How do software engineering researchers assess the credibility of grey literature?

Ashley Williams, Austen Rainer
Department of Computer Science and Software Engineering
University of Canterbury, NZ
ashley.williams@pg.canterbury.ac.nz, austen.rainer@canterbury.ac.nz

October 2, 2018

Background: Grey literature is potentially valuable as a source of evidence for software engineering research e.g. providing insights on state–of–practice. We are particularly interested in the value of the better–quality blog articles.

Objective: To identify a set of criteria that can be used by software engineering researchers to evaluate the credibility of blog articles.

Method: We conduct a literature review of prior research to generate a set of criteria for evaluating the credibility of online content. We then conduct a complementary survey, of software engineering researchers (n=44), to identify more specific criteria for evaluating the credibility of software engineering blog articles.

Results: In our literature review, we find few definitions of the concept of credibility and no definitions for any criteria for credibility. For our survey, software engineering researchers generally recognise that at least some blog articles are credible and, by implication, have value for research. The criteria most highly ranked by researchers are: reasoning, empirical data, clear writing, reporting of data collection methods and links to prior research. Further, 60% of survey respondents think that the criteria generalise to other practitioner–generated content, and (surprisingly) 58% think that the criteria also apply to researcher–generated content.

Conclusion: Our literature review and survey together present an initial set of criteria for assessing the credibility of blog articles in software engineering. This criteria appear to have wider applicability for practice and research.

Keywords: evidence based software engineering, credibility, evidence, blogs, argumentation, experience, grey literature review

Version History

Version No.	Date	Comment
1.0.0	May 2018	Initial version
1.0.1	June 2018	Changes and further analysis following internal review
1.0.2	July 2018	Changes and further analysis following external review

1 Introduction & Background

1.1 Introduction

Software engineering researchers often use practitioners as a source of evidence in their studies. This evidence is often collected by interviewing practitioners, or by surveying and observing practitioners in the work environment. With the emergence of the web, practitioner knowledge has been increasingly reported in social media (see [56]). The dissemination of practitioner knowledge through social media presents additional opportunities to the research community to gather evidence from practitioners. There have been many studies that have analysed these new sources of evidence, for example; detecting trends in Stack Overflow [6]; analysing public health in Twitter [45]; and running sentiment analysis over GitHub commit comments [28].

The focus of this paper is on assessing the *credibility* of grey literature. Specifically, we focus on practitioner blog articles, so as to better evaluate their value as evidence. We are particularly interested in the personal blog articles of practitioners (in contrast to corporate or product blog articles) as such articles more likely contain the opinions, reasoning and personal experiences of the blog article author. But even practitioner articles may contain biased opinions, depending on the intention of the author. Opinions, reasoning and personal experiences are the kinds of information traditionally gathered using research methods such as interviews and surveys. Having collected the data, opinions are then selected, extracted, analysed and reported [34]. A similar approach applies to analysing blog articles, but we need to develop a set of selection criteria for assessing blog article content. In this paper, we propose a set of credibility criteria to support researchers in their blog article selection. We believe that the extraction of opinion, reasoning and experience from multiple practitioner blog articles can be considered as a case survey. Case surveys combine the benefits of case studies and surveys [32].

1.2 Grey literature in software engineering research

The use of social media as evidence for research may be understood as a potential source or type of grey literature. Interest in the value of grey literature within software engineering research is increasing. For example, Garousi and his colleagues [23, 24] have investigated the development of multi–vocal literature reviews (MLRs). MLRs combine grey literature reviews (GLRs) with systematic literature reviews (SLRs) so that the state–of–practice can be reported along side the state–of–art.

Software engineering researchers typically seek grey literature that contains some element of empirical data. For example, Bailey et al. [5] reported a literature survey of evidence for object oriented design. Their inclusion criteria included, "... books, papers, technical reports and 'grey literature' describing *empirical studies* regarding OO software design..." ([5], p. 483; emphasis added here). There is the implication that grey literature should contain some element of empirical study.

Adams et al. [2] identify a hierarchy of credibility within grey literature. However, their reported definitions–by–example for each tier within the hierarchy are based on the source of the data which implies that, for example, all Stack Overflow posts are more credible than all blog articles. This is not always the case and we instead look at the credibility of document content rather than judging it based on its source. Adams et al. [1] present a contrasting model that distinguishes between different types of grey literature (grey literature, grey information and grey data). However, in this paper we refer to all types as grey literature.

1.3 Blog articles as a source of grey literature

Blog articles differ from other forms of social media in that they provide a platform which is not restricted to a character limit (e.g. microblogging sites like Twitter) or one particular form of medium (i.e Instagram/Pinterest). Blog articles, like news articles, are usually presented in a broadcast format. This contrasts to other types of grey literature, for example a conversational format (Stack Overflow posts and comments/email chains), a documentation format (GitHub commit comments) and even a mixture of these formats (Twitter posts can be both broadcasts or conversational). Authors of personal blogs write blog articles on topics of their choosing. Some practitioners (e.g. https://devdactic.com/blog-as-a-software-developer/) advocate that all practitioners should write a blog as a form of contributing to their community. Chau and Xu [12] show that mining and analysing blog articles is a valuable method for carrying out research for marketing purposes and some software engineering researchers have acknowledged blogs in their studies, for example, Briand [9] references Bertrand Meyers' blog in discussing the impact of software engineering research on practice.

Blog articles are a widely available source that are often untapped by research. This is partly because there are problems with using blogs:

- There is no established process of factual verification such as in traditional news outlets.
- Unlike research outputs, blog articles vary in language formality and structure.
- There is no process for extracting the high quality articles from the vast quantity available.
- There needs to be a method for presenting large amounts of data back to researchers in a way that is usable but also shows the process taken for traceability and to support repeatability.

However, despite these challenges, there are advantages to using blogs as a source of evidence in research:

- Often a technology has to be developed, released and adopted by the community before data is available for researchers to analyse. Analysing practitioner blog articles allows researchers to gain insights earlier in the emergence of technology.
- Analysing what is said in practitioner blogs could potentially provide new insights into how decisions are made within industry. Sharma et al. [54] have studied email content to better understand how decisions are made in relation to Python Enhancement Proposals (PEP).
- Analysing blog articles for opinions on, and sentiment toward, new technologies over time would allow researchers to examine technology trends.

Garousi et al. [22] provide several reasons for utilising grey literature: grey literature provides current perspectives and complements gaps in the formal literature; grey literature may help avoid publication bias (although Garousi et al. acknowledge that the grey literature found may be not representative); and grey literature provides an important perspective on topics. We believe that all these reasons apply to blog articles: blog articles provide current perspectives on practitioners working in the topic being studied; blog articles can help avoid publication bias as they contain both positive and negative views on particular topics; and blog articles provide the practitioner perspective on topics.

1.4 Motivation

We are working towards a methodology for identifying blog articles that are of high quality for software engineering researchers. We have previously conducted a pilot study [63] where we looked at three credibility criteria (relevance, rigour and experience) for identifying these high quality articles based on previous research [49]. From this pilot study, we identified a need for a more formal review of the literature to justify our chosen criteria and find any new criteria. We constrain the scope of our literature review to any broadcast–oriented written online media so that we can gather criteria from related media, as well as blogs. (See appendix for further information.) A pilot literature review was also conducted, and this identified 'quality of writing' as a candidate criterion. Therefore, we begin this literature review with four existing criteria:

- Relevance: is the article relevant to the researcher's study?
- Reasoning: is the practitioner writing the article conveying some argument about the relevant topic?
- Evidence–backed (experience–backed): is the practitioner supporting her or his argument with evidence? We are interested in personal experience because Devanbu [14] and Rainer et al. [50] found that whilst researchers argue based on empirical data, software practitioners often form opinions based on their personal experience.
- Quality of writing: is the article well–written?

1.5 Research questions & objectives

Our aim in this paper is to identify criteria (and definitions) for evaluating the credibility of blog articles. We derive two specific objectives: 1) to review prior research literature to identify existing criteria; and 2) to complement the literature review with a short survey investigating software engineering researchers' opinions on evaluating the credibility of blog articles. The overall research question for this paper is:

- **RQ1: How do researchers assess the credibility of broadcast–oriented written online media?**

To address this question, we break it down into sub–questions:

- RQ1.1: How is 'credibility' defined in research?
- RQ1.2: What are the criteria used to assess the credibility of broadcast–oriented written online media?
- RQ1.3: To what degree to software engineering researchers identify with these criteria?

1.6 Contribution

We present a set of credibility criteria, derived from a literature review of software engineering and non–software engineering research, for assessing the credibility of broadcast–oriented written online medi. We also present a more specific and complementary set of criteria for blogs, derived from a survey of software engineering researchers. These criteria could potentially be used to evaluate the credibility of blogs for GLRs and MLRs, and could help to clarify and strengthen the classification schemes proposed by Adams et al. [2] and Adams et al. [1]. Our findings also corroborate previous research; that their is a lack of definitions for credibility and credibility criteria in research.

1.7 Structure of the paper

The remainder of this paper is structured as follows: in section two, we conduct a literature review of existing work on credibility and present the criteria that we believe hold value for researchers; in section three we validate our findings by conducting a survey of SE researchers; in section four, we present our synthesis of the findings from sections two and three and discuss the aims and objectives. Our conclusions, threats and future research directions are provided in section four.

2 Literature review

2.1 Rationale for the structured review

In this paper, we use the principles found in Systematic Literature Reviews (SLRs) and Systematic Mapping Studies (SMSs) to semi–systematically assess the literature to determine what criteria are widely used to assess the credibility of broadcast–oriented written online media. We then cross reference these criteria with how researchers assess the quality of their evidence to refine our generated list of criteria into a set of criteria that can be used to aid the extraction of high quality blog articles in future studies.

In Kitchenham et al's widely cited guidelines [30], they present three reasons for conducting systematic literature reviews (SLRs): to summarise existing evidence, the identify gaps in the research and to provide a framework for positioning new research.

Systematic mapping studies (SMSs) share similarities in their approach to systematic literature reviews. However, they differ in terms of their goals and outputs. Where SLRs aim to synthesise evidence, SMSs provide an overview of a research area. This overview includes details of the topics that have been covered in the research area and where they have been published [30, 46, 47].

The goals of our literature review differ from that of systematic literature reviews in that we are analysing the papers to extract specific information e.g. the credibility criteria used, the definition of credibility adopted, the user group operated on. However, the study cannot be categorised as a mapping study as we do more than map the literature. Also, given the relative maturity of our research, conducting an SLR would be premature as there is a bootstrapping problem that takes place i.e we need to understand the domain well enough to be able to design a SLR.

We instead opt for conducting a carefully structured literature review, adopting some SLR criteria and omitting others. Similar adaptations have been conducted by others. For example, Mantyla et al. [37] conduct what they call a semi–systematic literature review of the concept of rapid releases.

A more comprehensive literature review is one avenue for further research. We briefly return to this point later in the paper.

