com with the following query:

"fault-based testing without the need of oracles" "Information and Software Technology" "44" "15" "923" "931" "2002"

or by directly visiting the following link:
https://www.google.com/search?source=hp&ei=3SR5W-_cHpuC-QaivZ6IBg&q=%22fault-based+testing+without+the+need+of+oracles%22+%22Information+and+Software+Technology%22+%2244%22+%2215%22+%22923%22+%22931%22+%222002%22&oq=%22fault-based+testing+without+the+need+of+oracles%22+%22Information+and+Software+Technology%22+%2244%22+%2215%22+%22923%22+%22931%22+%222002%22&gs_l=psy-ab.3...2247.2247.0.2749.1.1.0.0.0.0.0.0..0.0....0...1.1.64.psy-ab..1.0.0....0.MA_ba4RmjFQ

At the time of writing, the Google search engine had returned about 10 different articles (published in different years and venues) whose reference lists contain exactly the same citation mistake "**Fault-based testing without the need of oracles.** ***Information and Software Technology,*** **44(15):923–931, 2002**" (the same mistake as in Fig. 1a). These search results clearly show that a single citation mistake "44(15):923–931, 2002" (which should be "45(1): 1–9, 2003" [5]) had been propagated to different citing articles.

Further discussions on the validity of our analysis results are presented in Sections 6 and 8.

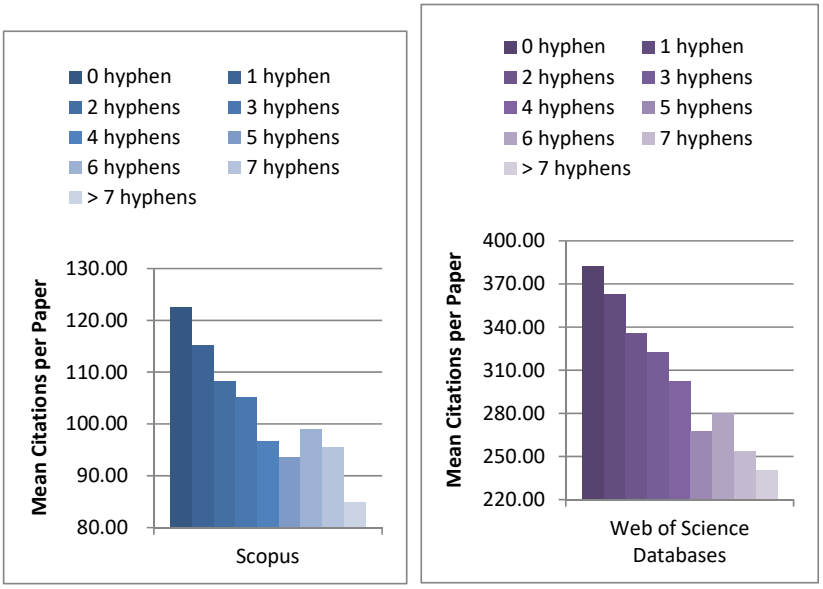
5.2.2 A progressive tax: the higher your (citation) income level, the higher a (hyphen) tax rate you pay

In the real world, a *progressive tax* is a tax in which the tax rate increases as the income increases. Here, the word *progressive* means "increasing in rate as the base increases"[2]. If we consider a cited paper to be a person, the paper's citation count calculated by a citation database system to be the person's after-tax income, and the paper's lost citations to be a *hyphen tax*, then we find that **the hyphen tax is a progressive tax**.

2. https://www.merriam-webster.com/dictionary/progressive

TABLE 1: Descriptive statistics of datasets of Section 5 (Scopus and Web of Science (WoS)).

No. of hyphens	0	1	2	3	4	5	6	7	> 7	Overall
No. of papers (Scopus)	67,659	43,537	18,895	6,630	2,144	696	247	89	103	140,000
Percentage (Scopus)	48.33%	31.10%	13.50%	4.74%	1.53%	0.50%	0.18%	0.06%	0.07%	100.00%
No. of papers (WoS)	19,298	10,121	3,808	1,209	375	97	51	14	9	34,982
Percentage (WoS)	55.17%	28.93%	10.89%	3.46%	1.07%	0.28%	0.15%	0.04%	0.03%	100.00%
(WoS Percentage)÷(Scopus Percentage)	114.15%	93.02%	80.67%	73.00%	69.93%	56.00%	83.33%	66.67%	42.86%	100.00%



(a) Scopus dataset. (b) Web of Science dataset.

Fig. 2: How are citation counts affected by numbers of hyphens in paper titles?

First, consider Fig. 2. The statistics shows that, on average, papers in the Web of Science dataset have much higher citation counts than those in the Scopus dataset: The mean citations of the former are all above 220, whereas those of the latter are all below 130. This is because the Scopus dataset was collected in an earlier year and included the top $20,000$ most cited papers per year, whereas the Web of Science dataset was collected in a later year and only included the top $5,000$ papers per year. Therefore, although these two datasets were collected from different sources, papers in the Web of Science dataset can be generally considered to have higher citations. Next, consider Table 1: The second column shows that the Scopus and the Web of Science (WoS) datasets include 48.33% and 55.17% 0-hyphen papers, respectively, and hence (WoS Percentage)÷(Scopus Percentage) $= (55.17\% \div 48.33\%) = 114.15\%$, as shown in the last row. This means that the WoS dataset includes a higher percentage of 0-hyphen papers than the Scopus dataset. The respective values of (WoS Percentage)÷(Scopus Percentage) for the 1- and 2-hyphen groups are 93.02%, and 80.67%, respectively, which means that the WoS dataset includes a smaller percentage of 1-hyphen papers, and an even smaller percentage of 2-hyphen papers, than the Scopus dataset. Generally speaking, the last row of Table 1 shows a descending trend of the ratio (decreases from 114.15% to 42.86% when the number of hyphens increases from 0 to > 7. Because all hyphen groups of the WoS dataset have higher mean citations than those of the Scopus dataset, we might be able to hypothesize that the more the citations papers receive, the stronger the impact of hyphens on the citations.

To investigate the above hypothesis, we have further analyzed the Scopus dataset. We define $P_{i,j}$ as the ratio of "the number of j-hyphen papers that have a citation count greater than i" to "the total number of j-hyphen papers," where $i = 20$, 40, 60, 80, 100, 120, 140, 160, 180, and 200; and $j = 0$, 1, …, 7, and "> 7." For example, the Scopus dataset contains a total of $67,659$ papers whose titles do not contain any hyphen, of which $63,907$ papers have a citation count greater than 20. Therefore, $P_{20,0} = 63907 \div 67659 = 94.45\%$ (which means that 94.45% 0-hyphen papers in the Scopus dataset have a citation count greater than 20). We find that, for all values of i (the citation threshold), $P_{i,0}$ is always the largest among all hyphen groups, which means that the 0-hyphen group always contains the largest percentage of papers whose citation counts are greater than the given threshold. We therefore normalize each $P_{i,j}$ value by calculating its ratio to $P_{i,0}$. That is, we calculate a normalized value $R_{i,j}$, defined as $R_{i,j} = P_{i,j} \div P_{i,0}$, where $i = 20$, 40, 60, 80, 100, 120, 140, 160, 180, and 200; and $j = 0$, 1, …, 7, and "> 7." When $j = 0$, $R_{i,j} = 100\%$; when $j > 0$, all $R_{i,j}$ values are smaller than 100%. A higher value of

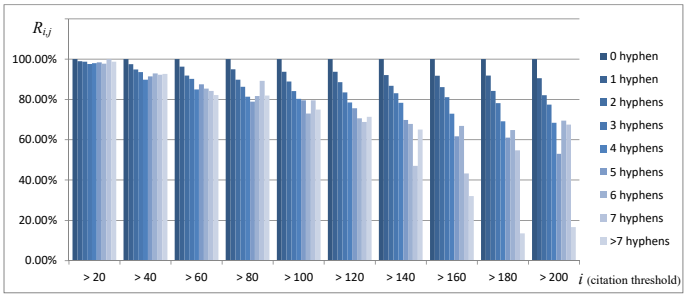
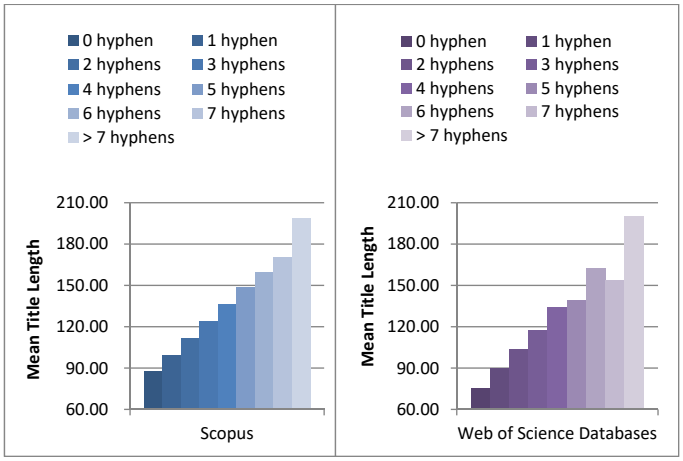


Fig. 3: The distribution of $R_{i,j}$, grouped by i (the citation threshold), and then by j (the number of hyphens). The figure shows that, in general, the higher the citation count a paper has, the stronger the impact of hyphens on the citation count.



(a) Scopus dataset. (b) Web of Science dataset.

Fig. 4: How does mean title length vary with no. of hyphens in the titles?

$R_{i,j}$ means a relatively higher percentage of papers that have citation counts above the threshold i. Fig. 3 shows the distribution of $R_{i,j}$, which indicates that the impact of hyphens (measured by the difference of $R_{i,j}$ values between various hyphen groups) is greater for highly cited articles. For example, Fig. 3 shows that, when the citation threshold is "$>$ 20," the number of hyphens in paper titles has had little impact on the citations (that is, the differences of $R_{i,j}$ values between various hyphen groups are small); but when the citation threshold increases, the slope becomes deeper and deeper. When the citation threshold reaches "$>$ 180," the differences between the various hyphen groups become very large.

5.2.3 *Greater of two evils: Title length or hyphens in title?*

We find in Section 5.2.1 that the mean citation count of an article is adversely affected by the number of hyphens in the title. We also recall the results in Letchford et al. [53] that the mean citation count is adversely affected by the title length. Of course, longer titles are more likely to include more hyphens. This is confirmed in Figs. 4a and 4b, which show that title length and the number of hyphens in the title are strongly correlated. A question naturally arises: Which is the more dominating factor for the reduced citation count — the title length or the hyphens in the title?

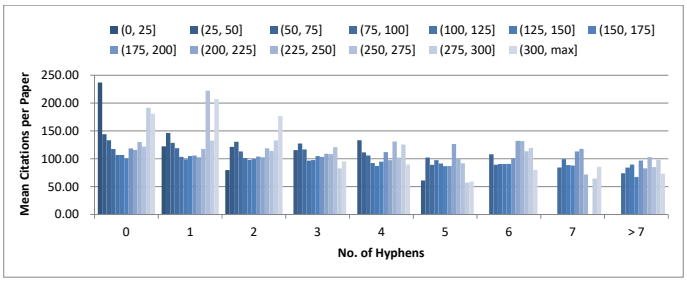
We have conducted further analyses to answer this question. We divide the collected statistics into nine hyphen groups (with 0, 1, ..., 7, and $>$ 7 hyphens in the paper titles). Each group is further divided into subgroups according to the title lengths of the papers. Fig. 5a shows how the citation counts are affected by the title lengths for each and every hyphen group based on the Scopus dataset, where the step length is set to 25 characters and the last subgroup covers all the papers with more than 300 characters in their titles. We observe that for papers with the same number of hyphens in the titles, there is no systematic trend between the title length and the mean citation count per paper. In other words, when the number of hyphens in paper titles is fixed, the title length does *not* have an obvious impact on citation count.

