Tools to Support Systematic Reviews in Software Engineering: A Cross-Domain Survey using Semi-structured Interviews

Christopher Marshall
School of Computing and Mathematics
Keele University
Staffordshire UK
c.marshall@keele.ac.uk

Pearl Brereton
School of Computing and Mathematics
Keele University
Staffordshire UK
o.p.brereton@keele.ac.uk

Barbara Kitchenham
School of Computing and Mathematics
Keele University
Staffordshire UK
b.a.kitchenham@keele.ac.uk

ABSTRACT

Background: A number of software tools are being developed to support systematic reviewers within the software engineering domain. However, at present, we are not sure which aspects of the review process can most usefully be supported by such tools or what characteristics of the tools are most important to reviewers. **Aim:** The aim of the study is to explore the scope and practice of tool support for systematic reviewers in other disciplines. **Method:** Researchers with experience of performing systematic reviews in Healthcare and the Social Sciences were surveyed. Qualitative data was collected through semi-structured interviews and data analysis followed an inductive approach. **Results:** 13 interviews were carried out. 21 software tools categorised into one of seven types were identified. Reference managers were the most commonly mentioned tools. Features considered particularly important by participants were support for multiple users, support for data extraction and support for tool maintenance. The features and importance levels identified by participants were compared with those proposed for tools to support systematic reviews in software engineering. **Conclusions:** Many problems faced by systematic reviewers in other disciplines are similar to those faced in software engineering. There is general consensus across domains that improved tools are needed.

Categories and Subject Descriptors

D.2.m [**Software Engineering**]: Miscellaneous

Keywords

Systematic review, automated tools, survey

1. INTRODUCTION

Systematic Reviews (SRs) involve the systematic storage, management, validation and analysis of large quantities of data, activities which can be error prone and time consuming [1, 2, 3, 4]. A range of software tools have been used to assist systematic reviewers in software engineering (SE) and in other disciplines. These include basic productivity tools, such as word processors and spreadsheets, reference managers, statistics packages and purpose-built tools which target all (or most) of the stages of the review process.

EASE '15, April 27 - 29, 2015, Nanjing, China Copyright 2015 ACM 978-1-4503-3350-4/15/04…$15.00
http://dx.doi.org/10.1145/2745802.2745827

Research has investigated the use of tools to support systematic reviewers. For example, within the Healthcare domain, a survey of information systems to support or automate SR tasks found a wide range of tools [5]. Tools discussed by Tsafnat et al. include the Cochrane Commission's Review manager (RevMan)[1], federated search engines such as Quick Clinical, citation managers (such as Endnote and ProCite), the Abstrackr system to support screening of abstracts, and meta-analysis tools (which are "already in wide use"). A more focused cross-domain mapping study of visual data mining support for SRs found that "most of the studies (16 out of 20 studies) have been conducted in the field of medicine" [6]. The authors of the study reported that data extraction and data synthesis were the most likely stages of the SR process to be supported by visual data mining tools.

Within the SE domain, a mapping study of tools for SRs (other than basic productivity tools, spreadsheets and reference managers) also found that a range of visualisation and text mining tools had been used to support study selection, data extraction and data synthesis [7].

The study reported in this paper is part of a research programme to develop and validate an evaluation framework for tools to support SRs in SE. The framework is composed of a set of features (see Table 1), associated importance weightings and scoring instruments, and has been used as part of a feature analysis that compared four tools designed to support most of the stages of the SR process in SE [8]. The features (as summarised in Table 1) were based on the experiences of performing SRs in SE reported in the literature [1, 2, 3, 9], a preliminary screening of candidate tools and discussion amongst the researchers who performed the feature analysis. This study aims to explore the experiences and opinions of systematic reviewers in domains other than SE with a particular focus on their use of and views about support tools. The goals of the study are to:

1) explore what tools are currently available and used to support SRs in other domains.
2) identify what participants consider to be the most important characteristics (or features) of tools to support SRs.
3) compare the features and importance levels identified in the survey with those forming part of our proposed evaluation framework (see Table 1).