2.2 Review structure

2.2.1 Searching & filtering

Google Scholar was used to search for relevant articles. We kept our search criteria broad on purpose so that we can collect and report on a wide spectrum of candidate articles (Table 1). Our searches retrieved 833 results. After duplicate results are removed there are 762 results.

Criteria	Description
Timely	Published between 2007 and 2017 (the searches were carried out in early 2017). This is to ensure that we are looking at criteria that is currently used
Content	Content includes one of the following sources; blog, blogger, news, twitter, web, website, content, media, social media. Also must be written in English.
Verbs	Title includes one of the following verbs; assessing, measuring, evaluating, assess, measure, evaluate.
Keywords	Title includes one of the following keywords; credibility, credible, truthfulness, truthful, truth, believability, believable, belief

Table 1: The search criteria to be used.

The titles of all 762 results were read to identify those which appear relevant to assessing the credibility of online media. Where the relevance was unclear from the title, the abstract was read. If it still wasn't clear from the abstract (for example, where the paper was discussing the credibility of other media), then the introduction and conclusion were read. The filtering was conducted by the first author, consulting with the second author on any paper that was not easy to assess. After filtering, 142 papers remained.

In a second pass of these 142 papers, all abstracts were read. Where the abstract wasn't clear, the introduction and conclusion was read. If further clarification was needed then the entire paper was read. This second pass led to a further 29 papers being removed from the study (Table 2) and 113 papers going forward into the next phase.

Status	No. of results
Relevant	113
Not relevant	6
Citation	8
Cannot access	14
Not in English	1

Table 2: The results of the filtering after the second pass

2.2.2 Classification

The 113 relevant articles were then organised into three sets; those that are peer reviewed and that include an empirical study (n=36), those that are peer reviewed but do not include an empirical study (n=47) and those that are not peer reviewed (n=28). This literature review focuses on the peer reviewed articles that contain an empirical study as we are interested in the criteria that people use to assess credibility.

In due course, we intend to also assess the peer reviewed, non–empirical studies as most of these are automated assessments of credibility and will show us how the community maps from the criteria (this paper) to the physical measures used. The grey literature category (i.e. not peer reviewed) will act as a potential contrast data set.

The 36 articles that are peer reviewed and contain an empirical study were then classified by the data source which they analysed (e.g. blogs, news sites, forums). We then accepted or rejected classifications based on how related to blog articles they were (i.e. whether they were broadcast–oriented written online media or not). We rejected product reviews/online recommendations (n=4), YouTube/video (n=2), social media profiles (n=2), multiple media sources (n=2), forums/conversational (n=1), web searching e.g. Google search results (n=1) and generic websites e.g where the study says it anaylses 'the web' (n=12) as they are either too generic, or not the type of content that we are interested in i.e a written online article. We accepted blog articles (n=3), online news articles (n=3), Wikipedia (n=1) and websites that have explicit focus e.g. online health information (n=6). Each segment is explained in the Appendix (Table 20) with a justification of why it was accepted/rejected. Table 3 provides details of the number of papers under each segment. The final number of papers accepted for analysis was 13. Figure 1 provides an overview of the review process.

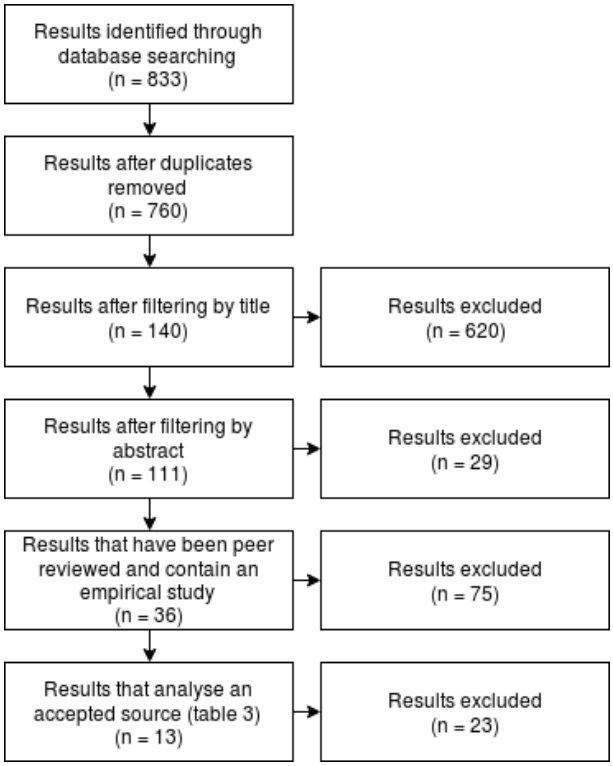

Figure 1: A summary of the literature search and synthesis process

2.2.3 Analysis

We reviewed the 13 papers identified from the four accepted classification sets of papers. For each paper, we identified the criteria that the paper used for credibility assessment (papers have been double bracketed for

Segment	# Results
Blogs	3
Online news	3
Wikipedia	1
Websites (explicit focus)	6
Product reviews/online recommendations	4
YouTube/video	2
Social media profiles	1
Multiple media sources	2
Forums/conversational	1
Websites (generic)	12
Web searching	1

Table 3: The number of results for each type of online media included

identification e.g. [[65]]. See Table 4 for citations to all 13 papers). In some studies, a list of criteria was initially reported and then the authors removed criteria due to their chosen user group finding it irrelevant or not as useful as other criteria for assessing credibility. For example, in [[4]], the authors narrow down their initially reported 31 criteria into its most 'parsimonious form' due to it being too much work for researchers using them in the future. At this stage, we have included these removed criteria as although the initial studies found them to be irrelevant, our user group (i.e researchers) may find them important to assessing credibility.

Year	Citation	Author(s)
2016	[58]	Wee–Khen Tan and Yun–Ghang Chang
2016	[65]	Quan Yuan and Qin Gao
2016	[4]	Alyssa Appelman and S Shyam Sunda
2015	[61]	Rita Zaharah and Wan–Chik.
2015	[60]	Oana Tugulea et al.
2014	[10]	D Jasun Carr et al.
2012	[44]	Katrina L Pariera
2011	[3]	Sharifah Aliman, Saadiah Yahya, and Syed Ahmad Aljunid
2011	[57]	Wee–Kheng Tan and Yu–Chung Chang
2011	[39]	Ericka Menchen–Trevino and Eszter Hargittai
2011	[55]	Beth St Jean et al
2010	[33]	Qingzi Vera Liao
2009	[66]	Dan Zhao, Chunhui Tan, and Yutao Zhang

Table 4: The 13 papers that were analysed.

2.3 Review results

2.3.1 Candidate credibility criteria

Overall, we identified 88 criteria from the 13 studies. We organised these criteria into the Source, Message, Channel, Receiver (SMCR) model presented by Berlo in 1960 [7]. However, this is for presentation purposes only as these are overlapping concepts [11] and without formal definitions being provided for each criteria in many of the studies, this classification is subjective to the opinions of the author. The full list of classified criteria are given in the Appendix (Tables 21, 23, 24 and 22 present the criteria that have been classified under the source, receiver, message and channel criteria respectively). As an example, the most frequently occurring criteria are: accurate, balanced, unbiased, reliable, trustworthiness, professional, complete, experience, fairness, credibility, well written, reputation, visual design, expertise, authority and believable (Table 5). Out of the 13 studies analysed, 7 provide no definitions at all for their criteria, 3 provide partial or implied definitions and only 2 of the studies provide definitions. However, even then, these definitions are by example. With no explicit definitions, it is difficult to determine whether and how the criteria overlap with each other, or whether there are duplicate criteria. For example, Tan and Chang [[57]] present 'unbiased' and 'fairness' as two separate criteria. However, these could be considered synonymous. In another example, Yuan and Gao [[65]] present 'equal,' 'neutral,' and 'objective.' This is a substantial finding of this review; that researchers need to formally define the criteria that they employ.

2.3.2 User groups

Due to the subjective nature of credibility, most credibility research is reported as being specific to a certain user group. For the analysis conducted in this paper, the most common user groups were students (and on one

Criteria	Frequency	Criteria	Frequency
accurate	8	balanced	3
unbiased	7	reliable	3
trustworthiness	6	professional	3
complete	5	experience	3
fairness	4	credibility	3
well written	4	reputation	3
design look	4	expertise	3
authority	3	believable	3

Table 5: A summary of the most popular criteria (n $\geq$ 3)

occasion, staff) from a university (n=6) and Amazons Mechanical Turk (n=2). This is probably due to these user groups being the most accessible to academic researchers. Table 6 provides the full list of user groups.

Study	User group
[58]	Students, between 18 and 25 years old and with experience reading travel blogs
[3]	academic staff and undergraduate students who are computer and internet literate, at the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA Johore (because the previous top management of that university used blogs as one of the alternative communication with the students and staff)
[57]	Taiwanese, 20 or above and general information consumers
[65]	Five experts in news information credibility assessment, including four Chinese editors of newspapers and one Journalism professor from Tongji University.
[4]	Amazons Mechanical Turk
[10]	Amazons Mechanical Turk
[39]	210 college students from two Mid–western US universities
[55]	29 content contributors out of the 333 participants from a previous study by Rieh et al. [52]. Indicated in their first study that they had either; created or posted original content to blogs or forums, commented on a blog or forum, rating, voting or tagging online content, or uploaded photos, music, video or items for sale. However, excluded contributors whose only contribution involved social networking sites and Twitter
[66]	consumers; that should have enough online shopping experience
[61]	users who have experienced accessing Islamic materials on the web
[60]	second and third year students
[33]	24 participants from a university community (12 younger, 12 older)
[44]	71 undergraduates enrolled in a psychology course at a private university in the eastern US

Table 6: The user groups used as participants for each of the studies analysed

Interestingly, none of the studies analysed used academic researchers as their user group. Appelman and Sundar [[4]] is the only paper that discusses the benefits that an effective list of credibility criteria can have on research. They describe how researchers in certain areas could benefit from an exclusive measure of message (content) credibility. However, missing from their list is an explanation of the effects that an exclusive measure of message credibility could have on the way that researchers collect and analyse evidence for their studies. This is ultimately an area in which we plan to explore throughout the longevity of our research.

2.4 Criteria source

The empirical studies from the 13 publications analysed typically fell under one of two categories; they had a pre–existing list of criteria from another study or literature review of many studies and then asked participants to judge them, or they created their list of criteria from the responses of their participants. However, as reported for each study's definition of credibility, some studies failed to report the source of their criteria, or only gave a partial or implied source (Table 7). Just as with the criteria, definitions of criteria and definition of credibility, explicitly stating the sources of the credibility criteria is imperative for the repeat–ability and systematic rigour of credibility studies.