Next, we regroup all the papers according to the title lengths and then according to the numbers of hyphens in the titles. The results are depicted in Fig. 5b, where a general trend can be observed: For papers in the same title length group, in general, the mean citation count per paper is adversely affected by the number of hyphens in the title.

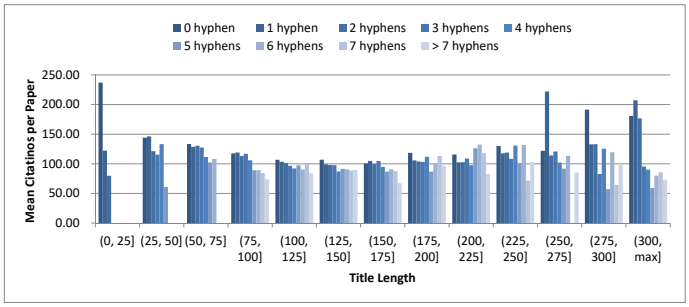
Figs. 5c and 5d show the corresponding plots for the Web of Science dataset. As this dataset is much smaller than the Scopus dataset, a slightly larger step length of 35 is used. A similar observation can be made: The title length does not have an impact on the mean citations once the number of hyphens in the title is fixed, whereas in general, the number of hyphens in the title adversely affects the mean citation for papers in the same title length group. It should also be noted that inside each and every title length group in Fig. 5b and Fig. 5d, the left most bar (corresponding to papers without hyphens in the titles) is consistently longer than the right most bar (corresponding to papers with more than seven hyphens in the titles).

We have also tried different step lengths and observed a similar pattern in the resulting charts.

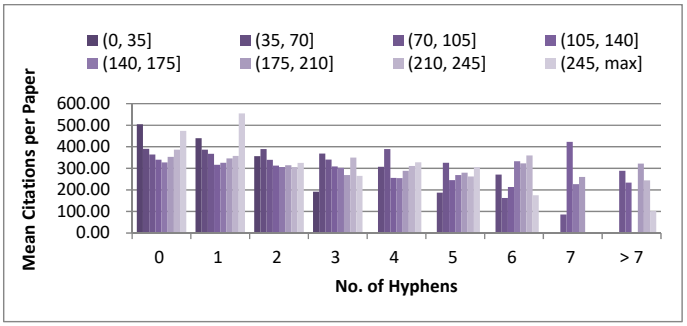
Letchford et al. [53] proposed three possible reasons for the adverse effect of title lengths: "One potential explanation is that high-impact journals might restrict the length of their papers' titles. Similarly, incremental research might be published under longer titles in less prestigious journals. A third possible explanation is that shorter titles may be easier to understand, enabling wider readership and increasing the influence of a paper." Unfortunately, these quality aspects of the papers and their publication venues cannot compete with a more dominating factor completely unrelated to excellence, namely, the number of hyphens in the paper title.



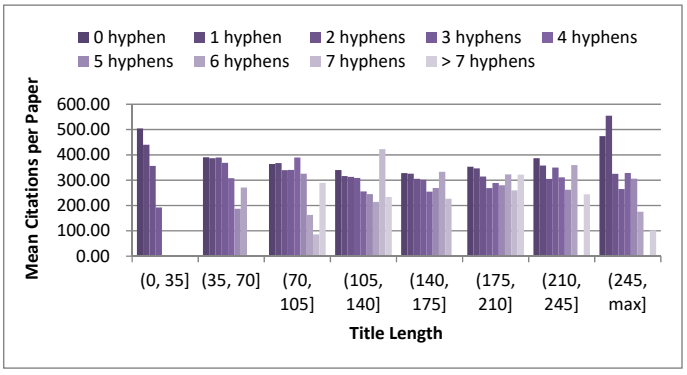
(a) How are citation counts affected by title lengths for various nos. of hyphens in the titles? (Scopus dataset.)



(b) How are citation counts affected by nos. of hyphens in the titles for various title lengths? (Scopus dataset.)

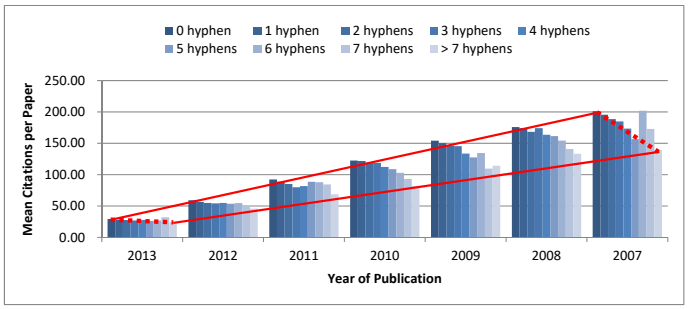


(c) How are citation counts affected by title lengths for various nos. of hyphens in the titles? (Web of Science dataset.)

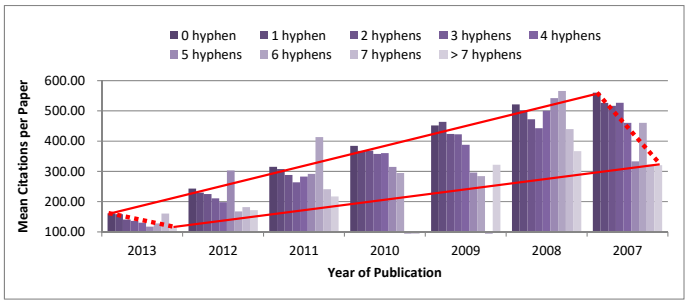


(d) How are citation counts affected by nos. of hyphens in the titles for various title lengths? (Web of Science dataset.)

Fig. 5: Nos. of hyphens in paper titles versus title lengths: Which factor dominates?



(a) Scopus dataset.



(b) Web of Science dataset.

Fig. 6: How are citation counts affected by nos. of hyphens in paper titles for various years of publication?

5.3 Violation of MR_{older} and MR_{aging}: Hyphens in paper titles "impede" impact of aging

Fig. 6a shows how the mean citation count per article in the Scopus dataset varies in relation to the number of hyphens in the paper title for various years of publication. Let P_0 be the set of publications without hyphens in the paper titles and $P_{>7}$ be the set of publications with more than seven hyphens in each title. The upper solid red line shows the escalation trend in mean citation count per article in P_0 while the lower solid red line shows the escalation trend in $P_{>7}$. Both P_0 and $P_{>7}$ are large sets that do not have any systematic difference in factors associated with potential impacts or plausible citations. The only difference is the number of hyphens in each paper title. According to metamorphic relation MR_{aging}, we expect the escalation trends in their mean citation counts per article $cite(P_0)$ and $cite(P_{>7})$ to have little systematic difference. We find from Fig. 6a, however, that $cite(P_{>7})$ increases at an observably lower rate than $cite(P_0)$ as the papers become older in the period from 2013 to 2007. This evidently violates MR_{aging}. In other words, articles with more hyphens in the titles have a less significant increase in citations over the years under study. Similarly, Fig. 6b shows a violation of MR_{aging} in the Web of Science dataset.

It is observed that MR_{older} is also violated: In Fig. 6a, several bars on the right side of the 2007 group are shorter than the left most bar of the 2008 group; similarly, the right most bars of the 2008, 2009, and 2010 groups are shorter than the left most bars of the 2009, 2010, and 2011

groups, respectively. A similar observation is made with the Web of Science dataset shown in Fig. 6b.

The dashed red lines on the left of Fig. 6a and Fig. 6b show the escalation trends in the mean citation counts of the papers in 2013 as the number of hyphens in paper titles increases from zero to more than seven. The dashed red lines on the right of the figures show the trends of the mean citation counts of the papers in 2007. They indicate that more aged articles have a more significant decrease in mean citation counts as the number of hyphens in the paper titles increases.

All these violations of MR_{aging} and MR_{older} arouse serious concerns because the escalation trends in citation counts of aging articles are reduced by a small symbol completely unrelated to paper quality. This may suggest that the citation database systems do not tackle robustness, or lack the ability to deal with incorrect data. As explained earlier, when authors refer to a paper, they may miss out some of the hyphens in the title. The systems cannot locate the original paper and, therefore, the citation count is adversely affected. The systems must be robust enough to deal with such mistakes. These mistakes are again reinforced by Simkin and Roychowdhury [60], who suggested that mistakes in paper titles appear more often in an older article than a newer one because authors simply copy the entries in a previous reference list to the present list.

6 Empirical Study at Discipline Level

A threat to the validity of our analysis results is that there can be large differences between fields in citation practices, resulting in publications in some fields having systematically higher citation counts than publications in other fields. It could be argued that papers in a field such as chemistry (where paper titles often carry hyphens that are standard chemical nomenclature) might only receive relatively limited numbers of citations, which could give rise to a spurious negative correlation between hyphens and citation counts.

We have therefore conducted a focused study on the chemistry journals in the Scopus dataset of $140,000$ entries. The metamorphic relation used in this study is $MR_{similar}$. We extract the records of all the *journals* whose titles contain the string "Chem," which is the search key for "Chemistry," "Chemical," "Chemotherapy," and so on. The results are shown in Fig. 7f, which indicates that hyphens adversely affect citation counts of papers even when we limit the study only to the discipline of chemistry. Because of the significantly reduced sample size as compared with the original dataset of $140,000$ records, we use six instead of nine hyphen groups to achieve more statistically meaningful results.

We have further investigated the citation data in other research areas: "Bio," "Comput," "Math," "Medic" and "Physic," all of which indicate a clear pattern that hyphens in paper titles adversely affect citation counts in the respective field. The results are shown in Figs. 7a to 7e.