[1] http://tech.cochrane.org/Revman

Table 1. Set of Features

id	*Feature Set*	*id*	*Feature*
F1	Economic	F1-F01	The tool does not require financial payment to use
		F1-F02	Maintenance
F2	Ease of introduction and setup	F2-F01	Simple installation and setup
		F2-F02	The tool is self-contained
F3	SR activity support	F3-F01	Protocol development
		F3-F02	Protocol validation
		F3-F03	Supports automated searches
		F3-F04	Study selection and validation
		F3-F05	Quality assessment and validation
		F3-F06	Data extraction and validation
		F3-F07	Automated analysis
		F3-F08	Text analysis
		F3-F09	Meta-analysis
		F3-F10	Report write up
		F3-F11	Report validation
F4	Process Management	F4-F01	Support for multiple users
		F4-F02	Document management
		F4-F03	Security
		F4-F04	Management of roles
		F4-F05	Re-use of data from past projects

The paper is organised as follows. Section two describes the methodology used. Section three presents the results. This is followed by a discussion of the results in section four along with some of the study's limitations. Finally, conclusions are drawn in section five with details of on-going and future work.

2. METHODOLOGY

The study takes the form of a survey and uses semi-structured interviews for data collection. Survey research is a particularly suitable method of gathering self-reported quantitative and qualitative data [10]. In this section, we describe the data collection, the approach taken for the selection of participants, the interview procedures and the data analysis strategy.

2.1 Data Collection

Since the goal of the study is to explore the experiences and opinions of systematic reviewers, it can be considered primarily as being qualitative in nature. Qualitative research focuses on investigating and understanding social and cultural phenomena in context [11] and is appropriate where the purpose is to explore a topic and obtain an overview of a complex area [12]. Semi-structured interviews are particularly suitable for collecting qualitative data because, unlike self-administered questionnaires, they provide the opportunity for discussion or exploration of new topics that arise during data collection.

Semi-structured interviews allow for considerable freedom in the sequencing of questions and in the amount of time and attention given to each topic. Questions can be open-ended, allowing for a variety of responses. This approach to data collection helps to reduce the risk of bias relating to the researcher's preconceptions and it allows for the use of elaboration probes to encourage the participant to keep talking about a particular subject [13].

2.1.1 Questions

Questions driving the interviews were grouped into four categories as shown in Table 2, which includes some examples of questions for each group. For Group 4 questions, participants were asked to rate each feature (see Table 1) as either *mandatory*, *highly desirable*, *desirable*, *nice to have* or *not necessary*.

2.1.2 Pilot Interview

The interview instruments and procedures were piloted with a PhD student who had undertaken two SRs. This experience confirmed our expectation that interviews would take approximately 45 minutes and also led to some changes in the delivery and sequencing of questions.

2.2 Selection of Participants

Participants were researchers in Healthcare and Social Sciences with knowledge and experience of the SR methodology. A combination of convenience and snowballing sampling techniques was used to recruit participants. An email invitation, which described the research project, the aim of the study and the commitment required was sent to 49 potential participants. 13 researchers from six institutions across the UK agreed to be interviewed. Table 3 summarises the role, field of interest and SR experience for each participant.

2.3 Interview Procedures

Interviews were carried out between June 2014 and September 2014. Prior to interview, each participant was sent an *Interview Preparation Sheet*. This document outlined the main themes to be covered during the interview, the expected duration, and measures which would be taken to ensure privacy and confidentiality. All interviews were carried out face-to-face by a single interviewer and recorded using a digital audio recorder. The researcher took notes throughout each interview. On average, each interview lasted 45 minutes. The shortest interview took 32 minutes and the longest interview lasted for 68 minutes.

2.4 Data Analysis Strategy

The raw data (i.e. recordings, field notes) was processed prior to analysis. Analysis took place concurrently with data collection, as recommended by Miles, Huberman & Saldana [14]. Analysis was an inductive process, which allowed for categories and codes to emerge progressively during the data collection [14].

3. RESULTS

3.1 Automated Tools to Support SRs

In this section, the tools referenced by participants are presented and have been classified by type. A summary of these results is presented in Table 4.