2.4.1 Definition of credibility

As well as extracting the criteria from the 13 studies, we also make note of the definitions each study uses to define the term 'credibility' (c.f. the definitions of credibility criteria mentioned previously). Appelman and Sundar [[4]] identify a lack of formal definition for credibility as one of the major problems in credibility research. They also note that a concern is that some credibility studies fail to provide their definition of credibility. Our study makes

Year	Study	Criteria source(s)
2016	[58]	Chesney & Su, 2010 [13]
2016	[65]	Metzger, 2007 [40]
2016	[4]	Various and focus groups (pretest)
2015	[61]	interviews
2015	[60]	Manolica et al, 2011 [36]
2014	[10]	Meyer 1988 [41], Ognianova 1998 [42] and Petty and Cacioppo 1986 [48]
2012	[44]	Fogg et al, 2001 [21]
2011	[3]	None explicit but implied from literature
2011	[57]	None explicit but stated from "an extensive literature review"
2011	[39]	N/A - this paper did not conduct a study which dealt with credibility criteria
2011	[55]	unspecified and from participants
2010	[33]	Petty and Cacioppo 1986 [48]
2009	[66]	consumer questionnaire

Table 7: The originating source of the criteria used for each of the studies analysed

a similar observation; the lack of definitions. Of the thirteen studies analysed, two provide no definition at all. Others take their definition of credibility from other studies but there is still a great variation between studies that makes reporting objective results on a combination of these studies difficult (Table 8).

Year	Study	Credibility definition (quoted from each paper)
2016	[58]	Credibility is often defined as believability, trust, reliability, accuracy, fairness, objectivity and many other concepts and combinations thereof [53]
2016	[65]	is what degree a person believes the information is true when receiving it. It is content–based, but different from different people and different environment, and easily affected by kinds of reasons
2016	[4]	message credibility is an individuals' judgment of the veracity of the content of communication
2015	[61]	None provided
2015	[60]	information that can be trusted, believed to be secure [20, 59]
2014	[10]	the assessment of believability and trustworthiness of a message based on a multitude of factors involved in communication, such as message source, content and the medium through which the message is presented
2012	[44]	early research on how people assess credibility defined the concept simply as trustworthiness, expertise and believability [29], terms which are still used by credibility researchers today [19]
2011	[3]	Credibility is a multidimensional construct that represents a composite of several characteristics that perceivers perceive in a source [20, 25, 38]
2011	[57]	Credibility is a complex concept with "dozens of other concepts and combinations" [53]. Conceptually, credibility is often classified as source, message and medium credibility [16]. Though conceptually tidy, credibility dimensions may overlap [11]. Information consumers often do not differentiate between these dimensions [18]. Many studies also do not make clear distinctions among these dimensions or focus on one or two of them selectively.
2011	[39]	we see credibility as believability because this provides an operational definition of trust in an information–seeking context, information that the respondent believes
2011	[55]	[51] defines credibility as "people's assessment of whether information is trustworthy based on their own expertise and knowledge" (p. 1338). Under this definition, people ultimately recognize and make judgments about information credibility rather than being led to make such assessments by specific characteristics of an information object, source, or person. In this paper, it is presumed that information credibility judgments are highly subjective and entail multidimensional assessment processes.
2010	[33]	None provided
2009	[66]	Enterprise website credibility is: the degree of trustworthiness of one enterprise website which concluded by the online consumers, from their own psychological point of view and based on their own experiences.

Table 8: Credibility definitions from each of the 13 studies analysed

The problem is that credibility is a complex concept that is subjective to each individual at each specific moment in time [66]. For example, an article that has been previously disregarded by a researcher may be seen as

more credible once emailed to them by a colleague. This problem with subjectivity within the research has been acknowledged in ten of the thirteen studies analysed (Table 9). The de facto standard for dealing with this in credibility research has been to report credibility criteria for a specific user group [[3]]. However, even this level of granularity can often be subjective within itself.

Year	Study	Mention of subjectivity (quoted from each paper)
2016	[58]	credibility is a subjective receiver–based construct rather than an objective measure of the actual quality of the information [27].
2016	[65]	Information credibility is what degree a person believe the information is true when receiving it. It is content–based, but different from different people and different environment, and easily affected by kinds of reasons.
2016	[4]	It is worth noting that two of these three measures, accuracy and authenticity, could be considered to be more objective, whereas the third, believability, could be considered to be more subjective. Because the proposed measure is based on self–report perceptions, these measures are all, in fact, subjective. In other words, we could view the three indicators as perceived accuracy, perceived authenticity, and believability.
2015	[61]	No explicit mention, but it is implied from their results (e.g. "They also mentioned that there is a need to have prior knowledge on Islam or the Quran for one to be able to better evaluate the retrieved information").
2015	[60]	No explicit mention, but it is stated that different contexts favour different dimensions of credibility; "The credibility's dimensions in various contexts are identified using the exploratory factor analysis."
2014	[10]	Not discussed
2012	[44]	Studies in audience credulity focus on characteristics of the audience that affect their subjective assessments of credibility.
2011	[3]	Credibility is not only perceptual phenomenon [26, 43], but also situational or contextual phenomenon [26]. Besides, credibility is dynamic or can change over time [26]
2011	[57]	This study also suggests that there is no "one size fits all" answer as to how information consumers assess credibility. It will depend on the types of information source and even travel blogs cannot be viewed as a monolithic type.
2011	[39]	Not discussed
2011	[55]	In this paper, it is presumed that information credibility judgments are highly subjective and entail multidimensional assessment processes.
2010	[33]	Not discussed
2009	[66]	Since the credibility of enterprise website is a sort of subjective feeling, different people will have different interpretations from different perspectives.

Table 9: Mentions of credibility being subjective for each of the 13 studies analysed

2.5 Aligning our criteria with the requirements of software engineering researchers

Now that we have our list of 88 criteria, we can start to cross reference them with the requirements that researchers have for the evidence that they use in their studies. This exercise is again subjective to the views of the author and therefore we complement our findings with the survey explained in the next section.

Although, like Chaffe [11], we acknowledge that there is an overlap between the different categories of credibility criteria (source, message, readers and medium). We begin by ignoring any criteria that falls under the *medium* category as this category is well–defined (i.e. any criteria that is a attribute of the *medium* itself and not related to what is being written on the site), and where blogs have been previously classified as a less credible source of media [35], our research looks to argue that certain types of blog articles do potentially hold value in research. Furthermore, when using grey–literature, research cannot disregard any based on the type of media under which it is published as it is the content that holds the value, not the location of where that content has been published.

Using the *source* criteria to judge an articles' credibility is also problematic. Of course, the majority of the general population is likely to find a blogger with many qualifications and years of experience in a relevant field more credible than someone who has recently started out in the industry, but that does not necessarily mean that this new starter cannot write credible content that provides new insights to research. Conversely, it also does not mean that the more experienced practitioner is writing content that is insightful. The same can be said of 'engagement.' An author that engages often with their audience through social media does not necessarily produce more credible insights than the author who writes few blog posts sporadically and with little other interaction with their readership.

Another issue with the *source* criteria presented is the subjectivity of some of the criteria. Reputation may for example be measured through the number of active subscribers to a particular blog. However, a large audience is not necessarily an indication of insightful valuable opinions as different readers may subscribe for different

reasons e.g. humour. Ideally, we would want to measure the reputation in terms of the articles value to research. Traditionally in research, this is often assessed through the citation count of peer reviewed articles. However, with no such measure in place for blog articles, this becomes difficult (we assess the use of URL citations to research and practitioner sources in a different study [62]). For these reasons, we also ignore the source criteria from this study. This leaves only the *message* and *receiver* criteria to be analysed.

For the *receivers* criteria, we create three sets; the prior beliefs of the reader, the influence that others have over the reader and the relevance to what the reader is looking for. For the *message* criteria, we create five sets; supported by evidence (e.g. experience, URL citations), quality of writing, strength of argument, prior beliefs of the reader and the influence of others on the reader (Table 10). Here you can see an overlap between the different categories of credibility criteria as noted by Chaffe [11].

Judgments of the individual reader	**Judgments of others who then recommend the content**	**Relevance to what the reader is looking for**
past experience with site	recommended	relevance
general suspicions	endorsed	cue in the content
general dislike		usefulness
aligns with own knowledge		
location of user		
trust		
name recognition		

Table 10: The three abstract groups of receiver (reader) criteria

Any system that looks towards aiding researchers with assessing the credibility of broadcast–oriented written online media is not applicable to the first group (the judgments of the individual reader) as this is what such a system is trying to achieve. The same can be said of the second group (the judgments of others who recommend the content) as there is no way to monitor these recommendations, and the ultimate aim of such a system would be to recommend articles to the reader based on other criteria. However, relevance is a major factor in selecting good articles for research. An article may contain highly credible and valuable insights, but if it is not relevant to the research of the reader then it is not useful. In previous pilot studies, we have turned to topic detection and utilising existing search engines to determine relevance [63].

As with the receiver criteria, the message criteria can be organised into abstract groups. In the case of message criteria, we have identified five abstract groups: the content is supported by evidence, the quality of writing, the strength of the argument, the judgment of the individual, and the judgment of others (Table 11). Interestingly, here we see some of the overlap that we have mentioned previously. Judgment of the individual and judgment of others are abstract groups of the reader criteria also. However, the criteria mentioned here are concerned with the content (message) and not with the initial desires of the reader prior to reading the article (intrinsic plausibility, believability, will have impact and popularity). This supports existing credibility assessment models such as the Elaboration Likelihood Model (ELM) designed by Petty and Cacioppo in 1986 [48]. ELM states that there are two routes to assessing credibility; peripheral cues and central cues. St Jean et al. [55] found that participants often favour heuristic credibility judgments (such as the peripheral cues in ELM) and then transition to strategy–based credibility judgments if heuristics are insufficient for the situation. Furthermore, they found that heuristics play an important role when making a predictive judgments of credibility. Conversely, Pariera [44] found that people are more likely to use central cues when they have a greater stake in the argument, are knowledgeable about the argument, and are motivated and able to process the information. The study found that students look first for textual cues and then supplement these credibility judgments through visual (peripheral) cues. Like the students in [44], researchers should also be more concerned with these central cues when assessing credibility.

The remaining abstract groups; supported by evidence, quality of writing and argument strength are also features important to researchers. Wohlin [64] states that software engineering should be evidence based but acknowledges that it is a challenge to synthesise the evidence available, even in tightly controlled experiments. He presents a series of criteria for good research evidence that we think also needs to be considered when analysing blogs for researchers:

1. Quality of evidence (how reliable is the evidence?)
2. Relevance of evidence
3. Aging of evidence (evidence that has aged too much may no longer be relevant)
4. Vested interest (is the evidence unbiased?)
5. Strength of evidence

Similarly to this, Fenton, Pfleeger and Glass [17] provide 5 questions that should be asked about any claim made in software engineering research:

Supported by evidence	**Quality of Writing**	**Strength of argument**	**Judgments of the individual**	**Judgments of others**
truthful	writing tone	argument strength/content	intrinsic plausibility	popularity
authentic	well written	reliable	believable	
experience	update	comprehensive	will have impact	
cite external source	corrections	consistent		
trusted sources	authority	detailed		
multiple sources	error–free	credibility		
verified	sincere	representative		
cited	etiquette	balanced		
accurate	professional	equal		
currency	focus	neutral		
transparent	clarity	objective		
trustworthiness	motivation	not opinionated		
honest		unbiased		
factual		complete		
		fairness		
		truth–seeking intentions		
		spin–free		
		partisan nature		

Table 11: The five abstract groups of content (message) criteria

1. Is the claim based on empirical evaluation and data?
2. Was the empirical study designed correctly?
3. Is the claim based on a toy or real situation?
4. Were the measurements used appropriate to the goals of the empirical study?
5. Was the empirical study run for a long enough time?