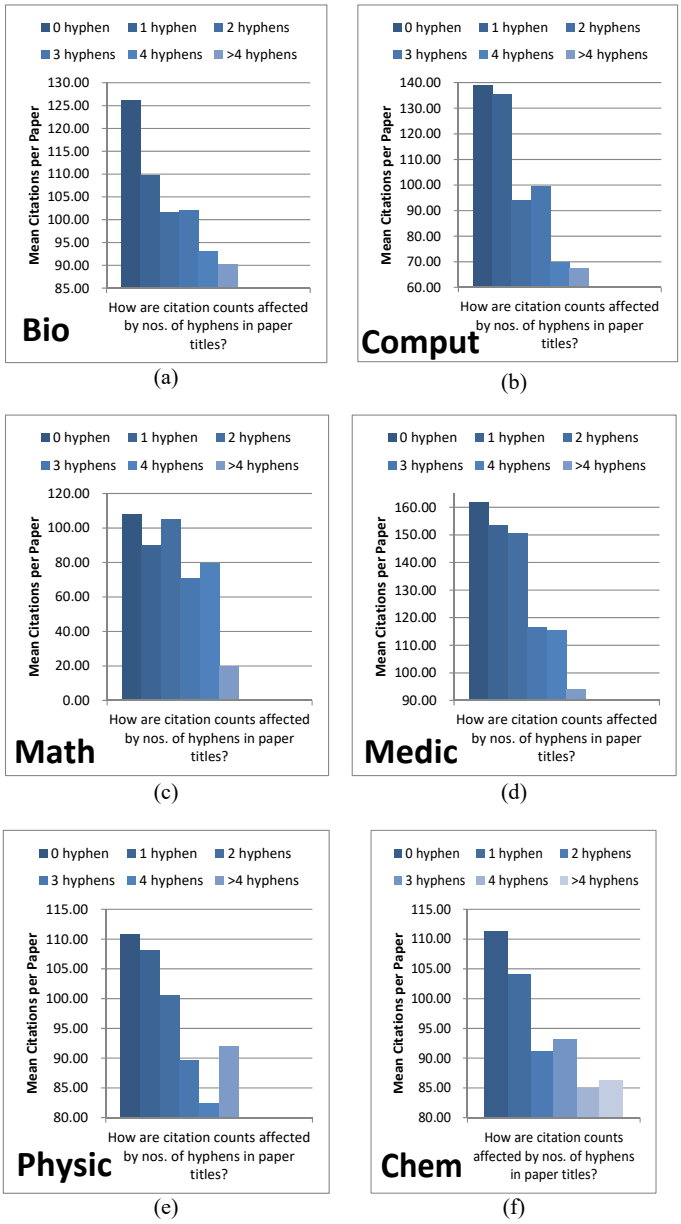


Fig. 7: How are citation counts affected by nos. of hyphens in paper titles for various research areas? (Scopus dataset.)

7 Empirical Study at Journal Level

To build on our findings at the article and discipline levels, we have further investigated the impact of hyphens in paper titles on *journal impact factors* (JIFs) within the field of *software engineering*, by collecting and analyzing a new set of journal-level data from the Web of Science database. In this section, we first explain the concept of *journal impact factors*, and then conduct case studies using "the two flagship software engineering journals" [62, p. 2]: *IEEE Transactions on Software Engineering* and *ACM Transactions on Software Engineering and Methodology*. After these two case studies, we conduct a larger scale empirical study by aggregating journal-level data from the Web of Science database involving all 106 journals in the category "COMPUTER SCIENCE, SOFTWARE

ENGINEERING" (as listed in the 2016 *Journal Citation Reports* (JCR) — the 2016 edition of JCR was the newest edition when we completed our data collection in March 2018). All these studies suggest that hyphens in paper titles have a negative impact on the *journal impact factors*.

The metamorphic relation used at the journal-level study is $\text{MR}_{\text{similar}}$.

7.1 Journal Impact Factor (JIF)

The JIF is a metric for determining citation frequency of an academic journal [51]. It is frequently used as the primary parameter for the relative importance of a journal within its field. The JIF is calculated and published annually by Clarivate Analytics (previously known as the Institute for Scientific Information (ISI)) in their *Journal Citation Reports*. The JIF for a specific year x is calculated as follows [3]:

$$\frac{\text{Year } x \text{ citation count to articles published in year } x-1 \text{ or year } x-2}{\text{Count of citable articles published in year } x-1 \text{ or year } x-2}.$$

The time-sensitive data, that is the citations gained in the specific 12 months, is not publicly accessible from Clarivate Analytics or the Web of Science databases, so the JIFs are calculated and reported by the organization themselves.

7.2 Case Study of IEEE Transactions on Software Engineering

We have taken the reported JIF of *IEEE Transactions on Software Engineering* (TSE) for each year from 2007 to 2016 (hence obtaining ten years' JIFs from the *Journal Citation Reports*). These JIFs correspond to 11 years' papers published in TSE (from 2005 to 2015): The 2007 JIF corresponds to the 2005 and 2006 papers; the 2008 JIF corresponds to the 2006 and 2007 papers; and so on. We have downloaded all the article data from the Web of Science database. There is a total of 723 articles (titles) from TSE in the span of 2005 to 2015.

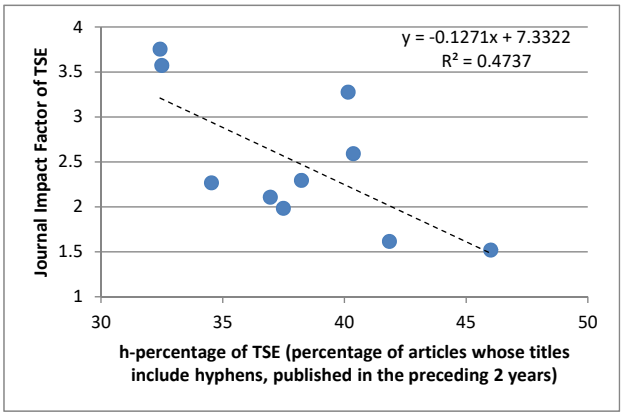
For each of the ten JIF years, we calculate the proportion of hyphenated paper titles of the preceding two years. More specifically, we define *h-percentage* of year x as $A \div B$, where A is the number of papers whose title contains at least one hyphen, published in TSE in year $x-1$ or year $x-2$, and B is the total number of papers published in TSE in year $x-1$ or year $x-2$. We thus obtain ten *h-percentage* scores corresponding to the ten JIFs, as shown in Fig. 8a.

Fig. 8a shows that TSE achieved the highest JIF (3.750) in 2009 — also in this year, TSE's *h-percentage* was the smallest (32.432), and that TSE achieved the second highest JIF (3.569) in 2008 — also in this year, TSE's *h-percentage* was the second smallest (32.500). Furthermore, Fig. 8a shows that TSE had the lowest JIF (1.516) in 2015 — in this year, TSE's *h-percentage* was the highest (46.012), and that TSE had the second lowest JIF (1.614) in 2014 — in this year, TSE's *h-percentage* was the second highest (41.848). Hence, there appears to be a negative correlation between TSE's JIF and *h-percentage*.

3. https://clarivate.com/essays/impact-factor/

Year	JIF	h-percentage (%)
2016	3.272	40.157
2015	1.516	46.012
2014	1.614	41.848
2013	2.292	38.235
2012	2.588	40.367
2011	1.980	37.500
2010	2.265	34.545
2009	3.750	32.432
2008	3.569	32.500
2007	2.105	36.957

(a) Ten years' JIFs and *h-percentages* of TSE. **h-percentage**: percentage of articles with hyphenated titles published in the preceding 2 years.



(b) There is a strong and significant negative correlation between TSE's *journal impact factor* and the percentage of papers with hyphenated titles published in the preceding 2 years: Pearson correlation $= -0.688$, $p = 0.028$; Spearman's *rho* $= -0.636$, $p = 0.048$. The correlations are significant at the 0.05 level (2-tailed).

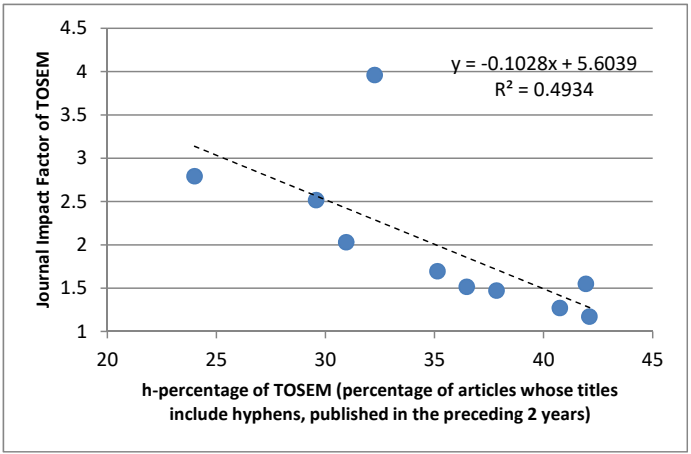
Fig. 8: How is the *journal impact factor* (JIF) of *IEEE Transactions on Software Engineering* (TSE) affected by hyphens in paper titles?

Fig. 8b shows the scatter chart of Fig. 8a. Both the Pearson's correlation coefficient (-0.688, $p = 0.028$) and the non-parametric statistic, Spearman's correlation coefficient (-0.636, $p = 0.048$) confirm a strong and significant negative correlation between TSE's *journal impact factor* and *h-percentage*.

Fig. 8b shows that the correlation coefficient squared (that is, *coefficient of determination*, R^2) has a value of 0.4737, indicating that the *h-percentage* accounts for (or "explains" [63]) 47.37% of the variation in TSE's *journal impact factors*, which is quite surprising as it is all about a simple hyphen in paper titles.

Year	JIF	h-percentage (%)
2016	2.516	29.577
2015	1.513	36.471
2014	1.170	42.105
2013	1.472	37.838
2012	1.548	41.935
2011	1.269	40.741
2010	1.694	35.135
2009	2.029	30.952
2008	3.958	32.258
2007	2.792	24.000

(a) Ten years' JIFs and *h-percentages* of TOSEM. **h-percentage**: percentage of articles with hyphenated titles published in the preceding 2 years.



(b) There is a strong and significant negative correlation between TOSEM's *journal impact factor* and the percentage of papers with hyphenated titles published in the preceding 2 years: Pearson correlation $= -0.702$, $p = 0.024$; Spearman's *rho* $= -0.855$, $p = 0.002$. The correlations are significant at the 0.05 level (2-tailed).

Fig. 9: How is the *journal impact factor* (JIF) of *ACM Transactions on Software Engineering and Methodology* (TOSEM) affected by hyphens in paper titles?

7.3 Case Study of ACM Transactions on Software Engineering and Methodology

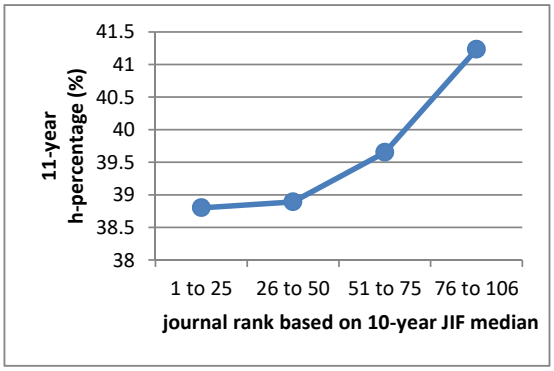
We have conducted a further case study using *ACM Transactions on Software Engineering and Methodology* (TOSEM). The analysis procedure is identical to that for TSE, and the results are presented in Fig. 9.

Fig. 9 shows that the TOSEM results are similar to (actually, stronger than) those of TSE: Both the Pearson's correlation coefficient (-0.702, $p = 0.024$) and the Spearman's correlation coefficient (-0.855, $p = 0.002$) confirm a strong and significant negative correlation between TOSEM's *journal impact factor* and *h-percentage*.

Fig. 9b shows that $R^2 = 0.4934$, indicating that the *h-percentage* explains nearly fifty percent of the variation in TOSEM's *journal impact factors*.

Journal rank based on 10-year JIF median	11-year h-percentage (%)	Total papers (>15 cites)
1 to 25	38.799	7,111
26 to 50	38.891	2,759
51 to 75	39.650	2,227
76 to 106	41.232	1,169

(a) 10-year JIF median versus 11-year h-percentage of papers with >15 cites (all paper data were collected from the Web of Science database during the period of February 28 to March 13, 2018). (**11-year h-percentage**: the percentage of hyphenated titles over the 11-year span of 2005 to 2015).