There were 21 tools identified by participants, which have been classified into seven categories as shown in Table 4.

Table 2. Example Questions Grouped by Topic

Question Group	*Example Questions*
[Group 1] Domain Context	*Could you tell me about the domain you are currently situated in and some of the work that you do?*
	How do systematic reviews play a role within your discipline?
[Group 2] Personal Experiences with SRs	*What types of SRs have you had experience with (e.g. were they primarily qualitative or quantitative)?*
	What, in your opinion, are the main challenges when undertaking a SR?
[Group 3] Experiences with Tools	*What tools have you used to support yourself whilst undertaking a SR?*
	What were some of the main strengths and weaknesses of the tool(s)?
[Group 4] Features of an SR Tool	*How important is support for the development of a review protocol?*
	How important is support for multiple users to work on a single review (i.e. collaboration)?

Table 3. Participant Information

id	*Role*	*Domain*	*No. of SRs*	*Type of SR (Qualitative or Quantitative)*
P-01	Research Associate	Healthcare	6 – 10	Both
P-02	Research Associate	Healthcare	1 – 5	Quantitative
P-03	PhD Student	Healthcare	1 – 5	Qualitative
P-04	Senior Lecturer	Healthcare	1 – 5	Qualitative
P-05	Information Officer	Healthcare	11 – 15	Quantitative
P-06	Lecturer	Healthcare	1 – 5	Quantitative
P-07	Lecturer	Social Science	1 – 5	Quantitative
P-08	Information Officer	Social Science	15+	Both
P-09	Professor	Social Science	15+	Both
P-10	Systematic Reviewer	Social Science	6 – 10	Both
P-11	Research Associate	Social Science	1 – 5	Both
P-12	Professor	Social Science	15+	Qualitative
P-13	Information Specialist	Healthcare	15+	Both

The majority of tools identified by participants were reference managers. In particular, *RefWorks* and *EndNote* were mentioned most often. *RefWorks* was praised by participants for its ability to ***"aid your systematic search process"*** and being able to ***"check for duplication"*** of papers. To some extent, *RefWorks* could also support study selection, with one participant explaining how they ***"classified studies using folders"*** to manage included and excluded papers. *RefWorks*, however, was criticised for the lack of a bulk export feature (***"you cannot export all your searches in one go.")*** and poor usability ***("I don't think it's easy to use at all. There is a lot compacted onto one screen")***.

EndNote was praised for having a web-based interface for remote access ***("I can access it anywhere, which is good.")***. Similar to *RefWorks*, some participants used *EndNote* to support study selection even though a feature to support this stage is not explicitly supported ***("I don't think it's built to do that, it's just the way I use it.")***. Participants also liked having ***"discrete databases for each review."*** This is not the case in RefWorks, which uses a ***"folder driven system."*** Participants at times, however, felt restricted by the tool, with some feeling they were unable to take their data to the ***"next stage of the review"*** due to weak export capabilities. Some raised concerns about poor support for team-based SRs ***("It's not ideal when you've got a big team.")*** and whether the system could effectively handle large numbers of papers/studies ***("people are concerned that it doesn't have the capacity to deal with huge numbers of references.").***

Two special-purpose tools designed to support particular stages of an SR (or the whole process) were identified; namely, *EPPI-Reviewer* and *RevMan*. The current version of *EPPI-Reviewer*, *EPPI-Reviewer 4*, is a comprehensive single or multi-user web-based system for managing SRs across Healthcare and Social Science domains. During the interviews, participants were very positive about the variety of ways in which the tool can support the SR process. For example, *EPPI-Reviewer* includes a feature aiming to improve the efficiency of a SR, which uses text mining ***"to prioritise the most relevant studies."*** This feature ***"pulls the most relevant ones*** [studies] ***to the beginning"*** and allows the review team ***"to start the full data extraction of the studies before finishing the screening."*** *EPPI-Reviewer* also uses visualisation techniques to support thematic analysis. This feature, which allows users to ***"depict the relationships between concepts,"*** was also considered useful. Participants, however, felt *EPPI-Reviewer* had a steep learning curve and that it ***"takes a while to learn all of the different things."*** In addition, some participants felt the ***"training could be improved."***

