e-Informatica Software Engineering Journal, Volume 12, Issue 1, 2018, pages: 79–115, DOI 10.5277/e-Inf180104

Tool Features to Support Systematic Reviews in Software Engineering – A Cross Domain Study

Chris Marshall*, Barbara Kitchenham**, Pearl Brereton**
* *York Health Economics Consortium Ltd., University of York*
** *School of Computing and Mathematics, Keele University*
chris.marshall@york.ac.uk, b.a.kitchenham@keele.ac.uk, o.p.brereton@keele.ac.uk

Abstract

Context: Previously, the authors had developed and evaluated a framework to evaluate systematic review (SR) lifecycle tools.
Goal: The goal of this study was to use the experiences of researchers in other domains to further evaluate and refine the evaluation framework.
Method: The authors investigated the opinions of researchers with experience of systematic reviews in the healthcare and social sciences domains.
They used semi-structured interviews to elicit their experiences of systematic reviews and SR support tools.
Results: Study participants found broadly the same problems as software engineering (SE) researchers with the SR process. They agreed with the tool features included in the evaluation framework. Furthermore, although there were some differences, the majority of the importance assessments were very close.
Conclusions: In the context of SRs, the experiences of researchers in other domains can be useful to software engineering researchers. The evaluation framework for SR lifecycle tools appeared quite robust.

Keywords: software engineering, systematic review tools, cross-domain survey, qualitative analysis

1. Introduction

A systematic review (SR) is a formal, repeatable method for identifying, evaluating and interpreting all available research regarding a particular problem or topic of interest. The rigorous and impartial nature of a systematic review increases the scientific value of its findings in comparison with expert-based literature reviews [1–3], which makes it an important tool for obtaining and appraising evidence in a reliable, transparent and objective way. Systematic reviews were first established in Clinical Medicine [4, 5]. Medical researchers defined the systematic review process to help mitigate the drawbacks of a conventional literature review [1]. A cautionary note needs to be added here that systematic reviews have received some criticism, in particular, that they are sometimes of quite poor quality and can reap high rewards in terms of citation counts despite biases and vested interests [6]. Also, the synthesis of outcomes, particularly in the software engineering field, can be problematic [7].

With a growing emphasis on empirical software engineering research, the popularity and importance of systematic reviews has grown considerably [8,9]. Despite their potential usefulness and importance to empirical software engineering research, undertaking a systematic review remains a highly manual and labour intensive process resulting in the possibility of process errors (such as misclassifying primary studies or wrongly excluding a primary study). In particular, there are challenges concerning the study selection, data extraction and data synthesis stages, amongst other collaborative activities

[10–14]. Furthermore, systematic reviews have only recently been adopted by software engineering researchers, and, as a result, there have been problems surrounding the provision of appropriate support for novices [11–14]. These drawbacks, along with others, make the systematic review methodology a prime candidate to benefit from an automated tool support [12–16].

In our experience, it is certainly possible to undertake a systematic review without too much automation. Furthermore, Kitchenham and Brereton were involved in the revision of the systematic review guidelines that emphasised human processes and decision making [17]. Thus, the authors believe it is important to have a balanced view of the benefits of automating the systematic review process. In this study, attitudes to automation in domains that have more practical experience of systematic reviews and their automation than software engineering were investigated.

In earlier research, the authors developed and validated a framework for evaluating tools intended to support the full systematic review process [18]. The framework was based on a set of tool features identified as important for systematic reviews in software engineering based on the SR guidelines, the authors experiences, and the experiences of other SE researchers reported in the literature. This paper reports on the results of a cross-domain study of researchers who undertake systematic reviews as part of their normal research practice, which was intended to further validate our framework.

Some of this research has already been reported [19], however, this paper provides a more detailed analysis of our study results relating to the impact of participant's experience level and the identification of trends among their comments (the additional analyses are itemized in Section 4.2.3).

Section 2 describes the evaluation framework and explains particular interest in systematic review lifecycle tools. Section 3 discusses SE research that used results from other disciplines, that investigated benefits and problems with the SR process, and discussed tools to support the SR process. Section 4 discusses the goals of the study and the methodology used to address these goals. Section 5 presents the results of the cross domain study. Section 6 discusses the results and conclusions are presented in Section 7.