(b) Trend: The lower the journal JIF rank, the higher the 11-year h-percentage.

Fig. 10: Journals with higher JIFs contain a smaller percentage of hyphenated paper titles.

7.4 Software Engineering Field-Wide Study

The 106 *software engineering* journals as listed in the 2016 Journal Citation Reports are used for the field-wide study, as this was the newest edition of JCR at the time of data collection. Using this list of 106 software engineering journals, the JIFs for each year dating from 2007 to 2016 are obtained and the 10-year *median* of JIFs for this period per journal is calculated. Next, the journals are ranked on this ten-year JIF median, and then grouped as follows: rank $1-25$, $26-50$, $51-75$, and $76-106$ based on the JIF median (highest to lowest). We also collect the article data of these journals — these articles were published in the 11-year span of 2005 to 2015, corresponding to the ten years of JIFs from 2007 to 2016. A total of $82,048$ papers' records have been collected from the Web of Science database. From this dataset, we have deleted all papers whose citation counts are smaller than or equal to 15. This is to filter out the noise caused by lowly cited papers because, as analyzed in Section 5.2.2, the impact of hyphens on citation counts is much more serious for highly cited papers, and that the minimum citation count recorded in the datasets of Section 5 is 16. Hence, we obtain a total of $13,266$ articles with > 15 cites[4]. This

4. We did not apply this treatment in the case studies conducted in Sections 7.2 and 7.3 to avoid a too small sample size.

article data is pooled for each of the four groups to generate the percentage of titles with at least one hyphen (over the 11 years, which is called the *11-year h-percentage*). The results are shown in Figs. 10a and 10b.

Fig. 10b (which corresponds to the data of Fig. 10a) reveals a clear trend that the higher JIF-ranked journals are publishing a lower percentage of papers with hyphenated titles.

8 Validity of this Research

We discuss the validity of this research from the following perspectives: scope of research, correlation and causation, the selection of hyphens, and the difference between robustness in GUI and robustness in citation indexing. At the end of this section, we present a further case study in which a *differential testing* strategy is used to explore the ground truth.

8.1 Scope of research

The main data source of our study at the article level and the discipline level (Sections 5 and 6) is the Scopus dataset of $140,000$ articles published by Letchford et al. [53]. This dataset is generally considered to be large for the analysis of citation statistics [55]. To enhance the validity of this research, we have systematically collected an additional dataset ourselves from a different source, the Web of Science. The analyses of these two datasets at the article level show surprisingly similar patterns, revealing a lack of robustness in both the Scopus and the Web of Science database systems. Nevertheless, the lack of robustness in these two systems cannot be extrapolated to other database systems that have not been investigated in this research.

Similarly, at the journal level, although our case studies of TSE (Section 7.2) and TOSEM (Section 7.3) show consistent results, the findings should not be extrapolated to other journals that have not been investigated (that is, we should not assume that a strong negative correlation between hyphens in paper titles and *journal impact factors* can be found in each and every journal other than TSE and TOSEM). One of the reasons for the strong and significant negative correlations found in our case studies could be that TSE and TOSEM are generally regarded as the most prestigious journals in the software engineering discipline, and therefore could have received a higher number of second-hand citations, which propagate citation errors at a faster speed than first-hand citations (as analyzed in Section 5.2). Further investigation into this phenomenon would involve psychology studies and, hence, is beyond the scope of the present research. In any case, the software engineering field-wide study conducted in Section 7.4 suggests that hyphens in paper titles have a wide impact on *journal impact factors*, at least within the software engineering discipline.

8.2 Correlation should not be equated with causation

It should be noted that, in scientific research, there is a well-known trap of equating correlation with causation. To address this threat and make valid conclusions, in this research we have conducted in-depth data analyses from several different perspectives, by applying the following control variables: (1) citation database (Scopus and Web of Science); (2) minimum citation count of a paper, that is, the citation threshold (Fig. 3); (3) title length (Fig. 5); (4) year of publication (Fig. 6); and (5) field of research (Fig. 7). Furthermore, we have conducted two case studies with two specific software engineering journals (Figs. 8 and 9) as well as a software engineering field-wide study (Fig. 10). We have thus provided strong evidence supporting the conclusion that hyphens in paper titles are indeed the cause for the decreased citation counts, and that the root cause for this is the lack of robustness of the Scopus and Web of Science citation database systems in dealing with the missing-hyphen citation errors.

A plausible alternative interpretation could be that "hyphens in paper titles may affect readability and, therefore, could result in fewer citations." We wish to point out that this argument is invalid: First, Section 7.84: "Hyphens and readability" of *The Chicago Manual of Style*[5] clearly states that:

> A hyphen can make for easier reading by showing structure and, often, pronunciation. Words that might otherwise be misread, such as *re-creation* or *co-op*, should be hyphenated. Hyphens can also eliminate ambiguity. For example, the hyphen in *much-needed clothing* shows that the clothing is greatly needed rather than abundant and needed.

Second, even if hyphens in paper titles did adversely affect readability and, hence, citations, this could not explain the phenomena shown in Fig. 3 (the impact of hyphens in paper titles is more serious on the citations of more highly cited articles) and Fig. 6 (violation of MR_{aging}). In contrast, all these phenomena are well explained by the propagation of citation errors and the lack of system robustness in dealing with these mistakes: Highly cited or more aged publications are more likely to receive second-hand (and third-hand, and so on) citations, which amplify citation errors because (1) if the bibliographic data in the original reference list are erroneous, they will remain erroneous or become more erroneous when moved to the copying citer's reference list; (2) if the bibliographic data in the original reference list are correct, errors could still be introduced when these data are moved to the copying citer's reference list. When the second-hand citation is copied by another citer (resulting in third-hand citation, and so on [64]), citation errors will be further amplified — hence, a greater information distortion rate after each round of copying the bibliographic data.

5. https://www.chicagomanualofstyle.org, accessed on August 31, 2018.

8.3 Why hyphens are selected?

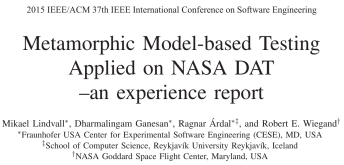
It may be argued that it is not justified to compare hyphens in paper titles that are standard chemical nomenclature, with hyphens that have been voluntarily inserted.

In fact, the *ambiguity* of the character '-' is one of the main reasons why we have selected this very character to test the *robustness* of the citation database systems. In the reference lists of citing articles, and in the citation databases, the plain text character '-' (ASCII code 45) could represent any of the following symbols: (1) hyphen, (2) subtraction or negative sign, (3) en dash, (4) em dash, (5) horizontal bar, (6) list icon, and so on.

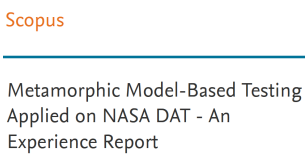
For example, Fig. 11a shows the original paper title of [6], which includes a hyphen (in "Model-based") and a dash (in "NASA DAT —an experience report"). Figs. 11b and 11c show that both the Scopus and Web of Science citation databases have converted the dash into a hyphen (and the Scopus database further added a white space between the hyphen and the word "An")[6]. In contrast, Fig. 11d shows that the IEEE digital library uses two consecutive hyphens "--" to represent the dash, and Figs. 11e and 11f show that both the ACM digital library and Google Scholar have converted the dash into a colon (in "NASA DAT: an experience report").

The above observation raises a question concerning the compatibility of the different representations of dashes among different citation databases. For example, even in a first-hand citation, the citer could easily take the bibliographic data directly from the ACM digital library and, hence, cite Lindvall et al.'s work [6] as "...NASA DAT: an experience report" rather than "...NASA DAT —an experience report." From the perspectives of the Scopus and Web of Science citation databases (both of which use a hyphen rather than a colon to represent the dash), this is a missing-hyphen citation error. Are these two citation database systems robust enough to correctly match the citation to the cited article?

Figs. 11g and 11h show two such cases: The citing articles are [65] and [66], and the cited article is [6]. In the reference lists of both [65] and [66], the dash in the cited article's title is printed as a colon. Fig. 12b reveals that the citation shown in Fig. 11g is lost in the Web of Science citation databases (although both the citing and the cited articles are within the coverage of Web of Science, as shown in Figs. 12a and 12b), hence demonstrating a lack of robustness of the citation database system.

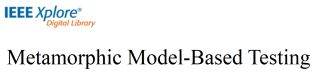


(a) Original paper title of [6].

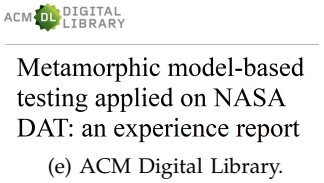


(b) Scopus.

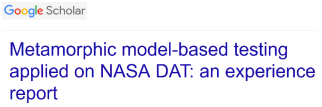
(c) Web of Science.



(d) IEEE Xplore Digital Library.



(e) ACM Digital Library.



(f) Google Scholar.

11 Lindvall M, Ganesan D, Árdal R, et al. Metamorphic model-based testing applied on NASA DAT: an experience report. In: Proceedings of the 37th International Conference on Software Engineering (ICSE'15), Florence, 2015. 129–138

(g) How [6] is cited in [65]. The *Web of Science* databases have failed to match this citation to the cited paper [6].

2. Lindvall, M., D. Ganesan, R. Ardal and R.E. Wiegand, 2015. Metamorphic model-based testing applied on NASA DAT: An experience report. Proceedings of the 37th International Conference on Software Engineering-Volume 2, May 16-24, 2015, Florence, Italy, pp: 129-138.

(h) How [6] is cited in [66].

Fig. 11: A case study of how the dash in the paper title of [6] is represented in various citation databases, and related software robustness / compatibility problems.

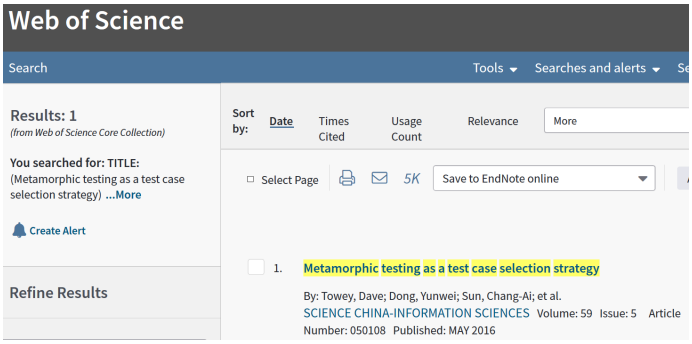
8.4 Robustness in GUI does not imply robustness in citation indexing

We find that the citation database systems under test are quite robust when queried with erroneous input through their Web-based GUI. For example, when we enter a paper title "Metamorphic model based testing applied on NASA DAT: an experience report" (which includes two missing-hyphen errors), the Web of Science website successfully returns Lindvall et al.'s paper [6], as shown in Fig. 13. It should be noted, however, that the system's GUI that parses user input, and the system's citation indexing interface that parses bibliographic data contained in citing articles' reference lists, are different modules, and the robustness of one cannot imply the robustness of the other.