Together, these two sources provide an indication of researchers' requirements of evidence and the inferences from evidence.

Quality of writing is also an important when assessing online credibility. Appelman and Sundar [[4]] found that writing quality "contributes significantly to perceptions of message credibility," and a study by Bird, McInerey and Mohr [8] found that writing quality was the most important factor in credibility assessments. Aliman et al. [[3]] found that writing tone was one of the common agreeable criteria in their study. In research, high quality writing is important to ensuring that evidence is unambiguous. Quality writing can also be an indication of the professionalism of the author. However, this is not always the case. Blogs are supposed to be a more relaxed form of discourse and as a result, the language is often informal and relaxed. For example, Joel Spolsky's blog, Joel on Software[1] contains many articles of varying writing quality. One article[2] contains 67 words, 66 of which are the word 'Dave' (the owner of the software company which founded Trello, and CEO of Stack Overflow).

The final group, 'strength of argument' is another group that is important in research as it is a measure of the rigour of the article. Argumentation mining is an emerging community in natural language processing research which aims to identify arguments from text and the relationships which the arguments have with each other (e.g. premise A and premise B lead to conclusion C). Given the subjectivity of credibility, argumentation mining is unable to provide researchers with articles that are of high value. However, presenting these identified argument maps to the reader can allow for quick assessment of the credibility and help the reader determine whether the article is worth considering further.

Within the criteria of the argument strength group, there is another sub–group that may also aid this decision making process around the fairness of the arguments presented (balanced, equal, neutral, objective, not opinionated, unbiased, complete, fairness, truth–seeking intentions, spin–free, partisan nature). Some of the criteria in this sub–group may appear synonymous with each other, but without any formal definitions within the analysed texts, there is no way to confidently group these together. In our previous studies [63], we have turned to indicator words to identify the presence of reasoning (e.g. the word 'because' always indicates that a reason will follow, 'therefore' always indicates that a conclusion will follow). However, we have identified that this method alone is too arbitrary. This literature review has identified another aspect of argumentation mining previously not

[1] https://www.joelonsoftware.com/
[2] https://www.joelonsoftware.com/2002/01/10/20020110/

Table 12: Final criteria from the literature review

Final criteria from the literature review
Relevance
Strength of argument
Evidence backed
Quality of writing
Prior beliefs of the reader
Prior beliefs of others who influence the reader

considered; that some check needs to be carried out to ensure the arguments presented are not biased. Previous research in natural language processing has looked at identifying bias in text. Doumit and Minai [15] use topic detection and NLP techniques to identify bias in online news media.

Overall, the literature review has found multiple criteria which we have then organised into groups and refined into the criteria that we believe are important to researchers. These are re–stated in the Table 12 for clarity. With this organisation and refinement being a subjective task, we next conduct a survey to complement our results.

3 Credibility Survey

3.1 Survey Design

3.1.1 Overview

To complement the findings and criteria (Table 12) from the literature review, we have conducted a survey of software engineering researchers to investigate how they would assess the credibility of practitioner written blogs.

The survey was run from the 13th February 2018 until the 26th of March 2018. This paper focuses on the quantitative data. An in–depth analysis of the qualitative data will be considered in a future publication. The full list of questions have been published online[3]. Respondents are asked for the number of years spent conducting research and their specific areas of research within software engineering. We ask a series of questions about how the researcher would assess the credibility of practitioner blogs and ask for the respondents contact details, whether they are willing to be contacted for a follow up interview, and whether they would like to receive an anonymised copy of the results on completion of the survey.

3.1.2 Development & refinement

The survey was first developed and sent to four colleagues that were familiar with the research for review. After making changes following the feedback from this review, we conducted a pilot study, inviting responses from a network of software engineering researchers within New Zealand (SIˆNZ[4]).

The SIˆNZ community comprised 27 members. The four colleagues from our internal review were removed so we had 23 possible participants. Of these, we received 8 fully completed responses. The feedback from the pilot study led us to clarifying some questions e.g. changing our Likert scales to an odd number so that participants could portray a neutral response. We also added an 'I don't know' option to these questions.

The survey instrument used was a online survey tool (Qualtrics[5]) and an anonymous link was provided in the invitation email. The survey was approved by the University of Canterbury under ethics reference: HEC 2017/68/LR-PS.

3.1.3 Participants

Invitations were sent out to the Programme Committees of two leading international software engineering conferences (in 2018), Evaluation and Assessment in Software Engineering (EASE) and Empirical Software Engineering and Measurement (ESEM). Members of each Programme Committee were removed from invitation if they were part of Software Innovation NZ due to their involvement with the pilot study.

Overall, 138 researchers were invited to participate. Four of these invitee's asked us whether they could forward the survey to colleagues. We approved these requests but did not track them, meaning that we are unable to report the total number of people who received the invitation. 57 invitees started the survey, but only 44 completed it, giving a response rate of 31.9%. The participants' experience in research ranged from 2 years to 35 years, with a mean average of 16.2 years. A summary of respondents research interests is given in Table 13.

The total time taken to complete the survey ranged from 2.4 minutes to 22 hours with an overall average of 75.7 minutes. Ignoring the completion times of the five responses that took longer than one hour to complete gives a range from 2.4 minutes to 47.1 minutes, with an average time of 11.7 minutes.

[3] https://www.researchgate.net/publication/324784268_Design_of_a_survey_on_credibility
[4] http://softwareinnovation.nz/
[5] https://qualtrics.com

Research interest	Frequency
mining and analytics	17
testing	15
empirical SE	14
human factors	14
other	13
requirements & requirements engineering	12
quality	10
software processes	9
agile	8
research	6
metrics	5
software engineering	4
EBSE	3
maintenance	3
evolution	2
global software development	2
open source	2
project management	2
security	2
software product	2
technical debt	2
usability	2
behavioral software engineering	1
programming	1
risk	1
safety	1
startups	1

Table 13: Summary of respondents' research interests

3.1.4 Post–survey follow–up

On closing the survey, we followed up with invitee's with an email asking to those who did not start the survey, or started but did not complete, the reasons why. A similar follow–up was also conducted with the SIˆNZ trial. The main reason given by both follow ups was that respondents were too busy to start the survey.

3.2 Results

3.2.1 Outliers

We identified one outlier from the survey responses. One respondent indicated that "... it is simply impossible to evaluate the value [of blog articles] since no real evidence is provided..." and "... if we start trusting blogs, we might as well stop doing scientific research in software engineering...". These qualitative comments are consistent with some of the respondent's answers to closed questions e.g. the respondent provided a score of 0 (no blogs are credible) when asked to provide their general credibility assessment of blog articles. However, these qualitative comments are not consistent with the way that the respondent scored all of the criteria of credibility i.e. when asked to score the importance of a given set of criteria for assessing credibility, the respondent provided a maximum score of 1 on a Likert scale of 0–6. One of the criteria was whether the blog article provides any empirical evidence. Given the respondent's comment on the lack of real evidence in blogs, we would expect a high/er score on the credibility score. A possible explanation for scoring all questions with a low score is that the the respondent has misunderstood the survey. On this basis we removed this response from our sample, so that we had 43 respondents. Removing the respondent from our results substantially alters the statistics (e.g. mean, standard deviation, and the minimum values).

3.2.2 General credibility assessment of blog articles written by practitioners

The results in Figure 2 indicate that at least some blog articles have a sufficient degree of credibility. Qualitative comments from respondents suggest that respondents recognise blog articles can vary considerably in their credibility. For example, 37 of the 43 respondents provided comments to support their answer. Of these 37, 10 of them indicated that "it depends" (for example, either on the topic, subject matter, author). We intend to explore these qualitative comments in more detail in a future paper.

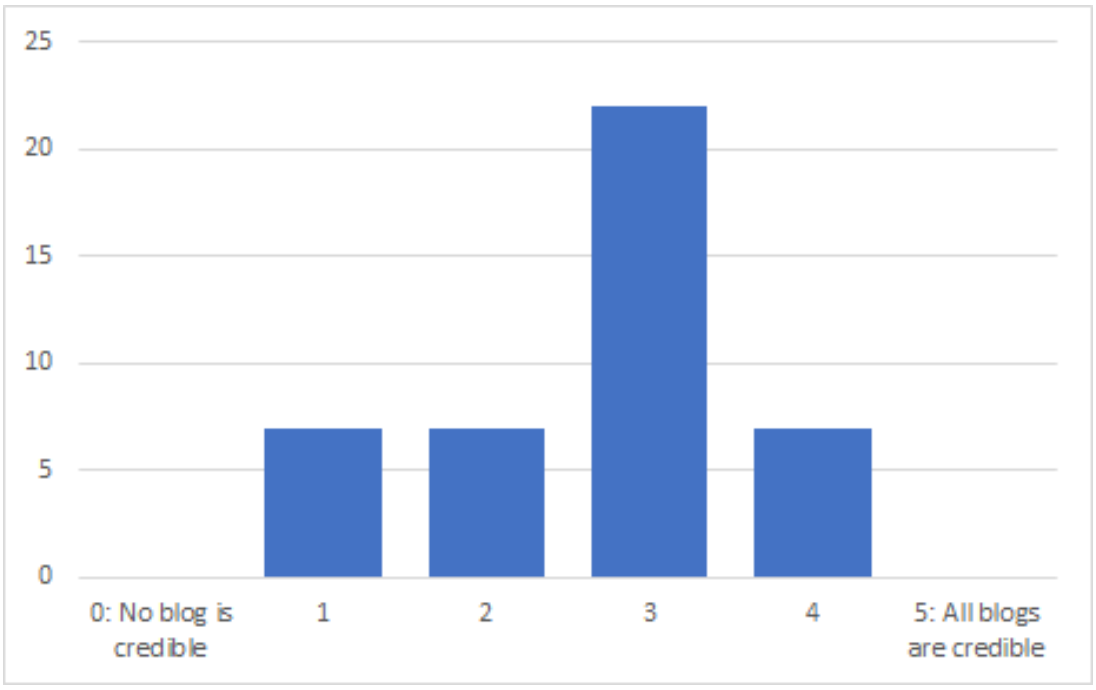

Figure 2: Researchers assessment of the general credibility of practitioner blogs (n=43, Mode=3, Median=3, Mean=2.7).