2. Framework for evaluating systematic review lifecycle tools

The developed evaluation framework was aimed at evaluating tools that support the full SR process in contrast to tools that assist a specific process or task. The reasons why the authors concentrated on these tools and developed a multi-criteria decision making framework are:

1. Large SRs are complex and hard to manage. In order to support the production and update of large scale (possibly distributed) SRs, standard tools such as reference managers and spread sheets become increasingly cumbersome and error prone. The developers of the SLuRp tool say "Our experience is that in order to produce reliable valid results, more than one reviewer is required. Maintaining large amounts of data in a team with several reviewers is time-consuming and error-prone. These errors are difficult to identify and eliminate without the use of a specific SLR tool like SLuRp." [20].
2. SR lifecycle tools cannot be easily evaluated. Tools that support a specific process or task can be evaluated in isolation using experiments or small case studies, in contrast SR lifecycle tools are more difficult to evaluate because they span the entire lifecycle of a review from initial planning to final reporting and even subsequent updating. This lifecycle process is made up of a series of individual processes that interact with one another and require validation and sometimes reworking. To maintain clarity within this paper we shall refer to these tools as SRLC (Systematic Review LifeCycle) tools.
3. Currently, there is interest among software engineering research groups in building SRLC tools. The initial search found four such tools [21] and later another one was found [22]. This interest suggests it is an appropriate time to consider how to evaluate such tools.

4. Adopting such tools is a major commitment. Research groups need to have some confidence that any tool they adopt will be able to support the sort of systematic reviews they perform and the way in which they manage their systematic review process.

The evaluation framework was based on feature analysis as proposed by the DESMET project [23]. Feature analysis is a type of multi-criteria decision analysis. It is a subjective method of evaluation. It is intended to provide a means of organising a subjective evaluation of a tool and making the components of that evaluation clear to, and auditable by other potential tool users.

In the context of SRLC tools, members of the same software engineering research group were expected to be other potential users. Thus, the authors envisage that our framework would provide a means by which researchers could make an informed, defensible decision together. One particular benefit of the DESMET feature analysis method is that it requires the users of the method to refine the evaluation process depending on their own requirements. Specifically it involves users of the feature analysis defining what they require of an acceptable tool with respect of each feature. So the users of the framework do not just evaluate a tool against a set of features, they also need to define the importance of each feature in terms of its importance to them. This means that although an evaluation exercise could involve a series of different candidate SRLC tools, the tools are not so much compared with each other as with the research group's specific set of requirements. This provides a feature analysis with a built-in element of flexibility, which allows users to tailor an evaluation to their own circumstances. The details of the initial version of the framework and its evaluation can be found in [18].

3. Related work

In 2004, Kitchenham et al. [24] introduced the concept of Evidence-Based Software Engineering (EBSE) as an approach to integrate academic research with industry needs and improve decision making regarding the development and maintenance of software. This initiative was based on the concept of Evidence-Based Medicine. Kitchenham et al. recommended the use of systematic reviews to support EBSE. Subsequently, Kitchenham [25] developed a set of guidelines for undertaking systematic reviews based on health care guidelines, which were updated in 2007 [3]. The 2007 guidelines were influenced both by a study of the use of systematic reviews in other disciplines and by guidelines developed for the social sciences [26], and were adapted to better reflect the use of systematic reviews in software engineering. A further update to the guidelines was released in 2015 (see Section III of [17]). This version of the guidelines was strongly oriented to addressing software engineering issues. In particular, it included more information about managing the collaboration aspects of systematic reviews and methods for synthesizing the results of quantitative and qualitative studies.

Since the release of the original guidelines and the publication of systematic reviews in software engineering journals, there has been substantial literature discussing how the software engineering community performs systematic reviews and how the process could be made more efficient. Kitchenham and Brereton [9] summarized this literature in a systematic review that included 45 papers published between January 2005 and June 2012. This study summarized the perceived benefits of doing SRs, problems SE researchers had found when undertaking SRs and the advice and techniques intended to assist in performing SR tasks. However, most of this work was fairly inward looking with relatively few papers discussing ideas from outside the software engineering community. The main exceptions were: Torres et al. [27] who trialled the methods of sentence classification used in scientific papers on SE data; Felizardo et al. [28] who undertook a cross-discipline mapping study to investigate the use of visual data mining techniques to support SRs; Ramampiaro et al. [16] who discussed the use of techniques from information retrieval and text mining to support the development of meta-searcher capabilities.

Since 2012, there have been two initiatives to investigate tools to support systematic reviews in software engineering undertaken independently by two groups of researchers:

1. Marshall and Brereton [21] performed a mapping study to identify tools available to support SRs in the SE community and identified 13 different tools of which three were intended to support the full lifecycle (i.e. were SRLC tools). They also introduced the systematic review toolbox which is a catalogue of tools to support systematic reviews [29]. All three authors of this paper presented an evaluation framework intended to assess SRLC tools and reported the results of using the evaluation framework to evaluate four different SRLC tools developed in the software engineering community [18]. They also published a preliminary analysis of data from our study of researchers in health care and social science [19].
2. Carver et al. [14] reported barriers to the SR process based on 52 responses to an online survey sent to authors who published SRs in SE venues and qualitative experiences from eight PhD students. Hassler et al. [30] reported the result of a community workshop that identified and ranked 37 barriers to the SR process that could be grouped into themes related to the SR process, primary studies, the practitioner community and tooling. Subsequently, Hassler et al. [31] reported a workshop-based study of SR tool needs based on information provided by 16 software engineering researchers. They compared the result of their study with the published preliminary results of our study of tool features [19].