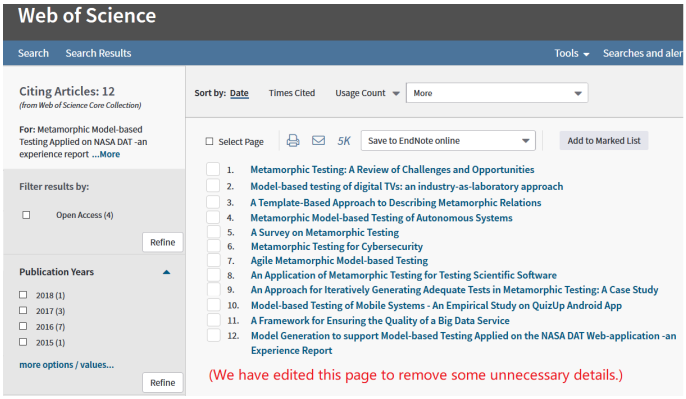
8.5 Exploring the ground truth using differential testing: A case study of Web of Science and Google Scholar

The analysis presented in this paper assumes that, for a given research field, the inclusion of a hyphen in paper titles in general is not a negative factor related to potential impact or likely citations of the paper. We have used this as a commonsense assumption without attempting a more systematic proof.

6. Because all datasets used in the present paper are downloaded from the Scopus and Web of Science databases, the "hyphens" discussed in this paper subsume dashes.



(a) The citing article [65] is covered by *Web of Science* (Core Collection).



(b) The *Web of Science* fails to identify [65] as a citing article for [6]. (Accessed in September 2018.)

Fig. 12: A lost citation in *Web of Science*: The system has failed to match the citation shown in Fig. 11g (where a colon rather than a hyphen is used to represent the dash) to the cited article (where a dash is included in the paper title), although both the citing article [65] and the cited article [6] fall within the coverage of the *Web of Science* citation databases.

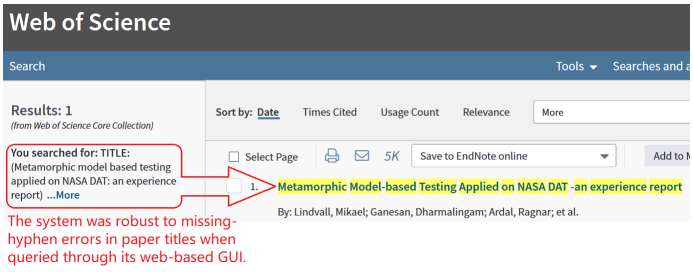


Fig. 13: When we searched for the paper title "Metamorphic model based testing applied on NASA DAT: an experience report" (which included two missing-hyphen errors) through the Web-based Graphical User Interface of Web of Science, the system found the article [6].

This research could be enhanced by exploring the above assumption through further empirical studies. For example, we could manually check the citation information for a sample of papers whose titles have varying numbers of hyphens to see whether they are given wrong citation counts. This kind of investigation, however, faces the oracle problem: It is not easy to find missing citations for a given paper's citation report. Although Section 8.4 has reported that the citation database systems under test are quite robust when queried with erroneous input through their Web-based GUI, this observed robustness is only with respect to paper title search (where the user enters a paper title p through the Web GUI and then the system returns a link to the actual paper titled p) rather than citation search (where the user enters a paper title p through the Web GUI and then the system returns all articles that cite p). The Web GUI uses the same citation databases as we have used in previous sections and, therefore, cannot provide a more accurate citation report.

In previous sections (for example, Sections 2, 5.2.1 and 8.3), we have reported some missing citation cases. For conducting a more systematic investigation, one potentially useful approach could be through *differential testing* [67], a strategy that tests multiple programs using the same input and observes differences in the programs' output. Since Scopus and Web of Science exhibit the same behavior pattern for hyphenated paper titles, differential testing that compares those two (as pseudo oracles for each other) would provide very limited help. However, there are other publicly available citation database systems, the most popular one being Google Scholar. We therefore decide to use Google Scholar as an independent source (pseudo oracle) to conduct differential testing. However, Google Scholar (as well as many other citation databases) has a major problem: It does not provide any public API or bulk download facility to enable quick collection of a large number of citation records. This means that we have to manually collect and process the citation data of each and every paper under study at the Google Scholar website https://scholar.google.com, which is a very time-consuming process. Worse, the outputs of the three systems (Scopus, Web of Science, and Google Scholar) are not directly comparable as they use different databases with different coverage. Moreover, it should be noted that, unlike Scopus and Web of Science, Google Scholar does not release information about its database's coverage [68], making it even more difficult to judge whether there are missing citations in Google-Scholar-generated citation reports.

Despite these challenges, we have managed to conduct a small-scale case study using differential testing, described as follows:

We randomly select four *cited papers* (including two 7-hyphen and two 0-hyphen ones) and collect their citation reports from both Google Scholar and Web of Science at around the same time. Details of these four papers and

their citation counts are shown in the first five columns of Table 2. As shown in the first column of the table, two of these papers are labeled as "low," indicating that they have *relatively* lower citation counts than the other two (which are labeled as "high"). The Google Scholar and Web of Science citation counts shown in columns 4 and 5 are not directly comparable because of the different coverage of these two citation database systems. Therefore, we perform the following checking for each cited paper (denoted by p_{cited}): For each citing paper p_{citing} included in the Google Scholar citation report of p_{cited}, we check whether p_{citing} is also included in the Web of Science (Core Collection) citation report; if not, then we go to the Web of Science website to check whether p_{citing} is an article indexed by the Web of Science (Core Collection) database — if yes then we download p_{citing} and check whether it has indeed cited p_{cited} and, if yes, then we have found a missing citation: p_{citing} is a citing article indexed by the Web of Science database but not counted in the Web of Science citation report for p_{cited}. Such missing citations are listed in the last column of Table 2, where the second, third, and fourth rows show that missing citations have been found for the two 7-hyphen papers but not for the 0-hyphen (low) paper. While this observation does seem to suggest an advantage of the non-hyphenated paper title, the last row of the table shows that the "0-hyphen (high)" paper also has three citations missing in the Web of Science citation report — however, a further investigation reveals that all these three citing articles that are missing in the Web of Science citation report do not include the cited paper's title in their reference lists and, more importantly, the first citing article includes wrong page numbers and the second citing article includes no journal name. Therefore, while we could say that Web of Science is not robust enough in finding citing articles, one could also argue that it may not necessarily be Web of Science's fault to miss the first two citing articles (whose erroneous reference list entries are highlighted in the cell at the lower-right corner of Table 2) because the underlying database of Web of Science might have adopted a strict filtering rule to ignore erroneous or incomplete citations. On the other hand, the inclusion of these three citing articles in Google Scholar's citation report might imply Google's stronger robustness and search capability (and/or a less strict filtering rule).

Having said that, we have also identified some major faults in Google Scholar, one of which is that it could mistakenly consider a reference list to be part of another article that appears on the same page as the reference list, as shown in Fig. 14. It should also be noted that we have decided not to use Web of Science as a pseudo oracle to identify missing citations in Google Scholar's citation reports, because the latter does not release information about its database's coverage.

In summary, this small case study has not only confirmed that missing citations of hyphenated paper titles are indeed part of a ground truth (as shown in the second, third, and fourth rows of Table 2) but has also revealed other potential robustness issues of Web of Science for dealing with erroneous or incomplete citations such as wrong page numbers and missing paper titles (as shown in the last row of Table 2). Furthermore, this case study has also revealed major defects in the functional correctness of Google Scholar, as shown in Fig. 14.

9 Related Work

This section reviews some of the related research in the areas of citation analysis and robustness testing.

9.1 Citation analysis

Citation analysis is a field that examines the frequency, patterns, and graphs of citations in documents. For example, at the article level, Habibzadeh and Yadollahie [69] found that longer titles were associated with more citations. This was confirmed by Jamali and Nikzad [70]. However, their finding was superseded by Paiva et al. [71], who found the opposite effect, while Fumania et al. [72] found no correlation between title lengths and citation counts.

It should be noted that the present research is **fundamentally different** from the citation analyses described above, as we aim to detect **software issues (robustness defects)** of citation database systems. For example, while the recent citation analysis reported that papers with shorter titles tended to be cited more than those with longer titles [53]–[55], we find that it is actually the number of hyphens in the title, not the title length, that serves as the dominating factor for citation counts, and that **this is a result of lack of robustness of the citation database systems**. Therefore, the present research belongs to the discipline of software engineering, not citation analysis.

9.2 Addressing the Oracle Problem in Robustness Testing

Fuzzing is a major approach for assessing software robustness. The term "fuzz testing" originates from a 1988 course project designed by Barton Miller at the University of Wisconsin [73], [74]. The "Fuzz Testing" website at the University of Wisconsin[7] lists the following unique characteristics of fuzzing: First, the input is random (in the original command-line studies, a fuzz test case is "simply random ASCII character streams"). Second, the pass criterion is simple: A failure is detected if the SUT *crashes* or *hangs*, otherwise it passes.

As a result of the above two characteristics, fuzzing has been successfully implemented into many automated testing tools, and detected a large number of software vulnerabilities in a variety of real-life systems. On the other hand, however, because of the oracle problem, fuzzing alone can hardly detect logic errors (which

7. http://pages.cs.wisc.edu/~bart/fuzz

TABLE 2: Case study results: Using Google Scholar to find missing citations in Web of Science (WoS) citation reports.

ID	Title of cited paper	Download link of cited paper	Google Scholar total cite	WoS Core Collection cite	WoS missing citations (citing papers that are not included in citatoin report)
7-hyphen (low)	Once-daily dolutegravir versus twice-daily raltegravir in antiretroviral-naive adults with HIV-1 infection (SPRING-2 study): 96 week results from a randomised, double-blind, non-inferiority trial	https://doi.org/10.1016/S1473-3099(13)70257-3	286	186	**(1) No clinically significant pharmacokinetic interactions between dolutegravir and daclatasvir in healthy adult subjects** **(2) SPRING-2 the future of antiretroviral therapy**
0-hyphen (low)	Glacial Survival of Boreal Trees in Northern Scandinavia	https://doi.org/10.1126/science.1216043	255	170	**nil**
7-hyphen (high)	Short- and Long-Term Outcomes With Drug-Eluting and Bare-Metal Coronary Stents A Mixed-Treatment Comparison Analysis of 117 762 Patient-Years of Follow-Up From Randomized Trials	https://doi.org/10.1161/CIRCULATIONAHA.112.097014	485	343	**(1) bivalirudin versus heparin in patients treated with percutaneous coronary intervention: a meta-analysis of randomised trials** **(2) incidence and implications of coronary artery disease in patients undergoing valvular heart surgery: the indian scenario** **(3) new concepts in the design of drug-eluting coronary stents**
0-hyphen (high)	Silk fibroin biomaterials for tissue regenerations (Banani Kundu, Rangam Rajkhowa, Subhas C. Kundu, and Xungai Wang, *Advanced Drug Delivery Reviews* 65 (2013) 457–470)	https://doi.org/10.1016/j.addr.2012.09.043	524	339	**(1) characterization of silk sponge in the wet state using 13 c solid state nmr for development of a porous silk vascular graft with small diameter** Reference list entry **(no paper title, wrong page number)**: B. Kundu, R. Rajkhowa, S. C. Kundu and X. Wang, Adv. Drug Delivery Rev., 2013, 65, 403–604. **(2) comparision of electro spun tassar silk fibroin-hydroxyapatite composite scaffold prepared by soaking and in-situ methods** Reference list entry **(no paper title, no journal name)**: Banani Kundu A, Rangam Rajkhowa B, Subhas C. Kunda, Xungai Wang (2013), 65: 457–470. **(3) effect of uv-light on the uniaxial tensile properties and structure of uncoated and tio2 coated bombyx mori silk fibers** Reference list entry: B. Kundu, R. Rajkhowa, S.C. Kundu, X. Wang, Adv. Drug Deliv. Rev. 65 (2013) 457–470.

produce erroneous output but do not cause the SUT to crash or hang) [21].