RevMan primarily supports the preparation and maintenance of Cochrane Reviews; although, it can be used to support other reviews. *RevMan* was praised by participants for its good support for statistical analysis techniques; in particular, meta-analysis ***("meta-analysis is quite easy")***. Support for protocol development was also considered useful ***("It helps with the protocol stage as well. It helps guide you.")***. Some users, however, felt, at times, restricted by the tool since some of its features were not accessible unless it was a Cochrane Review ***("if your review is not Cochrane commissioned then you can't use that feature of RevMan.").*** Other users also felt ***"confused"*** by the tool.

3.2 Rating the Features

In this section, the results of the feature rating exercise is presented. A summary of the key points raised by participants, for each feature, is given. The feature ratings are presented in Table 5 where the bold, underlined number is the modal response rating for the feature.

3.2.1 Feature Set 1 (F1): Economic

Concerning financial payment of a tool (F1-F01), some participants thought having the tool ***"free for personal use"*** with ***"different licenses for different*** [types] ***of user"*** would be a good idea. The majority of participants, however, felt having to pay for a tool was not an issue. One participant stated they would be ***"less inclined to use something if it was completely free"*** as they are placing trust in the tool to hold their valuable data. One participant commented about a lack of confidence in free; specifically, web-based tools, noting that they could ***"disappear tomorrow."***

Many participants felt maintenance of a tool (F1-F02), post development, was very important as there are ***"bound to be teething problems with something this massive."*** Also, as the ***"SR method changes"*** over time, the tool needs to ***"evolve"*** with those changes and bring new features and updates. Ratings for this feature are shown in rows 17 and 3 of Table 5 respectively.

Table 4. Tools Identified by Participants

Tool Type	*Tools*	*Participants (P)*	*Total*
Reference Management Tools	RefWorks	P-01; P-03; P-04; P-05; P-06	5
	EndNote / EndNote Web	P-04; P-05; P-08; P-09; P-13	5
	Mendeley	P-03; P-07; P-12; P-13	4
	Reference Manager	P-02; P-08; P-13	3
	ProCite	P-09	1
Special Purpose Tools	Review Manager (RevMan)	P-01; P-02; P-03; P-05; P-07; P-09; P-13	7
	EPPI-Reviewer	P-08; P-09; P-10; P-11	4
Basic Productivity Tools	Microsoft Word	P-02; P-04; P-09; P-13	4
	Microsoft Excel	P-02; P-07; P-12	3
Advanced Analysis Software	STATA	P-01; P-02; P-09	3
	NVivo	P-07; P-12	2
	SPSS	P-06; P-09	2
	Mplus	P-07	1
	ATLAS.ti	P-12	1
Other	FreeMind	P-04; P-13	2
	RIS conversion tool	P-08	1
	PubReMiner	P-13	1
Custom-built tool	Web-based coding tool	P-07	1
	Excel add-in	P-02	1
Meta-analysis tools	MetaEasy	P-07	1
	MetaLight	P-07	1

Table 5. Summary of Participant Ratings for each Feature

Row No.	*id*	*Feature*	*Mandatory*	*Highly Desirable*	*Desirable*	*Nice-to-have*	*Not Necessary*	*SE Feature Ratings* [8]
1	F4-F01	Multiple users	**9**	2	2	0	0	Mandatory
2	F3-F06	Data extraction	**7**	5	1	0	0	Highly Desirable
3	F1-F02	Maintenance	6	**7**	0	0	0	Highly Desirable
4	F2-F02	Simple installation and setup procedure	**6**	5	1	1	0	Highly Desirable
5	F4-F02	Document management	**6**	4	2	1	0	Mandatory
6	F4-F03	Security	**6**	2	1	3	1	Desirable
7	F3-F05	Quality assessment and validation	5	**7**	1	0	0	Highly Desirable
8	F3-F07	Automated analysis	5	**7**	1	0	0	Highly Desirable
9	F3-F04	Study selection and validation	5	**6**	2	0	0	Highly Desirable
10	F3-F09	Meta-analysis	4	**5**	2	2	0	Nice-to-have
11	F2-F05	Re-use of data from past projects	3	**7**	3	0	0	N/A
12	F3-F03	Search process	3	**4**	3	3	0	Highly Desirable
13	F4-F04	Role management	3	3	2	**4**	1	Highly Desirable
14	F3-F01	Development of review protocol	2	**4**	2	3	2	Desirable
15	F3-F02	Protocol validation	1	1	**5**	1	**5**	Desirable
16	F2-F05	Self-contained	0	**6**	**6**	0	1	Highly Desirable
17	F1-F01	No financial payment	0	**5**	3	1	4	Highly Desirable
18	F3-F11	Report validation	0	3	3	3	**4**	Nice-to-have
19	F3-F08	Text analysis	0	3	2	**5**	3	Nice-to-have
20	F3-F10	Report write-up	0	2	**6**	4	0	Nice-to-have