3.2.3 Criteria used for assessing the credibility of blog articles

Respondents were presented with a set of criteria that the literature review found important for assessing the credibility of online social media. For each criteria, respondents were asked to rate the importance on a 7 point Likert scale, ranging from 0 (not at all important) to 6 (extremely important). The results are provided in Figure 3 and Table 14. The results presented in Table 14 identify criteria to use. Note that the table is not saying that blog articles are generally well–reasoned, or have empirical data etc.

Figure 3: Researchers assessment of the importance of individual credibility criteria (n=43).

Table 14: Researchers assessment of the importance of individual credibility criteria on a scale from 0 (Not at all important) to 6 (Extremely important) (n=43)

		0	**1**	**2**	**3**	**4**	**5**	**6**	**I don't know**
Clarity of writing	***f***	0	0	2	8	7	13	13	0
	%	0.0	0.0	4.7	18.6	16.3	30.2	30.2	0.0
Reporting of empirical data	***f***	0	0	1	3	11	14	14	0
	%	0.0	0.0	2.3	7.0	25.6	32.6	32.6	0.0
Reporting of the method for data collection	***f***	0	1	2	4	13	11	12	0
	%	0.0	2.3	4.7	9.3	30.2	25.6	27.9	0.0
Professional experience	***f***	0	1	2	6	9	16	9	0
	%	0.0	2.3	4.7	14.0	20.9	37.2	20.9	0.0
Web links to other practitioner sources	***f***	1	2	4	5	12	14	4	1
	%	2.3	4.7	9.3	11.6	27.9	32.6	9.3	2.3
Web links to peer–reviewed research	***f***	1	2	4	1	10	19	6	0
	%	2.3	4.7	9.3	2.3	23.3	44.2	14.0	0.0
Reasoning	***f***	0	0	1	3	5	15	17	2
	%	0.0	0.0	2.3	7.0	11.6	34.9	39.5	4.7
Prior beliefs of the reader	***f***	6	3	3	9	7	7	3	5
	%	14.0	7.0	7.0	20.9	16.3	16.3	7.0	11.6
Influence of others on the reader's beliefs	***f***	5	5	9	5	9	7	3	0
	%	11.6	11.6	20.9	11.6	20.9	16.3	7.0	0.0

Ranking the criteria Our literature review identified a set of criteria to consider when evaluating the credibility of broadcast–oriented written online media. A natural question to consider is whether some of these criteria, or some combination of criteria, are more important for credibility. Or, alternatively, whether the presence of particular criteria are essential to demonstrating credibility. The Likert–scale data we have (Table 14) allows us to examine the frequency with which a criteria was considered to be important. We calculated three statistics: the mean, the median, and the percentage of respondents who ranked the criteria as extremely important (a value of six on our seven–point Likert scale). Looking at the percentage of respondents who ranked the criteria as extremely important provides an insight into criteria that may be essential for credibility. Table 15 presents the three statistics. The criteria in Table 15 are ranked according to the percentage of respondents who ranked that criteria as extremely important. We recognize the need to be cautious here however. The statistics presented in Table 15 are estimates drawn from one survey sample. The relative rankings could change based on a different sample.

The rankings presented in Table 15 raise three interesting observations. First, there appears to be several clusters of criteria: Reasoning (on its own); then the Reporting of empirical data, Clarity of writing, and Reporting of the methods for data collection; then Professional experience (on its own); linking to others sources; and finally, two categories relating to Beliefs. We noted earlier that the statistics are estimates. The clusters suggest there could be greater variation within a cluster rather than across clusters. In other words, that the rankings for the criteria of Reporting of empirical data, Clarity of writing, and Reporting of the methods for data collection may change, but Reasoning would be more likely to remain ranked higher than Reporting of empirical data, Clarity of writing, and Reporting of the methods for data collection.

Second, it is surprising how low (some of) the percentages are e.g. 'only' 32% of researchers consider the Reporting of empirical data to be extremely important. This is surprising for an empirical disciplines. Perhaps the 'low' percentage suggests that respondents consider that none of these criteria is essential to credibility, though some criteria are clearly more important for credibility.

Third, that Reasoning is the criteria most frequently considered extremely important. Software engineering researchers place considerable emphasis in their publications on evidence and empirical data. It is surprising that the quality of reasoning receives relatively little explicit consideration in software engineering research. One possible explanation here is that researchers may expect to see more reasoning, and better quality reasoning, in blog articles in the absence of empirical data.

Devanbu et al. [14] and Rainer er al. [50] found that software engineering practitioners formed opinions based on their own personal experience over empirical data. Prior beliefs and influence of others ranks low on our criteria. This contrasts noticeably with Devanbu et al. [14] who found that practitioners rank the influence of others highly as a source of credible information.

The criteria of: presence of reasoning, reporting of empirical data, clarity of writing and reporting of the method for collecting data all rank highly as important criteria. The presence of web links to research sources ranks higher than the presence of web links to practitioner sources. This is to be expected given that researchers completed the survey. The reporting of professional experience ranks fifth in our list (both by Mean ranking and by ranking on the percentage of respondents who scored the criteria as 'extremely important'). Prior beliefs and influence of others ranks low on our criteria. This contrasts noticeably with Devanbu et al. [14] who found that practitioners rank the influence of others highly as a source of credible information.

Table 15: Summary statistics and rankings for credibility criteria.

	Statistics					Rankings		
	Me	**Mo**	**Med**	**SD**	**%(6)**	**Med**	**Me**	**%(6)**
Reasoning	5.1	6	5	1.0	38.6	1	1	1
Reporting of empirical data	4.9	6	5	1.0	31.8	1	2	2
Clarity of writing	4.6	5	5	1.2	29.5	1	3	3
Reporting of the method for data collection	4.6	4	5	1.3	27.3	1	3	4
Professional experience	4.5	5	5	1.2	20.5	1	4	5
Web links to peer–reviewed research	4.3	5	5	1.5	13.6	1	5	6
Web links to other practitioner sources	4.0	5	4	1.4	9.1	2	6	7
Prior beliefs of the reader	3.1	3	3	1.9	6.8	3	7	8
Influence of others on the reader's beliefs	3.0	2	3	1.8	6.8	3	8	8

Me: Mean
Mo: Mode
Med: Median
SD: Standard deviation
%(6): Percentage of respondents rating the criterion as 6 *Extremely important*

Table 16 presents Spearman's rank order correlations for the nine criteria. In general, the criteria do not correlate with each other, which suggests we have independent constructs. Two pairs of criteria have strong correlations: *Reporting data* and *Reporting methods of data collection*, and *Prior beliefs* and *Influence of others*. For the *Reporting data* and *Reporting methods of data collection* criteria, the field of software engineering research recognises these as separate constructs, albeit closely related to each other. For the criteria of *Prior beliefs* and *Influence of others*, the respondents appear to consider both of these criteria as not important (see for example the statistics reported in Table 15).

Criteria		1	2	3	4	5	6	7	8	9
Writing	1	1	0.02	-0.13	0.01	0.03	0.09	*0.25*	-0.05	0.08
Data	2		1	**0.74**	-0.09	0.01	*0.27*	0.01	*-0.20*	*-0.21*
Method	3			1	0.03	0.20	*0.30*	0.04	-0.13	-0.24
Experience	4				1	*0.28*	-0.11	0.11	0.15	0.14
Practice	5					1	**0.55**	0.13	*0.25*	*0.27*
Research	6						1	*0.25*	-0.15	-0.01
Reasoning	7							1	-0.18	0.11
Beliefs	8								1	**0.78**
Others'	9									1

Table 16: Spearman rank correlations

3.3 Perceived general credibility vs. criteria importance

We checked whether the respondents answers to the general credibility of blog articles are associated with the scores those respondents gave to rating the importance of credibility criteria. In other words, is the general assessment of the credibility of blog articles associated with the scores given when rating each of the credibility criteria? Our hypothesis is that the answers are not associated. To answer our question and test our hypothesis, we organized our data into three subsamples (see Table 17) and present descriptive statistics for these subsamples in Table 25. (After removing our outlier, no respondents gave a general credibility rating of 0 or 5).

Due to the small sample sizes for subsample 1 and 3, we cannot with confidence conduct statistical tests comparing these samples e.g. Mann Whitney. Instead we simply compare the descriptive statistics. The statistics in the Appendix (Table 25) suggest that respondents answers to the general credibility of blog articles are not associated with the scores that those respondents gave to assessing the importance of credibility criteria. For example, for the majority of the criteria, the mean averages are similar across the three profiles, as are the

Table 17: The three subsamples, grouped by respondents general credibility score

Subsample	General credibility score	Sample size
1	0 or 1	7
2	2 or 3	29
3	4 or 5	7

		General credibility of blog articles												
		0		1		2		3		4		5		
Generalise	**Total**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	**Total %**
Yes	26	0	0	4	9.30	3	6.98	16	37.21	3	6.98	0	0	60.47
No	6	0	0	1	2.33	2	4.65	1	2.33	2	4.65	0	0	13.95
It depends	11	0	0	2	4.65	2	4.65	5	11.63	2	4.65	0	0	25.58
Total	43	0	0	7	16.28	7	16.28	22	51.16	7	16.28	0	0	100

Table 18: Does the model generalise to other practitioner–generated online content? (n=43)

		General credibility of blog articles												
		0		1		2		3		4		5		
Generalise	**Total**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	***f***	**%**	**Total %**
Yes	25	0	0	3	6.98	4	9.30	14	32.56	4	9.30	0	0	58.14
No	10	0	0	1	2.33	2	4.65	5	11.63	2	4.65	0	0	23.26
It depends	8	0	0	3	6.98	1	2.33	3	6.98	1	2.33	0	0	18.60
Total	43	0	0	7	16.28	7	16.28	22	51.16	7	16.28	0	0	100

Table 19: Does the model generalise to researcher–generated content? (n=43)

standard deviations. This indicates that respondents are in broad agreement of the value of the credibility criteria, and that agreement is separate from their opinion on whether the blog articles are actually credible.

In the table, about 71% of respondents in subsample 3 have rated 'reasoning' as extremely important (c.f. 43% in subsample 1 and 31% in subsample 2). This indicates that for researchers who find blog articles credible the amount of reasoning presented within blog articles is an important factor for that article's credibility."

3.4 Generalisation of the criteria

We are also interested in the degree to which the criteria could be used for assessing other kinds of practitioner–generated and researcher–generated content. We therefore asked the following two questions in the survey:

1. Do you think that the criteria identified generalise to assessing the quality of content written by practitioners, other than blogs e.g. emails, Q&A sites such as Stack Exchange, comments that have been provided in response to blog articles?

2. Do you think the criteria identified generalise to assessing the quality of content written by researchers e.g. journal articles, conference papers?