To address the oracle problem of fuzzing / random testing, Zhou et al. [15], [16] combined metamorphic testing and fuzzing by feeding the SUT with random ASCII characters and then checking the SUT's output against certain metamorphic relations — even if the SUT does not crash or hang, a failure can still be detected if the metamorphic relation is violated. This strategy was successful, and they reported on the detection of previously unknown bugs in real-life Web search engines. For example, they first entered a random string "GLIF," and the Microsoft search engine returned "11,783" results. They then generated an additional random string "5Y4W," setting the search criteria to be "any of these terms" (that is, Web pages that contain either "GLIF" or "5Y4W" should be returned), but this time the search engine returned 0 results, violating the expected metamorphic relation. Because the failure was repeatable, a bug in the search engine was revealed (and later confirmed by Microsoft).

Zhou et al.'s metamorphic relations for search engines [15] were later adopted and further developed by Murphy [75] for testing Apache Lucene, an open-source search software.[8] Murphy further developed a *metamorphic runtime checking* technique, in which the SUT was tested by checking the metamorphic relations of its individual functions while the entire system was run — the software execution was at the system level (that is, the full application level), although the metamorphic relation was at the function level [75]. This was an innovative strategy that differed from both traditional system testing

8. http://lucene.apache.org/

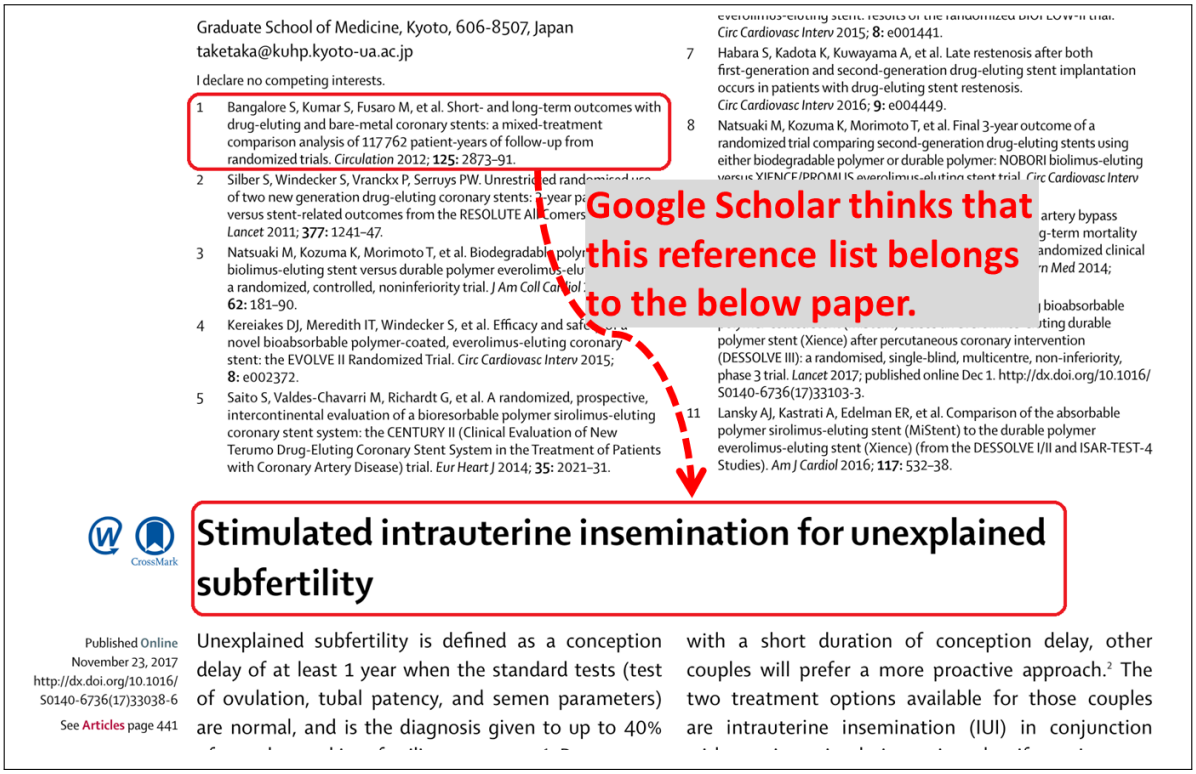


Fig. 14: Excerpt of a printed journal page: Google Scholar mistakenly considers the reference list entry "[1]" (by Bangalore et al., in the upper part of the page) as part of the paper titled "Stimulated intrauterine insemination for unexplained subfertility" that appears in the lower part of the page.

and traditional unit testing. In an empirical study [75], Murphy applied metamorphic runtime checking to PAYL, a network intrusion detection system. Murphy showed that, while traditional system-level testing could not modify the values of the bytes inside the payloads (because, at the system testing level, the front-end of the full PAYL application filtered out invalid inputs, which were syntactically or semantically malformed network packets, before they could reach most of the PAYL code), metamorphic runtime checking was able to feed both valid and invalid inputs (payload bytes) to internal functions, finding many more seeded bugs than when only valid inputs were used during mutation testing. In other words, metamorphic runtime checking enabled the tester to circumvent the restrictions imposed by the front-end of the full PAYL application and, therefore, the SUT could be tested using both valid and invalid inputs against the function-level metamorphic relations that involved changing the byte values.

In the area of automated test case generation, a related technique is known as *data mutation* [76]: For a given set of seed inputs, data mutation generates new inputs by changing the original inputs using *data mutation operators*, such as increasing or decreasing some input parameters' values. Data mutation and metamorphic testing can be combined by considering not only the changes to the input but also the impact of such changes on the output [77].

Chen et al. [21] explicitly stated that there is a "feasibility of combining MT and fuzzing," and they used this strategy to test real-life applications, detecting previously unknown logic errors in several security-critical software products (including both open-source and commercial software).

In recent years, a trend has emerged for applying metamorphic testing to address the oracle problem in testing machine learning and autonomous systems [11], [26], [78], [79]. Tian et al. [11] tested three different Deep Neural Network (DNN) models for autonomous driving. For each DNN model, the input was a picture from a camera, and the output was a steering angle. To check the correctness of the output, Tian et al. used metamorphic relations based on image transformations, such as by adding synthetic weather effects to road images. These transformations could generate valid but sometimes unexpected inputs. Tian et al. found a large number of corner case inputs leading to erroneous behavior in the three DNN models.

The software that Tian et al. tested was deep learning models that "won top positions in the Udacity self-driving challenge" [11]. In contrast to their work, Zhou and Sun [14] tested a *real-life* system, Baidu Apollo (a well-known real-world self-driving software system controlling many cars on the road today, http://apollo.auto). They combined metamorphic testing and fuzzing, and detected previously unknown fatal software bugs in

the LiDAR obstacle-perception module of Baidu Apollo, reporting the alarming findings eight days before Uber's deadly crash in Tempe, AZ, USA, in March 2018 [14].

10 Limitations

This section discusses some of the limitations of this research.

10.1 Selection of Metamorphic Relations

In this paper, we have presented three metamorphic relations: $MR_{similar}$, MR_{older}, and MR_{aging}. These MRs have turned out to be effective in revealing the hidden defects in the SUTs; however, we have not presented a strategy for the identification of effective MRs for a wider range of applications.

A recent trend in the research direction of MR identification is the development of *metamorphic relation patterns* [18], [80], [81]. A "metamorphic relation pattern" (MRP) is "an abstraction that characterizes a set of (possibly infinitely many) metamorphic relations" [81] and, hence, can be used to derive many concrete MRs. Zhou et al. [81] also defined a *symmetry* MRP as follows: "The *symmetry* MRP refers to the existence of different viewpoints from which the system appears the same" (note that this does not mean that the software system's source and follow-up outputs must have an equality or equivalence relation). They further hypothesized that "*symmetry* and *asymmetry* are two fundamental MR patterns that come in pairs for computer systems."

In the present research, the identification of the metamorphic relation $MR_{similar}$ is a direct application of the *symmetry* MRP to the citation indexing domain. In fact, $MR_{similar}$ could be understood as a sub-pattern under *symmetry*. Similarly, the identification of MR_{older} is a direct application of the *asymmetry* MRP, whereas MR_{aging} is identified when we think about *symmetry* with respect to time, in the context of considering how the citation counts would change over time.

Using the "pattern" concept, more MRs could be identified for testing citation database systems, and this will be a future research direction.

10.2 Determination of Sample Sizes

The definitions of the three MRs ($MR_{similar}$, MR_{older}, and MR_{aging}) include some phrases such as *large*, *systematic difference*, and *higher citation counts*. We have used these phrases without explaining how to measure them, because these concepts are studied in the field of statistical science and, therefore, an in-depth discussion is beyond the scope of this paper.

For a given field of study, there could be different ways of determining sample sizes, such as using experience, using a target confidence interval, using a confidence level (the larger the required confidence level, the larger the sample size), using a pilot study (to obtain necessary parameter estimates), and so on. Interested readers are referred to the literature of statistics and sample size determination for further information on this topic [82], [83].

Generally speaking, larger sample sizes result in higher precision of estimation. Fig. 2a, for example, shows an increase in mean citation counts of the 6- and 7-hyphen groups, and Fig. 2b shows an increase in the mean citation count of the 6-hyphen group. These "anomalies" are caused by the significantly reduced sample sizes of these groups (see Table 1), and can be eliminated by combining the last few small-sample high-hyphen groups (the rightmost bars of Figs. 2a and 2b). However, for our study, such a combination is unnecessary because, without any combination, the negative correlation between the number of hyphens and the mean citation count is already very strong and statistically significant: For the data presented in Figs. 2a and 2b, the Spearman's rank correlation coefficients (Spearman's rho) are -0.91667 and -0.98333, respectively, and the p-values (2-tailed) of both cases are below 0.001. These results mean that there is a systematic pattern that violates the expected metamorphic relation.

10.3 Other factors

In this research, we have reported on the impact of hyphens in paper titles. It is reasonable to further ask whether other factors, such as other symbols in paper titles, could have a similar impact. We have therefore conducted a preliminary study to investigate the impact of other symbols including the colon (:), semicolon (;), comma (,), and period (.). We have not found any systematic and statistically meaningful trend as found in hyphens.

It should however be noted that, due to the limited scope of this research, we cannot exclude the possibility of the impact of other factors such as the inclusion of non-English characters (for example, μ) in paper titles, incorrect spelling of author names, wrong page numbers, etc. Furthermore, in the present research, we have not investigated the impact of the incidental line-break (automated word-break) hyphen that often appears at the end of a line in a paper's reference list.