3.2.2 Feature Set 2 (F2): Ease of Introduction

Some participants felt that without a simple installation process (F2-F01), users would become ***"frustrated with it"*** One participant pointed out that you ***"you don't pick your collaborators based on their IT skills"*** and, therefore, a simple installation is important. Other participants, however, felt that ***"if the tool is good enough,"*** then, ***"some people are prepared to give*** [the difficult setup] ***a go"***. Many participants felt having a self-contained (F2-F02) tool (i.e. able to function, primarily, as a stand-alone application) was preferable and that, if this was the case, then as a user ***"you are more likely engage with the tool."*** Other participants, however, felt it wasn't an issue and that they'd ***"probably be quite happy installing other packages."*** if the tool ***"does stuff that nothing else can do."*** Ratings for these features are shown in rows 4 and 16 of Table 5 respectively

3.2.3 Feature Set 3 (F3): SR Activity Support

Participants stated that support for developing the review protocol (F3-F01) would be ***"highly useful"***; particularly, within a ***"large-scale review team"***. Some participants, however, were unsure of its usefulness, stating that there were ***"already resources*** (e.g. Cochrane Handbook) ***which support this"*** and that using ***"Word and track changes"*** is sufficient. Participants felt that tool support for protocol validation (F3-F02) would be useful for ***"making sure you don't miss anything"*** and that by having a ***"workable check-list,"*** it makes things easier. Some participants, however, felt that introducing automation might be ***"over-complicating the process."*** Ratings for these features are shown in rows 14 and 15 of Table 5 respectively.

Many participants felt that automated support for the search process (F3-F03) would be ***"very useful"*** and ***"save a lot of time"***. In particular, participants felt that automated support could be helpful for ***"developing the search strategy"*** particularly when ***"piloting your search terms."*** A number of participants, however, questioned the ***"reliability"*** of such a feature. Ratings for this feature are shown in row 12 of Table 5.

Participants felt that tool support for study selection and validation (F3-F04) has the potential to ***"reduce a lot of workload"*** and could ***"speed up the overall process."*** A facility for resolving disagreements was also praised. Some participants, however, felt that a lot of what the feature was targeting support for could be solved with a ***"quick conversation"*** between members of the review team. Concerning tool support for quality assessment (F3-F05), the majority of participants felt this would be another useful feature since ***"all these things otherwise require meetings and organisation."*** In particular, a facility to compare user assessments and ***"identify where your disagreements are, would be really good."*** Some participants raised concerns about the feature's ***"flexibility"*** and that, as a user, you'd need to be able to ***"tailor the quality criteria."*** Ratings for these features are shown in rows 9 and 7 of Table 5 respectively.

Concerning tool support for data extraction (F3-F06), many participants felt that ***"something to store all that information would be useful."*** In the context of an end-to-end tool, the ability to have extracted data ready to go ***"straight into the analysis"*** was also praised. Some participants, however, had a ***"hard time seeing how*** [the feature] ***would work properly in practice,"*** particularly when handling qualitative data. Concerning automated support for analysis (F3-F07), many participants felt this would be ***"very helpful"*** and would ***"save a lot of work."*** One participant felt that ***"less experienced reviewers would find*** [this feature] ***particularly useful."*** A number of participants mentioned that ***"data preparation"*** could be, potentially, more helpful. One participant stated it should be ***"mandatory for being able to get structured data out into different formats."*** Ratings for these features are shown in rows 2 and 8 of Table 5 respectively.