For both questions, the available answers were *In general, yes*, *In general, no*, or *It depends (please explain below)*.

Table 18 indicates that over 60% of respondents thought that the criteria generalise to other practitioner–generated content. Surprisingly, a very similar percentage (over 58%) think the criteria also apply to researcher–generated content too (see Table 19).

4 Discussion & conclusions

4.1 Addressing the research questions

In the introduction, we state our overriding research questions to be:

- **RQ1: How do researchers assess the credibility of broadcast–oriented written online media?**

In order to answer this, we break the question into three sub–questions:

- RQ1.1: How is 'credibility' defined in research?

A common theme throughout this study and the existing credibility research is the subjectivity of the credibility assessment. Previous research has accounted for this subjectivity by reporting on the credibility assessment of particular user groups. Therefore, much of the research reported around credibility assessment is specific to a particular group (e.g. senior citizens, college students, visually impaired). However, credibility assessment is actually subjective to each individual person in each individual circumstance. Two researchers may assess the credibility of a given article differently and draw different conclusions on the articles credibility given their own views and previous experience. Similarly, a single researcher may judge the same article differently at two different moments in time.

It is widely acknowledged that credibility assessment is subjective, but this study has identified that a major problem with existing credibility research is that it too can be interpreted subjectively in parts. For example, the

majority of research analysed provides no formal definitions for its criteria, leading to ambiguous interpretations of what the criteria is. We make the following four observations:

1. The term 'credibility' is interpreted subjectively. Some studies do not provide their definition for 'credibility.'
2. Some studies do not give detail of how their credibility assessment was conducted, or do not state the criteria that they have used for assessing credibility.
3. None of the studies analysed provide formal definitions for their credibility criteria. This ambiguity can cause problems for studies as participants may interpret criteria differently.
4. Some studies do not report the original source of their criteria (i.e where they got their list of criteria from).

Our literature review found no study that provides formal definitions for their criteria, but there are definitions given for the general concept of credibility. Research needs to work towards a formal definition for credibility so that the concept can be discussed and assessed more objectively. A possible method for doing so, may be to review the current definitions and analyse the similarities and contrasts. In this study we have gathered the definitions of 11 publications (Table 8) to serve as a starting point and example (two of the papers we analysed provided no definition). We leave a formal review and analysis for future research, but from our gathered definitions, we can see that 'believability' and 'trustworthiness' are the two most common descriptors for credibility.

- RQ1.2: What are the criteria used to assess the credibility of broadcast–oriented written online media?

In this study, we have collated the criteria presented in 12 publications (1 of the 13 publications analysed presented no criteria). These are presented in the Appendix as Tables 21, 23, 24 and 22. From this collation, we have presented the most frequently occurring criteria (Table 5). These are; accurate, unbiased, trustworthiness, complete, fairness, well written, design/look, authority, balanced, reliable, professional, experience, credibility, reputation, expertise, believable.

However, although we believe our list of criteria to be extensive, we recognise that 13 papers is a small subset of the research reporting around the credibility of broadcast–oriented written online media. Broadening our search criteria would have found more publications and possibly more criteria to consider.

- RQ1.3: To what degree to software engineering researchers identify with these criteria?

Given our collated criteria that were gathered for RQ1.2, we next grouped and refined the criteria down to those that are relevant to researchers. These are presented in Tables 10 and 11.

After refining, we find that the remaining criteria for assessing the credibility of broadcast–oriented written online media all tend to fall under one of six criteria; the strength of the message/argument, whether the article is supported by evidence, the quality of the writing, the judgment of the individual and the judgment of others (who recommend and as a result, influence the judgment of the individual). Of course, the article has to also be relevant to the researcher (Table 12).

Within these categories however, researchers may weight the importance of the criteria depending on the nature of the research being undertaken. For example, research looking at recent phenomenon may favour the timeliness of the data over other criteria as a measure of relevance. Relevance is still important however, as the evidence still needs to be on topic. Despite this subjectivity within the categories, all categories remain important and must be considered when assessing the credibility of broadcast–oriented written online media.

Following our literature review, we conducted a survey to discover how researchers assess the credibility of blog articles. We presented the criteria from the literature review and asked them to score the importance of each criterion. From this we are able to rank criteria based on their perceived importance. The importance scores are provided in Table 14 and the ranking is given in Table 15. Over 60% of the survey participants think that the criteria generalise to other practitioner–generated content and over 58% of the survey participants think that that criteria generalise to other researcher–generated content.

4.2 How does our research relate to how researchers currently assess evidence?

Kitchenham, Dyba and Jorgensen [31] suggested in 2004 that software engineering would benefit from being evidence–based. This evidence–based and systematic approach to research originates from medical sciences and promotes rigour and thoroughness throughout. Wohlin agrees that software engineering research should be evidence based, but acknowledges that it is difficult to synthesise evidence, even in tightly controlled experiments [64]. He presents five criteria for evaluating evidence and finding high quality evidence for research; the quality of the evidence, the relevance of the evidence, the aging of the evidence, the vested interest of the source and the strength of the evidence.

Fenton, Pfleeger and Glass report five questions that should be asked about any claim made in software engineering research [17]. The first of which asks whether the claim is based on empirical evaluation and data, thus supporting Kitchenham et al's proposition for evidence–based research ten years before its publication. The remaining four questions are about the particulars of the study to assess the trustworthiness and rigour of the evidence that the study presents.

Together, the two papers specified above provide a general idea of the requirements that researchers place on the evidence they use in their studies. Evidence must be rigourous, reliable, unbiased and relevant. However, a more formal review is needed to ensure that these requirements generalise to all research. This review is left for future research.

4.3 Threats to validity

There are several threats to validity of this study. During the literature review, our sample size for analysis only included 13 publications. This could be due to the strict search criteria and rejection criteria that we initially developed for collecting relevant publications. As a form of validation, we conducted a single iteration of backwards snowballing on the 13 publications. Together these 13 papers contain 509 references. We removed duplicates and assessed the relevance of each publication in the same way as described in our main analysis. This validation process yields an extra 30 papers which would have been good candidates for analysis. These 30 papers were not included in our study due to the strict requirements that our search criteria places on the title of accepted publications. Similarly, another weakness of this study is that we only used Google Scholar for the search. Querying multiple databases may have yielded more relevant results again.

Another threat of the literature review, and one discussed extensively throughout the study is one of subjectivity and bias in grouping the criteria, especially since there was only one rater. Three of the six groups have been identified by our previous work (relevance, evidence backed, message rigour) [63], then this literature review has added quality of writing, individual judgment and judgment of others to this list but it is possible that the criteria has been sorted into groups to fit into these existing categories (part of the rationale for conducting the survey was for one way to combat this threat). The lack of definitions reported in the analysed papers also contributed to this as assumptions needed to be made. Hence, the need to verify our refined criteria against a representative set of researchers.

There were also threats to validity within the survey. The survey participants were self selecting and few non–respondents gave feedback on the reasons why they didn't participate during our follow up. There are also questions as to the degree in which the respondents are representative of software researchers. The survey also asks respondents how they think they assess credibility rather then measuring the reality of what they actually do.

4.4 Future research

We plan to replicate the survey with software engineering practitioners so that we can compare the differences in credibility assessment between the two communities. The survey conducted also includes qualitative results. We intend to further analyse the qualitative data to better understand the reasons for the respondents quantitative scores, and to add further validation to our results.

With regards to the literature review, we can now address the threats by creating a set of less conservative search criteria and using them to add more studies to this review (such as the 30 papers identified during our snowballing validation). Threats can also be addressed by conducting a systematic literature review. As explained previously in this paper, we refrained from conducting a systematic literature review initially due to the maturity and objectives of the research. However, we have now been through a bootstrapping stage and have a clearer understanding of the problem domain.

4.5 Conclusions

We have investigated the criteria that are important to researchers in assessing the credibility of broadcast–oriented online media that has been written by practitioners. We have conducted a literature review, the outcome of which is a set of credibility criteria that have been used previously when researching different user groups, and that we believe are relevant to software engineering researchers to evaluate grey literature in GLRs and MLRs. We have also conducted a survey of software engineering researchers to validate our findings and create a more specific set of criteria for assessing blog articles in GLRs and MLRs. Both sets of criteria could help bring further organisation and rigour to the classification schemes proposed by previous research.

We have conducted a review of the literature to collate the criteria used by others for assessing the credibility of broadcast–oriented written online media. In doing so, we have found that the subjectivity of credibility assessment has led to researchers reporting criteria for individual user groups. However, none of the studies analysed looked at the criteria that aligns with the requirements that researchers hold over the evidence that they use. We have refined and grouped these criteria into areas that we believe are relevant to researchers, and then verified and refined our results further by conducting a survey of researchers. Our literature review reports that there are six distinct areas that exist for researchers assessing the credibility of their evidence; the relevance of the evidence to their research, the quality of the writing, whether the evidence is itself backed by evidence, the strength and rigour of what is being said and the arguments put forward, the judgment of the researcher, and the judgment of others that can influence the judgment of the researcher. The survey adds three further criteria; the reporting of the method in which the empirical data was collected, whether the article contains citations to other practitioner sources and whether the article contains citations to research (peer–reviewed) sources.

In the future, we plan to broaden our search criteria to allow for more studies to be added to our literature review. The next steps are also to look at how the criteria identified can be physically measured so that we can work towards a semi–automated method for filtering the high quality broadcast–oriented written online media (specifically blogs) from the vast quantity available.

Acknowledgement

We would like to thank the participants from the programme committees of EASE'18 and ESEM'18 who responded to this survey, as well as the respondents from SIN̂Z who participated in our pilot study. We would also like to thank those who reviewed the survey design prior to the pilot study, and Paul Ralph, Claes Wohlin and Sarah Beecham for reviewing an earlier version of this paper. The survey conducted in this paper has been approved by the University of Canterbury in ethics reference: HEC 2017/68/LR-PS

References

[1] Jean Adams, Frances C Hillier-Brown, Helen J Moore, Amelia A Lake, Vera Araujo-Soares, Martin White, and Carolyn Summerbell. "Searching and synthesising 'grey literature'and 'grey information'in public health: critical reflections on three case studies". In: *Systematic reviews* 5.1 (2016), p. 164.

[2] Richard J Adams, Palie Smart, and Anne Sigismund Huff. "Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies". In: *International Journal of Management Reviews* 19.4 (2017), pp. 432–454.

[3] Sharifah Aliman, Saadiah Yahya, and Syed Ahmad Aljunid. "Presage criteria for blog credibility assessment using Rasch analysis". In: *Journal of Media and Information Warfare* 4 (2011), pp. 59–77.

[4] Alyssa Appelman and S Shyam Sundar. "Measuring message credibility: Construction and validation of an exclusive scale". In: *Journalism & Mass Communication Quarterly* 93.1 (2016), pp. 59–79.