11 Conclusion

In this research, we have presented a metamorphic robustness testing approach, which examines the software's output for erroneous input. Using this approach, in combination with a fault-based testing strategy, we have analyzed large datasets that are outputs of two major citation database systems: Scopus and the Web of Science.

Our data analyses against three metamorphic relations ($MR_{similar}$, MR_{older}, and MR_{aging}) have revealed surprising hidden defects in these two citation database systems. At the article level, we find that the inclusion of hyphens in paper titles distorts the citation counts. The results are shown to be more serious in highly cited papers or those published earlier, which is consistent with the

finding by other researchers that errors in citations may be propagated from one author to another because the citers may not necessarily read the papers that they cite or verify the bibliographies. We have also shown that our results are valid even when we limit the scope of analysis to individual research fields such as chemistry. At the journal level, we have found that there is a strong and significant negative correlation between the *journal impact factor* of *IEEE Transactions on Software Engineering* and the percentage of hyphenated paper titles published in the journal. An even stronger negative correlation has been found for *ACM Transactions on Software Engineering and Methodology*. A further software engineering field-wide study shows a clear pattern that the higher JIF-ranked journals are publishing a lower percentage of papers with hyphenated titles.

We have shown that this research is fundamentally different from the field of citation analysis, where citation counts are generally regarded as a reliable measure for the assessment of papers. From the field of citation analysis, it was reported that papers with shorter titles tended to be cited more than those with longer titles [53], and this finding was widely reported by the media worldwide, including Science [54] and Nature [55]. In the present research, we have provided strong evidence to show that it is actually the number of hyphens in the title, not the title length, that serves as the dominating factor for citation counts, and that it is a result of lack of robustness in the underlying citation database systems. This paper, therefore, contributes to both the theory and practice of software engineering. We have provided a careful analysis of the validity of this research to avoid falling into the well-known trap of equating correlation with causation.

With regard to our four research questions, we have provided a negative answer to RQ4, which implies a negative answer to RQ2 and, hence, a negative answer to RQ1. Although we have not addressed RQ3 thoroughly, the explicit statement of RQ3 and the understanding of its relationship to RQ2 and RQ4 make the preceding reasoning straightforward.

As a consequence of this study, we question the reliability of citation statistics and *journal impact factors*, because the number of hyphens in paper titles should have no bearing on the actual quality of the respective articles and journals.

In future research, we plan to apply metamorphic robustness testing to other areas involving the collection and processing of big data.

Acknowledgments

This work was supported in part by a linkage grant of the Australian Research Council (Project ID: LP160101691) and an Australian Government Research Training Program scholarship. We wish to thank Morphick Pty Ltd for supporting this research. We are grateful to the anonymous reviewers for their valuable comments. All correspondence should be addressed to Dr. Z. Q. Zhou at the address shown on the first page of this paper.

References

[1] M. Pezzè and C. Zhang, "Automated test oracles: a survey," in *Advances in Computers*, A. Memon, Ed. Elsevier Science & Technology, 2014, vol. 95, ch. 1, pp. 1–48.

[2] E. T. Barr, M. Harman, P. McMinn, M. Shahbaz, and S. Yoo, "The oracle problem in software testing: A survey," *IEEE Transactions on Software Engineering*, vol. 41, no. 5, pp. 507–525, 2015.

[3] K. Patel and R. M. Hierons, "A mapping study on testing non-testable systems," *Software Quality Journal*, vol. 26, no. 4, pp. 1373–1413, 2018.

[4] T. Y. Chen, S. C. Cheung, and S. M. Yiu, "Metamorphic testing: A new approach for generating next test cases," Department of Computer Science, Hong Kong University of Science and Technology, Hong Kong, Tech. Rep. HKUST-CS98-01, 1998.

[5] T. Y. Chen, T. H. Tse, and Z. Q. Zhou, "Fault-based testing without the need of oracles," *Information and Software Technology*, vol. 45, no. 1, pp. 1–9, 2003.

[6] M. Lindvall, D. Ganesan, R. Árdal, and R. E. Wiegand, "Metamorphic model-based testing applied on NASA DAT —an experience report," in *Proceedings of the IEEE/ACM 37th International Conference on Software Engineering (ICSE '15)*, 2015, pp. 129–138.

[7] S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Cortés, "A survey on metamorphic testing," *IEEE Transactions on Software Engineering*, vol. 42, no. 9, pp. 805–824, 2016.

[8] U. Kanewala, L. L. Pullum, S. Segura, D. Towey, and Z. Q. Zhou, "Message from the workshop chairs," in *Proceedings of the IEEE/ACM 1st International Workshop on Metamorphic Testing (MET '16), in conjunction with the 38th International Conference on Software Engineering (ICSE '16)*. ACM, 2016.

[9] S. Segura and Z. Q. Zhou, "Metamorphic testing 20 years later: A hands-on introduction," in *Proceedings of the IEEE/ACM 40th International Conference on Software Engineering (ICSE '18 Companion)*. ACM, 2018.

[10] Z. Wang, D. Towey, Z. Q. Zhou, and T. Y. Chen, "Metamorphic testing for Adobe Analytics data collection JavaScript library," in *Proceedings of the IEEE/ACM 3rd International Workshop on Metamorphic Testing (MET '18), in conjunction with the 40th International Conference on Software Engineering (ICSE '18)*. ACM, 2018, pp. 34–37.

[11] Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: Automated testing of deep-neural-network-driven autonomous cars," in *Proceedings of the IEEE/ACM 40th International Conference on Software Engineering (ICSE '18)*. ACM, 2018, pp. 303–314.

[12] N. Mouha, M. S. Raunak, D. R. Kuhn, and R. Kacker, "Finding bugs in cryptographic hash function implementations," *IEEE Transactions on Reliability*, vol. 67, no. 3, pp. 870–884, 2018.

[13] T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, T. H. Tse, and Z. Q. Zhou, "Metamorphic testing: A review of challenges and opportunities," *ACM Computing Surveys*, vol. 51, no. 1, pp. 4:1–4:27, 2018.

[14] Z. Q. Zhou and L. Sun, "Metamorphic testing of driverless cars," *Communications of the ACM*, vol. 62, no. 3, pp. 61–67, March 2019. [Online]. Available: https://doi.org/10.1145/3241979

[15] Z. Q. Zhou, T. H. Tse, F.-C. Kuo, and T. Y. Chen, "Automated functional testing of web search engines in the absence of an oracle," Department of Computer Science, The University of Hong Kong, Tech. Rep. TR-2007-06, 2007.

[16] Z. Q. Zhou, S. Zhang, M. Hagenbuchner, T. H. Tse, F.-C. Kuo, and T. Y. Chen, "Automated functional testing of online search services," *Software Testing, Verification and Reliability*, vol. 22, no. 4, pp. 221–243, 2012.

[17] Z. Q. Zhou, S. Xiang, and T. Y. Chen, "Metamorphic testing for software quality assessment: A study of search engines," *IEEE Transactions on Software Engineering*, vol. 42, no. 3, pp. 264–284, 2016.

[18] S. Segura, J. A. Parejo, J. Troya, and A. Ruiz-Cortés, "Metamorphic testing of RESTful web APIs," *IEEE Transactions on Software Engineering*, vol. 44, no. 11, pp. 1083–1099, November 2018.

[19] V. Le, M. Afshari, and Z. Su, "Compiler validation via equivalence modulo inputs," in *Proceedings of the 35th ACM SIGPLAN Conference on Programming Language Design and Implementation (PLDI'14)*, 2014, pp. 216–226.

[20] J. Regehr, "Finding compiler bugs by removing dead code," http://blog.regehr.org/archives/1161, June 20, 2014.

[21] T. Y. Chen, F.-C. Kuo, W. Ma, W. Susilo, D. Towey, J. Voas, and Z. Q. Zhou, "Metamorphic testing for cybersecurity," *Computer*, vol. 49, no. 6, pp. 48–55, 2016.

[22] A. F. Donaldson, H. Evrard, A. Lascu, and P. Thomson, "Automated testing of graphics shader compilers," *Proceedings of the ACM on Programming Languages*, vol. 1, no. OOPSLA, pp. 93:1–93:29, 2017.

[23] J. Brown, Z. Q. Zhou, and Y.-W. Chow, "Metamorphic testing of navigation software: A pilot study with Google Maps," in *Proceedings of the 51st Annual Hawaii International Conference on System Sciences (HICSS-51)*, 2018, pp. 5687–5696, available: http://hdl.handle.net/10125/50602.

[24] J. Rothermel, M. Lindvall, A. Porter, and S. Bjorgvinsson, "A metamorphic testing approach to NASA GMSEC's flexible publish and subscribe functionality," in *Proceedings of the IEEE/ACM 3rd International Workshop on Metamorphic Testing (MET '18), in conjunction with the 40th International Conference on Software Engineering (ICSE '18)*. ACM, 2018, pp. 18–25.

[25] D. C. Jarman, Z. Q. Zhou, and T. Y. Chen, "Metamorphic testing for Adobe data analytics software," in *Proceedings of the IEEE/ACM 2nd International Workshop on Metamorphic Testing (MET '17), in conjunction with the 39th International Conference on Software Engineering (ICSE '17)*, 2017, pp. 21–27.

[26] A. Dwarakanath, M. Ahuja, S. Sikand, R. M. Rao, R. P. J. C. Bose, N. Dubash, and S. Podder, "Identifying implementation bugs in machine learning based image classifiers using metamorphic testing," in *Proceedings of the ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA '18)*. ACM, 2018, pp. 118–128.

[27] Accenture. (2018) Quality engineering in the new: A vision and R&D update from Accenture Labs and Accenture Testing Services. [Online]. Available: https://www.accenture.com/t20180627T065422Z__w__/us-en/_acnmedia/PDF-81/Accenture-Quality-Engineering-POV.pdf

[28] GraphicsFuzz homepage. [Online]. Available: https://www.graphicsfuzz.com

[29] GraphicsFuzz. How it works. [Online]. Available: https://www.graphicsfuzz.com/howitworks.html

[30] A. F. Donaldson and A. Lascu, "Metamorphic testing for (graphics) compilers," in *Proceedings of the IEEE/ACM 1st International Workshop on Metamorphic Testing (MET '16), in conjunction with the 38th International Conference on Software Engineering (ICSE '16)*. ACM, 2016, pp. 44–47.

[31] X. Xie, J. W. K. Ho, C. Murphy, G. Kaiser, B. Xu, and T. Y. Chen, "Testing and validating machine learning classifiers by metamorphic testing," *Journal of Systems and Software*, vol. 84, pp. 544–558, 2011.

[32] ISO/IEC 25010:2011, "Systems and software engineering – systems and software quality requirements and evaluation (SQuaRE) – system and software quality models," 2011.

[33] C. Ghezzi, M. Jazayeri, and D. Mandrioli, *Fundamentals of Software Engineering*, 2nd ed. Pearson, 2002.

[34] A. Vassilev and C. Celi, "Avoiding cyberspace catastrophes through smarter testing," *Computer*, vol. 47, no. 10, pp. 102–106, October 2014.

[35] V. Okun and E. Fong, "Fuzz testing for software assurance," *CrossTalk – The Journal of Defense Software Engineering*, vol. 28, no. 2, pp. 35–37, 2015.