Some participants felt text analysis (F3-F08) would be a useful aid to certain stages of an SR (e.g. study selection), and had potential to ***"cut down on time for very big reviews."*** One participant felt that text analysis would become ***"increasingly more important as the complexity of the literature increases."*** Participants felt tool support for meta-analysis (F3-F09) was ***"very important"*** particularly for novices as, ***"for a lot of people undertaking a SR for the first time, meta-analysis is their biggest fear."*** Some participants, however, challenged the importance of support for meta-analysis as ***"not all reviews need it."*** Ratings for these features are shown in rows 19 and 10 of Table 5 respectively.

Participants felt that tool support for writing the report (F3-F10) would give reviewers a ***"starting point*** and a ***"good template."*** Many participants, however, felt such a feature would suffer since there are ***"so many different journals, which have so many different ways that they want you to present your work,"*** that

having a feature, which could ***"map to all of them,"*** would be ***"difficult."*** Concerning tool support for report validation (F3-F11), one participant felt that this might be a useful feature if, for example, ***"the validation itself is done by the team members, but the framework for the validation is generated by the tool, possibly through previous sets of criteria."*** Many participants, however, felt that there were already ***"plenty of resources"*** that already supported this aspect of an SR. Ratings for these features are shown in rows 20 and 18 of Table 5 respectively.

3.2.4 Feature Set 4 (F4): Process Management

Many participants felt support for multiple users (F4-F01) within a tool was really important. In particular, allowing users to collaborate within ***"large-scale teams"*** was considered very useful. Therefore, in order for other features such as study selection, data extraction and quality assessment to be fully supported by a tool, support for collaboration would need to be in place. Ratings for this feature are shown in row 1 of Table 5.

Many participants felt that tool support for document management (F4-F02) would be a useful feature. In particular, having the relationships between the papers and studies ***"closely integrated"*** would be ***"really helpful."*** Furthermore, such a feature might help transition the tool from a ***"reference manager to a study-based system."*** A key issue raised by one participant was copyright. With multiple users collaborating and sharing documents, problems concerning permissions/access of certain papers may occur. Ratings for this feature are shown in row 5 of Table 5.

Many participants felt a feature, which supports security (F4-F03), should be included in a tool. One participant argued, however, that since SRs deal with ***"published studies"*** that have ***"already been anonymised"***, security wouldn't be necessary. Another participant, however, felt security was important because ***"you might include unpublished stuff that the authors have let you use."*** Similarly, another participant noted that ***"some reviews use industry supplied data, which is not in the public domain."*** Tool support for role management (F4-F04) where, for example, you could ***"see all the people in the team and what their roles were"*** was generally considered a useful feature. One participant raised concerns about allowing others to see your role and contribution within the project. Another participant, however, points out that ***"it's not necessarily that you don't trust people to do a good job, it would just cut down the chances of a mistake."*** Ratings for these features are shown in rows 6 and 13 of Table 5 respectively.

Many participants felt that tool support for re-using data from past SRs (F4-F05) would be useful; particularly, when updating SRs (which ***"is happening more and more now."***) The potential for time-saving was also praised. In particular, speeding up quality assessment by including a previously assessed study (from a past SR) might mean that ***"you wouldn't have to quality assess it again."*** Similarly, participants note that it could also help during the search. For example, ***"you run the search and it automatically excludes any paper that was found in a previous SR."*** Ratings for this feature are shown in row 11 of Table 5.

4. DICUSSION

4.1 Tools Identified

As shown in Table 3, the most common type of tool identified by participants were reference managers. The systematic storage and management of citations is a critical part of any SR (in any domain) and it was, therefore, unsurprising that these types of tool were mentioned most frequently.