[5] John Bailey, David Budgen, Mark Turner, Barbara Kitchenham, Pearl Brereton, and Stephen Linkman. "Evidence relating to Object-Oriented software design: A survey". In: *Empirical Software Engineering and Measurement, 2007. ESEM 2007. First International Symposium on*. IEEE. 2007, pp. 482–484.

[6] Anton Barua, Stephen W Thomas, and Ahmed E Hassan. "What are developers talking about? an analysis of topics and trends in stack overflow". In: *Empirical Software Engineering* 19.3 (2014), pp. 619–654.

[7] David K Berlo. *The Process of Communication: An Introduction to Thery and Practice*. Rinehart Press, 1960.

[8] Nora J Bird, Claire R McInerney, and Stewart Mohr. "SOURCE EVALUATION AND INFORMATION LITERACY." In: *Communications in Information Literacy* 4.2 (2010).

[9] Lionel Briand. "Embracing the engineering side of software engineering". In: *IEEE software* 29.4 (2012), pp. 96–96.

[10] D Jasun Carr, Matthew Barnidge, Byung Gu Lee, and Stephanie Jean Tsang. "Cynics and skeptics: Evaluating the credibility of mainstream and citizen journalism". In: *Journalism & Mass Communication Quarterly* 91.3 (2014), pp. 452–470.

[11] Steven H Chaffee. "Mass media and interpersonal channels: Competitive, convergent, or complementary". In: *Inter/media: Interpersonal communication in a media world* 57 (1982), p. 77.

[12] Michael Chau and Jennifer Xu. "Business intelligence in blogs: Understanding consumer interactions and communities". In: *MIS quarterly* 36.4 (2012), pp. 1189–1216.

[13] Thomas Chesney and Daniel KS Su. "The impact of anonymity on weblog credibility". In: *International journal of human-computer studies* 68.10 (2010), pp. 710–718.

[14] Premkumar Devanbu, Thomas Zimmermann, and Christian Bird. "Belief & evidence in empirical software engineering". In: *Software Engineering (ICSE), 2016 IEEE/ACM 38th International Conference on*. IEEE. 2016, pp. 108–119.

[15] Sarjoun Doumit and Ali Minai. "Online news media bias analysis using an LDA-NLP approach". In: *International Conference on Complex Systems*. 2011.

[16] Gunther Eysenbach. "Credibility of health information and digital media: New perspectives and implications for youth". In: *Digital media, youth, and credibility* (2008), pp. 123–154.

[17] Norman Fenton, Shari Lawrence Pfleeger, and Robert L. Glass. "Science and substance: A challenge to software engineers". In: *IEEE software* 11.4 (1994), pp. 86–95.

[18] Andrew J Flanagin and Miriam J Metzger. "Digital media and youth: Unparalleled opportunity and unprecedented responsibility". In: *Digital media, youth, and credibility* (2008), pp. 5–27.

[19] BJ Fogg, Cathy Soohoo, David R Danielson, Leslie Marable, Julianne Stanford, and Ellen R Tauber. "How do users evaluate the credibility of Web sites?: a study with over 2,500 participants". In: *Proceedings of the 2003 conference on Designing for user experiences*. ACM. 2003, pp. 1–15.

[20] BJ Fogg and Hsiang Tseng. "The elements of computer credibility". In: *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM. 1999, pp. 80–87.

[21] BJ Fogg, Jonathan Marshall, Othman Laraki, Alex Osipovich, Chris Varma, Nicholas Fang, Jyoti Paul, Akshay Rangnekar, John Shon, Preeti Swani, et al. "What makes Web sites credible?: a report on a large quantitative study". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM. 2001, pp. 61–68.

[22] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. "Guidelines for including the grey literature and conducting multivocal literature reviews in software engineering". In: *arXiv preprint arXiv:1707.02553* (2017).

[23] Vahid Garousi, Michael Felderer, and Mika V Mäntylä. "The need for multivocal literature reviews in software engineering: complementing systematic literature reviews with grey literature". In: *Proceedings of the 20th International Conference on Evaluation and Assessment in Software Engineering*. ACM. 2016, p. 26.

[24] Vahid Garousi and Mika V Mäntylä. "When and what to automate in software testing? A multi-vocal literature review". In: *Information and Software Technology* 76 (2016), pp. 92–117.

[25] Robert H Gass and John S Seiter. "Credibility and public diplomacy". In: *Routledge handbook of public diplomacy* (2009), pp. 154–165.

[26] Robert H Gass and John S Seiter. *Persuasion: Social influence and compliance gaining*. Routledge, 2015.

[27] Albert C Gunther. "Biased press or biased public? Attitudes toward media coverage of social groups". In: *Public Opinion Quarterly* 56.2 (1992), pp. 147–167.

[28] Emitza Guzman, David Azócar, and Yang Li. "Sentiment analysis of commit comments in GitHub: an empirical study". In: *Proceedings of the 11th Working Conference on Mining Software Repositories*. ACM. 2014, pp. 352–355.

[29] Carl I Hovland, Irving L Janis, and Harold H Kelley. "Communication and persuasion; psychological studies of opinion change." In: (1953).

[30] Staffs Keele et al. "Guidelines for performing systematic literature reviews in software engineering". In: *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. sn, 2007.

[31] Barbara A Kitchenham, Tore Dyba, and Magne Jorgensen. "Evidence-based software engineering". In: *Proceedings of the 26th international conference on software engineering*. IEEE Computer Society. 2004, pp. 273–281.

[32] Rikard Larsson. "Case survey methodology: Quantitative analysis of patterns across case studies". In: *Academy of management Journal* 36.6 (1993), pp. 1515–1546.

[33] Qingzi Vera Liao. "Effects of cognitive aging on credibility assessment of online health information". In: *CHI'10 Extended Abstracts on Human Factors in Computing Systems*. ACM. 2010, pp. 4321–4326.

[34] Johan Linâker, Sardar Muhammad Sulaman, Rafael Maiani de Mello, and Martin Hˆst. "Guidelines for conducting surveys in software engineering". In: (2015).

[35] Rhonda W Mack, Julia E Blose, and Bing Pan. "Believe it or not: Credibility of blogs in tourism". In: *Journal of Vacation marketing* 14.2 (2008), pp. 133–144.

[36] A Manolică, O Ciobanu, C Bobâlcă, and C Sasu. "A Method to Asses Credibility of Commercial Web SItes. One level to Change Consumers' Attitude and Behaviour"". In: *Proceedings of the International Conference on Management of Technological Changes*. 2011.

[37] Mika V Mäntylä, Bram Adams, Foutse Khomh, Emelie Engström, and Kai Petersen. "On rapid releases and software testing: a case study and a semi-systematic literature review". In: *Empirical Software Engineering* 20.5 (2015), pp. 1384–1425.

[38] James C McCroskey and Thomas J Young. "Ethos and credibility: The construct and its measurement after three decades". In: *Communication Studies* 32.1 (1981), pp. 24–34.

[39] Ericka Menchen-Trevino and Eszter Hargittai. "Young adults' credibility assessment of wikipedia". In: *Information, Communication & Society* 14.1 (2011), pp. 24–51.

[40] Miriam J Metzger. "Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research". In: *Journal of the Association for Information Science and Technology* 58.13 (2007), pp. 2078–2091.

[41] Philip Meyer. "Defining and measuring credibility of newspapers: Developing an index". In: *Journalism quarterly* 65.3 (1988), pp. 567–574.

[42] Ekaterina Ognianova. "Effects of the content provider's perceived credibility and identity on ad processing in computer-mediated environments". In: *PROCEEDINGS OF THE CONFERENCE-AMERICAN ACADEMY OF ADVERTISING*. AMERICAN ACADEMY OF ADVERTISING. 1998, pp. 155–156.

[43] Daniel J O'keefe. *Persuasion: Theory and research*. Vol. 2. Sage, 2002.

[44] Katrina L Pariera. "INFORMATION LITERACY ON THE WEB How College Students Use Visual and Textual Clues to Assess Credibility on Health Websites". In: (2012).

[45] Michael J Paul and Mark Dredze. "You are what you Tweet: Analyzing Twitter for public health." In: *Icwsm* 20 (2011), pp. 265–272.

[46] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. "Systematic Mapping Studies in Software Engineering." In: *EASE*. Vol. 8. 2008, pp. 68–77.

[47] Kai Petersen, Sairam Vakkalanka, and Ludwik Kuzniarz. "Guidelines for conducting systematic mapping studies in software engineering: An update". In: *Information and Software Technology* 64 (2015), pp. 1–18.

[48] Richard E Petty and John T Cacioppo. "The elaboration likelihood model of persuasion". In: *Advances in experimental social psychology* 19 (1986), pp. 123–205.

[49] Austen Rainer. "Using argumentation theory to analyse software practitioners' defeasible evidence, inference and belief". In: *Information and Software Technology* (2017).

[50] Austen Rainer, Tracy Hall, and Nathan Baddoo. "Persuading developers to" buy into" software process improvement: a local opinion and empirical evidence". In: *Empirical Software Engineering, 2003. ISESE 2003. Proceedings. 2003 International Symposium on*. IEEE. 2003, pp. 326–335.

[51] Soo Young Rieh. "Credibility and cognitive authority of information". In: (2010).

[52] Soo Young Rieh, Yong-Mi Kim, Ji Yeon Yang, and Beth St Jean. "A diary study of credibility assessment in everyday life information activities on the Web: Preliminary findings". In: *Proceedings of the Association for Information Science and Technology* 47.1 (2010), pp. 1–10.

[53] Charles C Self. "Credibility". In: *An integrated approach to communication theory and research* 1 (1996), pp. 421–441.

[54] Pankajeshwara Sharma, Bastin Tony Roy Savarimuthu, Nigel Stanger, Sherlock A Licorish, and Austen Rainer. "Investigating developers' email discussions during decision-making in Python language evolution". In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. ACM. 2017, pp. 286–291.

[55] Beth St Jean, Soo Young Rieh, Ji Yeon Yang, and Yong-Mi Kim. "How content contributors assess and establish credibility on the web". In: *Proceedings of the Association for Information Science and Technology* 48.1 (2011), pp. 1–11.

[56] Margaret-Anne Storey, Leif Singer, Brendan Cleary, Fernando Figueira Filho, and Alexey Zagalsky. "The (r) evolution of social media in software engineering". In: *Proceedings of the on Future of Software Engineering*. ACM. 2014, pp. 100–116.

[57] Wee-Kheng Tan and Yu-Chung Chang. "Credibility assessment model of travel information sources: An exploratory study on travel blogs". In: *Information and Communication Technologies in Tourism 2011*. Springer, 2011, pp. 457–469.

[58] Wee-Kheng Tan and Yun-Ghang Chang. "Place Familiarity and Attachment: Moderators of The Relationship Between Readers' Credibility Assessment of A Travel Blog and Review Acceptance". In: *Journal of Travel & Tourism Marketing* 33.4 (2016), pp. 453–470.