[36] B. Hjørland, "The importance of theories of knowledge: Indexing and information retrieval as an example," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 1, pp. 72–77, 2011.

[37] A. D. Cleveland and D. B. Cleveland, *Introduction to Indexing and Abstracting*, 4th ed. Libraries Unlimited, 2013.

[38] Scopus. [Online]. Available: https://www.elsevier.com/solutions/scopus

[39] Web of Science. [Online]. Available: https://clarivate.com/products/web-of-science/

[40] "San francisco declaration on research assessment." [Online]. Available: https://sfdora.org

[41] R. Just and F. Schweiggert, "Evaluating testing strategies for imaging software by means of mutation analysis," in *Proceedings of the IEEE International Conference on Software Testing, Verification, and Validation Workshops*. IEEE, 2009, pp. 205–209.

[42] U. Y. Raja and J. G. Cooper, "How accurate are the references in Emergency Medical Journal?" *Emergency Medicine Journal*, vol. 23, no. 8, pp. 625–626, 2006.

[43] B. Ghai, A. K. Saxena, and J. K. Makkar, "A guide to reducing citation errors in bibliographies," *Emergency Medicine Journal*, vol. 24, no. 3, pp. 232–233, 2007.

[44] J. Mayer and R. Guderlei, "On random testing of image processing applications," in *Proceedings of the 6th International Conference on Quality Software (QSIC '06)*. IEEE, 2006.

[45] T. Y. Chen, T. H. Tse, and Z. Zhou, "Fault-based testing in the absence of an oracle," in *Proceedings of the 25th Annual International Computer Software and Applications Conference (COMPSAC '01)*. IEEE, 2001, pp. 172–178.

[46] T. Y. Chen, T. H. Tse, and Z. Q. Zhou, "Semi-proving: An integrated method for program proving, testing, and debugging," *IEEE Transactions on Software Engineering*, vol. 37, no. 1, pp. 109–125, 2011.

[47] J. Wilsdon, *The Metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. London: SAGE Publications, 2016.

[48] R. Adler, J. Ewing, and P. Taylor, "Citation statistics: A report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS)," *Statistical Science*, vol. 24, no. 1, pp. 1–14, 2009.

[49] S. Lehmann, A. D. Jackson, and B. E. Lautrup, "Measures for measures," *Nature*, vol. 444, no. 7122, pp. 1003–1004, 2006.

[50] C. Neylon and S. Wu, "Article-level metrics and the evolution of scientific impact," *PLOS Biology*, vol. 7, no. 11, 2009.

[51] E. Garfield, "The history and meaning of the journal impact factor," *The Journal of the American Medical Association*, vol. 295, no. 1, pp. 90–93, 2006.

[52] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[53] A. Letchford, H. S. Moat, and T. Preis, "The advantage of short paper titles," *Royal Society Open Science*, vol. 2, no. 8, 2015. [Online]. Available: https://doi.org/10.1098/rsos.150266

[54] D. S. Chawla, "In brief, papers with shorter titles get more citations, study suggests." *ScienceInsider*, August 25, 2015. [Online]. Available: https://doi.org/10.1126/science.aad1669

[55] B. Deng, "Papers with shorter titles get more citations: Intriguing correlation mined from 140,000 papers." *Nature*, August 26, 2015. [Online]. Available: https://doi.org/10.1038/nature.2015.18246

[56] A. Letchford, H. S. Moat, and T. Preis, "Data from: The advantage of short paper titles," 2015. [Online]. Available: https://doi.org/10.5061/dryad.hg3j0

[57] S. Hanafi and S. Boucherie, "Discover the data behind the Times Higher Education World University Rankings," January 18, 2018. [Online]. Available: https://www.elsevier.com/connect/discover-the-data-behind-the-times-higher-education-world-university-rankings

[58] QS Intelligence Unit, "Citations per faculty," 2018. [Online]. Available: http://www.iu.qs.com/university-rankings/indicator-citations-per-faculty

[59] A. A. Chadegani, H. Salehi, M. M. Yunus, H. Farhadi, M. Fooladi, M. Farhadi, and N. A. Ebrahim, "A comparison between two main academic literature collections: Web of Science and Scopus databases," *Asian Social Science*, vol. 9, no. 5, pp. 18–26, 2013.

[60] M. V. Simkin and V. P. Roychowdhury, "Do you sincerely want to be cited? Or: Read before you cite," *Significance*, vol. 3, no. 4, pp. 179–181, 2006.

[61] ——, "A mathematical theory of citing," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 11, pp. 1661–1673, 2007.

[62] M. Dwyer, "State of the journal," *IEEE Transactions on Software Engineering*, vol. 44, no. 1, pp. 1–2, 2018.

[63] A. Field, *Discovering Statistics Using IBM SPSS Statistics: North American Edition*, 5th ed. SAGE Publications Ltd, 2017.

[64] C. G. Scanes, "Editorial: Professional ethics and publishing," *Poultry Science*, vol. 86, no. 4, pp. 603–604, 2007.

[65] D. Towey, Y. Dong, C.-A. Sun, and T. Y. Chen, "Metamorphic testing as a test case selection strategy," *Science China Information Sciences*, vol. 59, pp. 050 108:1–050 108:2, 2016.

[66] A. Meriem and M. Abdelaziz, "A meta-model for model-based testing technique: A review," *Journal of Software Engineering*, vol. 12, no. 1, pp. 1–11, 2018.

[67] W. M. McKeeman, "Differential testing for software," *Digital Technical Journal*, vol. 10, no. 1, pp. 100–107, 1998. [Online]. Available: http://www.hpl.hp.com/hpjournal/dtj/vol10num1/toc.htm

[68] J. C. Fagan, "An evidence-based review of academic web search engines, 2014-2016: Implications for librarians' practice and research agenda," *Information Technology and Libraries*, vol. 36, no. 2, pp. 7–47, Jun 2017. [Online]. Available: https://doi.org/10.6017/ital.v36i2.9718

[69] F. Habibzadeh and M. Yadollahie, "Are shorter article titles more attractive for citations? Cross-sectional study of 22 scientific journals," *Croatian Medical Journal*, vol. 51, no. 2, pp. 165–170, 2010.

[70] H. R. Jamali and M. Nikzad, "Article title type and its relation with the number of downloads and citations," *Scientometrics*, vol. 88, pp. 653–661, 2011.

[71] C. E. Paiva, J. P. da Silveira Nogueira Lima, and B. S. R. Paiva, "Articles with short titles describing the results are cited more often," *Clinics*, vol. 67, no. 5, pp. 509–513, 2012.

[72] M. R. F. Q. Fumania, M. Goltajib, and P. Partoc, "The impact of title length and punctuation marks on article citations," *Annals of Library and Information Studies*, vol. 62, pp. 126–132, 2015.

[73] B. P. Miller, L. Fredriksen, and B. So, "An empirical study of the reliability of Unix utilities," *Communications of the ACM*, vol. 33, no. 12, pp. 32–44, 1990.

[74] A. Takanen, J. D. DeMott, C. Miller, and A. Kettunen, *Fuzzing for Software Security Testing and Quality Assurance*, 2nd ed. Artech House, 2018.

[75] C. Murphy, "Metamorphic testing techniques to detect defects in applications without test oracles," Ph.D. dissertation, Columbia University Computer Science Technical Reports CUCS-010-10, USA, 2010. [Online]. Available: https://academiccommons.columbia.edu/doi/10.7916/D81261JW

[76] L. Shan and H. Zhu, "Generating structurally complex test cases by data mutation: A case study of testing an automated modelling tool," *The Computer Journal*, vol. 52, no. 5, pp. 571–588, 2009.

[77] H. Zhu, "JFuzz: A tool for automated Java unit testing based on data mutation and metamorphic testing methods," in *Proceedings of the 2nd International Conference on Trustworthy Systems and Their Applications*. IEEE, 2015, pp. 8–15.

[78] M. Lindvall, A. Porter, G. Magnusson, and C. Schulze, "Metamorphic model-based testing of autonomous systems," in *Proceedings of the IEEE/ACM 2nd International Workshop on Metamorphic Testing (MET '17), in conjunction with the 39th International Conference on Software Engineering (ICSE '17)*, 2017, pp. 35–41.

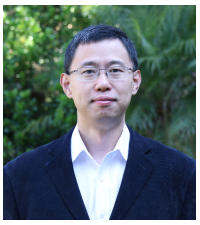
[79] M. Zhang, Y. Zhang, L. Zhang, C. Liu, and S. Khurshid, "DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE '18)*. ACM, 2018, pp. 132–142.

[80] S. Segura, "Metamorphic testing: Challenges ahead (keynote speech)," in *Proceedings of the IEEE/ACM 3nd International Workshop on Metamorphic Testing (MET '18), in conjunction with the 40th International Conference on Software Engineering (ICSE '18)*, May 27, 2018, pp. 1–1.

[81] Z. Q. Zhou, L. Sun, T. Y. Chen, and D. Towey, "Metamorphic relations for enhancing system understanding and use," *IEEE Transactions on Software Engineering*, in press. [Online]. Available: https://doi.org/10.1109/TSE.2018.2876433

[82] T. P. Ryan, *Sample Size Determination and Power*, ser. Wiley Series in Probability and Statistics. Hoboken, New Jersey: Wiley, 2013.

[83] E. L. Lehmann, *Elements of Large-Sample Theory*, ser. Springer Texts in Statistics. New York: Springer, 2004.



Zhi Quan Zhou received the BSc degree in computer science from Peking University, China, and the PhD degree in software engineering from The University of Hong Kong. He is currently an associate professor and director of the Bachelor of Computer Science degree at the University of Wollongong, Australia. His current research interests include software testing and debugging; the interplay among software testing, machine learning, and big data; and self-driving vehicles. Zhou was a main contributor to some of the earliest research papers on metamorphic testing, and was one of the few pioneers who opened up and established the metamorphic testing research field. In 2016, he co-founded and chaired the IEEE/ACM 1st International Workshop on Metamorphic Testing, in conjunction with ICSE (ICSE MET '16). He was an invited keynote speaker at ICSE MET '17 and at the IEEE International Conference on Artificial Intelligence Testing, 2019. He was an invited ACM SIGSOFT Webinar speaker, an ICSE '18 Technical Briefings speaker, and an ICSE '16 and ICSE '19 journal-first speaker, introducing metamorphic testing through all these venues. He was selected for a Virtual Earth Award by Microsoft Research, Redmond, USA, and a 2018 Researcher of the Year Gold Disruptor Award by the Australian Computer Society.

T.H. Tse received the PhD degree from the London School of Economics and was a visiting fellow at the University of Oxford. He is an honorary professor in computer science at The University of Hong Kong after retiring from the full professorship in 2014. His research interest is in program testing and debugging. He is the steering committee chair of the IEEE International Conference on Software Quality, Reliability, and Security, an associate editor of *IEEE Transactions on Reliability*, and an editorial board member of *Software Testing, Verification and Reliability* and *Software: Practice and Experience*. He also served on the search committee for the editor-in-chief of *IEEE Transactions on Software Engineering* in 2013. He is a fellow of the British Computer Society. He was awarded an MBE by The Queen of the United Kingdom.

Matt Witheridge received the Bachelor of Computer Science (Honours) degree from the University of Wollongong, Australia, where he is currently a PhD student in computer science. His research interests include software testing and data analytics.