Interestingly two custom-built tools were reported. These tools (i.e. a web-based coding tool that supports collaborative study selection and a customised excel add-in that supports analysis), were developed by their respective review teams, as they felt that available tools did not provide sufficient support for the complexity of their reviews. It may be, however, that suitable tools were available but were not known to the teams. A web-based catalogue *(Systematic Review Toolbox*[2]), which aims to help reviewers identify appropriate tools, has been developed.

4.2 Feature Ratings

The set of features, ranked by level of importance, are shown in Table 5. Features considered by participants to be particularly important (i.e. features that received many ratings of Mandatory or Highly Desirable) include support for multiple users, data extraction and maintenance. Clearly, collaboration is a key aspect of SRs and is recommended for many stages in the process to ensure maximum reliability and validity.

Some features (i.e. F3-F01, F3-F02 and F4-F04) generated a wide range of opinions and, thus, resulted in little consensus amongst participants. We checked whether the lack of consensus could be explained by participants' different experience levels or areas of work. However, no patterns relating to these factors were found. One possible explanation could be that although some participants thought that tool support for a particular stage would be useful, they gave it a low rating because they were not able to imagine how such support could be provided (e.g. ***"I have a hard time seeing how that would work properly."*** and ***"it would be highly difficult to automate all that."***). The issue of financial payment for a tool (or, rather, lack of) also received varying opinions amongst participants. We had assumed that having a tool free of financial cost would be a positive characteristic. Results show, however, that many participants suggest some payment for a tool provides a degree of confidence in the reliability and longevity of the tool (see Section 3.2.1).

Features not considered particularly important include support for writing the report, text analysis and report validation. Therefore, results seem to suggest that tool support for the reporting phase of a SR is not a high priority for reviewers.

4.3 Comparing the Feature Ratings

This section compares features and importance levels identified by participants with those proposed for tools to support SRs in SE [8] (see the last column of Table 5). In particular, some of the key disagreements are discussed.

Generally, there was a good level of agreement between ratings. Comparing the modal value from the 13 participants with ratings proposed for SE showed no disagreements for 11 features and only slight disagreements (i.e. one level of importance higher or lower) for five features (see Table 5). Results suggest, therefore, that many of the frequently raised difficulties faced by reviewers are shared by researchers in most domains. Clearly there is considerable commonality between SRs in SE and other disciplines, so it is not surprising that there is some agreement about the importance of tool features. There are, however, notable differences relating to three features; namely, meta-analysis, role management and security.

As shown in Table 4, the modal response by participants for a feature which supports meta-analysis indicates a 'Highly Desirable' level of importance. We, in SE, on the other hand,

[2] http://systematicreviewtools.com

considered this feature only 'Nice-to-have.' This reflects the fact that few meta-analyses are undertaken within the SE domain because the differences among outcome metrics, analysis methods and experimental designs are too great to make statistical meta-analysis feasible. In Healthcare, however, where reviewers often extract and analyse data from randomized controlled trials, synthesis tools and, in particular, meta-analysis tools are more important. This feature is, therefore, an example of a context-dependent feature, where its relative importance is influenced by the particular SR-related issues associated with a specific domain.

There were also differences about the importance of support for security and role management. In SE, we rated support for role management as a 'Highly Desirable' feature. The modal response from participants, however, rated this feature as 'Nice-to-have.' This was somewhat surprising since support for multiple users was rated highly by both SE researchers and participants in this study. It was expected, therefore, that being able to manage those users within the context of a review would be important to users in other domains as well. For security, we rated this as a 'Desirable' feature. The modal response from participants, however, considered security features as 'Mandatory.' This higher level of importance might be explained by the, sometimes, sensitive nature of data that is included in a SR (i.e. patient or industry related data). This, again, may be an example of a context dependent feature. Furthermore, it should be noted that in both these cases the modal value was only four and responses were spread fairly evenly over most of the categories. This is another indication of a context dependent feature. Other features showing a similar pattern are development of the review protocol, report validation and text analysis.