[59] Oana Tugulea et al. "Does a different year of study means different important credibility dimensions? A study on the dimensions of credibility of online sales websites". In: *Review of Economic And Business Studies* 7.2 (2014), pp. 31–49.

[60] Oana Ţugulea. "Different Web Credibility Assessment as a Result of One Year Difference in Education". In: *Review of Economic and Business Studies* 8.2 (2015), pp. 117–133.

[61] Rita Zaharah Wan-Chik. "Information credibility assessment of Islamic and Quranic information on the web". In: *Proceedings of the 9th International Conference on Ubiquitous Information Management and Communication*. ACM. 2015, p. 25.

[62] Ashley Williams. "Do software engineering practitioners cite research on software testing in their online articles?: A preliminary survey." In: *Proceedings of the 22nd International Conference on Evaluation and Assessment in Software Engineering 2018*. ACM. 2018, pp. 151–156.

[63] Ashley Williams and Austen Rainer. "Toward the use of blog articles as a source of evidence for software engineering research". In: *Proceedings of the 21st International Conference on Evaluation and Assessment in Software Engineering*. ACM. 2017, pp. 280–285.

[64] Claes Wohlin. "An evidence profile for software engineering research and practice". In: *Perspectives on the Future of Software Engineering*. Springer, 2013, pp. 145–157.

[65] Quan Yuan and Qin Gao. "The Analysis of Online News Information Credibility Assessment on Weibo Based on Analyzing Content". In: *International Conference on Engineering Psychology and Cognitive Ergonomics*. Springer. 2016, pp. 125–135.

[66] Dan Zhao, Chunhui Tan, and Yutao Zhang. "Evaluating the Enterprise Website Credibility from the Aspect of Online Consumers". In: *Management of e-Commerce and e-Government, 2009. ICMECG'09. International Conference on*. IEEE. 2009, pp. 14–17.

Appendix

Accepted/ Rejected	Data source	Description/justification
Accepted	Blogs	Includes studies that use any type of blog(s) as the sole source of their analysis. This could be personal blogs or blogs published by a company.
Accepted	Online news	Includes studies that emphasise an analysis of online news. This could be traditional online news outlets (such as the websites of CNN, BBC etc) or any other type of news outlet (such as blog articles or video's that report on news topics). One of the studies in this category [10] looks at the difference of perceived credibility between mainstream and citizen journalism. Their study included users watching videos reporting a news story from either; a suit-wearing reporter at a news-desk, or a more casually clothed vlogger in a home setting. The reason that this study was placed in the 'Online news' category and not the rejected 'YouTube/Video' category was because of its explicit focus on online news.
Accepted	Wikipedia	Includes studies that use Wikipedia as the sole source of analysis.
Accepted	Websites that have explicit focus	Includes studies that label their source of analysis as 'the web' but give an explicit focus to the type of information that they require (e.g. online health information). In this segment there are 6 studies; 2 assess online health information, 2 assess enterprise or commercial websites, 1 looks at user-generated content (UGC) on the web and 1 looks at how participants assess Islamic and Quaranic content on the web.
Rejected	Product reviews/ online recommendations	Includes studies that look at the credibility of product reviews and online recommendations (on sites such as Yelp!). This research is interested in the credibility of practitioners writing about their own opinions and experiences. Therefore, these studies are rejected from our analysis.
Rejected	YouTube/video	Includes studies that look at assessing the credibility of any online video or motion media. Again, these are rejected from our analysis due to their incompatibility with the overall aims of the research.
Rejected	Social media profiles	Includes studies that assess the credibility of individual social media profiles. Again, rejected from our analysis due to their incompatibility with the overall aims of the research.
Rejected	Multiple media sources	Includes studies that look at the credibility assessment of multiple types of media, or that use the term 'media' as a blanket statement for assessing all types of media. These studies are rejected as they are too vague to determine whether they apply to this research.
Rejected	Forums/ conversational	Includes studies that assess the credibility of online forums or websites where the content is conversations (e.g. a comments thread). It can be argued that microblogs such as Twitter also fall under this category as their character limit and interaction makes their content distribution more conversational rather than the article type publications typical of traditional blogs. For this reason, they too are rejected from our analysis.
Rejected	Generic websites or 'the web'	Unlike the studies that analyse the web with a specific focus on a particular topic. These studies simply label their unit of analysis as 'the web.' We reject these from our analysis as their vagueness makes it difficult to determine their relevance to the research.
Rejected	Web searching	Includes studies that assess the credibility of search results on the web, such as results from search engines. These studies are rejected due to their incompatibility with the overall aims of this research.

Table 20: The different data sources and a justification for why they were accepted/rejected

Criteria	[58]	[3]	[57]	[65]	[4]	[10]	[39]	[55]	[66]	[61]	[60]	[33]	[44]	Total
identity		x								x				2
credentials/qualifications		x								x				2
reputation		x			x					x				3
expertise			x		x						x			3
authorship												x		1
engagement		x									x			2
Totals	0	4	1	0	2	0	0	0	0	3	2	1	0	

Table 21: Criteria classified under the source category

Criteria	[58]	[3]	[57]	[65]	[4]	[10]	[39]	[55]	[66]	[61]	[60]	[33]	[44]	Total
past experience with site		x						x						2
general suspicions		x												1
general dislike		x												1
aligns with own knowledge										x				1
recommended										x				1
endorsed					x							x		2
location of user										x				1
source		x												1
usefulness		x												1
name recognition		x												1
relevance										x				1
cue in the content										x				1
trust			x											1
Totals	0	6	1	0	1	0	0	1	0	5	0	1	0	

Table 22: Criteria classified under the receiver (the reader) category

Criteria	[58]	[3]	[57]	[65]	[4]	[10]	[39]	[55]	[66]	[61]	[60]	[33]	[44]	Total
believable	x				x								x	3
motivation		x												1
focus		x			x									2
clarity		x			x									2
trustworthiness			x		x	x		x			x		x	6
currency				x										1
transparent					x									1
will have impact					x									1
professional					x	x							x	3
representative					x					x				2
spin-free					x									1
partisan nature						x								1
intrinsic plausibility								x						1
honest								x						1
popularity										x				1
sincere											x			1
etiquette											x			1
truthful					x			x						2
authentic					x					x				2
experience		x						x			x			3
cite external source								x		x				2
trusted sources								x						1
multiple sources								x						1
verified										x				1
cited										x				1
accurate	x	x	x	x	x	x		x		x				8
writing tone		x						x						2
well written		x			x			x		x				4
update		x												1
corrections		x												1
authority				x	x					x				3
error-free					x									1
argument strength/ content			x									x		2
balanced					x	x				x				3
equal					x									1
neutral					x									1
objective					x			x						2
not opinionated					x	x								2
reliable					x			x		x				3
comprehensive					x									1
consistent					x									1
detailed					x									1
unbiased	x	x	x	x	x	x		x						7
complete	x		x	x	x					x				5
factual	x													1
fairness			x		x	x		x						4
truth-seeking intentions						x								1
credibility						x			x	x				3
Totals	5	10	6	5	27	10	0	15	1	13	4	1	3	

Table 23: Criteria classified under the message category

Criteria	[58]	[3]	[57]	[65]	[4]	[10]	[39]	[55]	[66]	[61]	[60]	[33]	[44]	Total
sponsorship		x						x						2
affiliations		x								x				2
privacy policy		x										x		2
site functionality		x												1
customer service		x												1
advertising		x									x			2
image credibility									x					1
business function credibility									x					1
efficient admin										x				1
information support											x			1
navigation tools												x		1
site length											x			1
ease of use											x			1
real world feel											x			1
accessibility										x				1
URL								x						1
social relationship		x						x						2
performance		x												1
design look		x						x				x	x	4
design/ structure		x	x											2
attractiveness			x		x									2
Totals	0	10	2	0	1	0	0	4	2	3	5	3	1	

Table 24: Criteria classified under the channel (the medium) category

Table 25: Summary statistics for each of the three subsamples (subsample = S; clarity of writing = CoW; reporting of empirical data = ED; reporting of the method for data collection = M; professional experience = E; web links to other practitioner sources = PC; web links to peer-reviewed research = RC; reasoning = R; prior beliefs of the reader = PB; influence of others on the reader's beliefs = I)

S	Statistic	CoW	ED	M	E	PC	RC	R	PB	I
1	**COUNT**	7	7	7	7	7	7	7	7	7
2	**COUNT**	29	29	29	29	29	29	29	29	29
3	**COUNT**	7	7	7	7	7	7	7	7	7
1	**MIN**	3	5	2	2	2	4	4	2	2
2	**MIN**	2	3	2	1	0	0	2	0	0
3	**MIN**	3	2	1	3	2	1	4	2	2
1	**MAX**	6	6	6	6	6	6	6	5	5
2	**MAX**	6	6	6	6	6	6	6	6	6
3	**MAX**	6	6	5	6	5	5	6	6	5
1	**MEAN**	4.6	5.4	5	4.7	4.3	4.9	5.1	3.2	3.4
2	**MEAN**	4.6	4.9	4.7	4.3	3.9	4.3	4.9	2.8	2.7
3	**MEAN**	4.7	4.1	3.7	4.9	3.9	3.4	5.6	3.9	3.6
1	**MODE**	6	5	6	5	5	5	6	3	2
2	**MODE**	5	6	4	5	5	5	5	0	1
3	**MODE**	5	5	4	5	4	4	6	4	4
1	**MEDIAN**	5	5	5	5	5	5	5	3	3
2	**MEDIAN**	5	5	5	5	4	5	5	3	2
3	**MEDIAN**	5	4	4	5	4	4	6	4	4
1	**STDEV**	1.5	0.5	1.4	1.4	1.5	0.7	0.9	1	1.3
2	**STDEV**	1.2	1	1.1	1.3	1.6	1.5	1.1	2.1	2
3	**STDEV**	1.3	1.3	1.4	1.1	0.9	1.5	0.8	1.3	1.1
1	**Rank by median**	2	1	1	1	1	1	1	3	3
2	**Rank by median**	1	1	1	1	2	1	1	3	4
3	**Rank by median**	2	3	3	2	3	3	1	3	3
1	**Rank by median, then mean**	5	1	2	4	6	3	2	8	7
2	**Rank by median, then mean**	4	2	3	5	6	5	1	7	8
3	**Rank by median, then mean**	3	4	6	2	5	8	1	5	7
1	**% of extremely importants**	42.9	42.9	42.9	28.6	14.3	14.3	42.9	0	0
2	**% of extremely importants**	27.6	34.5	31	17.2	10.3	17.2	31	6.9	10.3
3	**% of extremely importants**	28.6	14.3	0	28.6	0	0	71.4	14.3	0
1	**Rank by % of Extremely importants**	1	1	1	2	3	3	1	4	4
2	**Rank by % of Extremely importants**	3	1	2	4	5	4	2	6	5
3	**Rank by % of Extremely importants**	2	3	4	2	4	4	1	3	4