4.4 Limitations of the Study

Semi-structured interviews rely heavily on the communication skills of the interviewer [15]. It is possible, therefore, that the quality of the data collected may be limited by the interviewer's lack of experience. This problem was at least partially addressed by performing a pilot interview (see Section 2.1.2). Furthermore, research suggests that people respond differently depending on how they perceive the interviewer ('*the interviewer effect')* [16]. Factors such as gender, age and the ethnic origins of the interviewer have a bearing on the amount of information people are willing to contribute [16]. In addition, participant's responses can be influenced by what they think the situation requires [17]. To try to address this, every effort was made to put participants at ease and to explain the purpose and the topics to be covered.

5. CONCLUSIONS

This study has explored the experiences and opinions of systematic reviewers in Healthcare and Social Science domains, with a particular focus on their use of and views about automated tools to support SRs; using, an interview-based survey.

21 software tools, which were each categorised into one of seven groups, were identified (see Table 4). Reference management tools were the most commonly mentioned forms of automated support. Special purpose tools (i.e. *EPPI-Reviewer* and *RevMan*) were the second most common. The top three most important features classified by participants were support for multiple users, data extraction and maintenance. The three least important features for a tool were support for writing the report, text analysis and report validation.

We compared the importance levels of features identified by participants with our ratings from an SE perspective. Generally, there was a good level of consensus, with only a small number of notable differences; specifically, ratings for meta-analysis, role management and security. However, we note that researchers wanting to use our tool evaluation framework should take care to determine the importance of context dependent features for their own particular circumstances, rather than using our weightings.

We plan to present the set of features and importance ratings to experts in SE for further refinement and validation.

6. REFERENCES

[1] Riaz, M., Sulayman, M., Salleh, N., & Mendes, E. (2010). Experiences conducting systematic reviews from novices' perspective. In *Proceedings. of EASE* Vol. 10, pp. 1-10.

[2] Babar, M. A., & Zhang, H. (2009). Systematic literature reviews in software engineering: Preliminary results from interviews with researchers. In *Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement*, pp. 346-355.

[3] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., & Khalil, M. (2007). Lessons from applying the systematic literature review process within the software engineering domain. *Journal of Systems and Software*, Vol. 80, no. 4, pp. 571-583.

[4] Carver, J. C., Hassler, E., Hernandes, E., & Kraft, N. A. (2013). Identifying Barriers to the Systematic Literature Review Process. In *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on* (pp. 203-212). IEEE.

[5] Tsafnat, G., Dunn, A., Glasziou, P., & Coiera, E. (2013). The automation of systematic reviews. *BMJ: British Medical Journal*, *346*

[6] Felizardo, K. R., MacDonell, S. G., Mendes, E., & Maldonado, J. C. (2012). A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews. *Journal of Software*, *7*(2), 450-461..

[7] Marshall, C., & Brereton, P. (2013). Tools to Support Systematic Literature Reviews in Software Engineering: A Mapping Study. In *Empirical Software Engineering and Measurement, 2013 ACM/IEEE International Symposium on* (pp. 296-299).

[8] Marshall, C., Brereton, P., & Kitchenham, B. (2014). Tools to support systematic reviews in software engineering: A feature analysis. In *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering* (p. 13). ACM.

[9] Staples, M., & Niazi, M. (2007). Experiences using systematic review guidelines. *Journal of Systems and Software*, *80*(9), 1425-1437

[10] Lethbridge, T. C., Sim, S. E., & Singer, J. (2005). Studying software engineers: Data collection techniques for software field studies. *Empirical software engineering*, *10*(3), 311-341.

[11] Myers, M. D., & Avison, D. (1997). Qualitative research in information systems. *Management Information Systems Quarterly*, *21*, 241-242.

[12] Robson, C. (2002). *Real word research*. Oxford: Blackwell.

[13] Patton, M. Q. (1990). *Qualitative evaluation and research methods*. SAGE Publications.

[14] Miles, M. B, Huberman, A. M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook*. Sage.

[15] Clough, P., & Nutbrown, C. (2012). *A student's guide to methodology*. Sage.

[16] Denscombe, M. (2010). *The Good Research Guide: For Small-Scale Social Research Projects: For small-scale social research projects*. McGraw-Hill International.

[17] Gomm, R. (2004). *Social research methodology*. New York: Palgrave Macmillan.