4. Goals and methodology of the cross-domain study

4.1. Goals

The objective was to see if the experiences of researchers from domains that have more extensive experience in the use of systematic reviews would be valuable to software engineering (SE) researchers and SR tool designers. In particular, the goals of this study were:

1. To assess whether the SR experiences of researchers in other domains are relevant to those of SE researchers.
2. To explore what tools were currently available and used to support systematic reviews in other domains.
3. To compare the features and importance levels identified by the participants with those in this SRLC tool evaluation framework.

These goals could best be addressed by a qualitative study aimed at eliciting the experiences of systematic reviewers on other domains. For this reason, Marshall undertook a series of cross-domain, semi-structured interviews, which were designed to explore the experiences and opinions of systematic reviewers in other domains (outside of software engineering) about support tools.

It should be noted that, as is common with qualitative studies, the goals are fairly general and do not map to detailed research questions and hypotheses. They exist to scope the qualitative study not to define questions and metrics.

4.2. Methodology of the cross-domain study

This section reports on the research strategy and research process.

4.2.1. Research strategy

Semi-structured interviews were used to elicit the opinions of researchers about systematic review support tools. This means that a number of questions were identified to ask the participants and also to encourage a discussion about the issues to follow the directions that the participants wanted. Semi-structured interviews were selected instead of a self-administered questionnaire for two main reasons:

1. The awareness that terminology differs between different domains and that face-to-face interviews would allow potential misunderstandings to be identified and resolved.
2. The need for certainty that the identified participants had appropriate experience.

Since the study was qualitative, no detailed research questions or research hypotheses were

derived, data collection and analysis procedures arose from the research goals and resulted from the expectations that:

- Viewpoints of researchers working in domains where systematic reviews are well-understood and considered a standard research practice would be valuable to software engineering researchers.
- Viewpoints of novices and experts would differ.
- Tool feature preferences of participants would be influenced by the type of systematic review they undertook.

Thus, the selected study participants covered various domains, different levels of researcher experience and different systematic review types. The aim was to interview both senior and junior researchers from several different domains. Originally, six topic areas were considered: Clinical Medicine, Criminology, Education, Empirical Psychology, Nursing & Midwifery, and Primary Care, however, in practice two high level domains became the focus: social sciences and health care. No restriction was placed on whether the researchers had performed quantitative or qualitative reviews. The goal was to interview researchers with experience of both types of review because issues related to data extraction and aggregation are very different for qualitative and quantitative reviews.

The inclusion criteria for participants were as follows:

- Researchers used systematic reviews as part of their standard research process.
- Researchers had a wide range of roles and responsibilities.

Initially it was planned to provide a *theoretical sample* covering the six topic areas. The theoretical sample is a type of purposeful sampling where researchers are seeking incidents/reports of the phenomenon they are studying which will supply useful data [32]. However, after the data was collected and tabulated, it was found out that the coverage of three dimensions had been achieved:

- The two domains (health care and social sciences).
- Three experience levels corresponding to 1–5 SRs (i.e. Low), 6–15 SRs (i.e. Medium), and > 15 SRs (i.e. High)[1].
- Types of SRs performed: Quantitative and Qualitative.

This coverage of three important dimensions allowed to extend the analysis of the study results.

4.2.2. Research process

Marshall developed the semi-structured interview plan after discussions with Kitchenham and Brereton. He, then, piloted the semi-structured interview procedure with a PhD student who had undertaken two SRs. This led to some changes to the delivery and sequencing of questions and also confirmed the expectation that interviews would take approximately 45 minutes. The interview plan included questions related to four concerns:

- Group 1: questions relating to the participant's background and domain.
- Group 2: questions about the participant's experience of undertaking systematic reviews.
- Group 3: questions about the participant's use of systematic review tools.
- Group 4: questions about SRLC tool features and their importance levels.

The detailed interview questions are reported in Appendix A.

In the research a combination of convenience and snowballing sampling techniques was used to identify 49 potential participants. Finally, 13 researchers from six institutions agreed to take part. Marshall carried out the interviews between June 2014 and September 2014. Prior to the interview, each participant was sent an Interview Preparation Form (see Appendix B). 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 and recorded using a digital audio recorder. Marshall took notes throughout each interview. The shortest interview took 32 minutes and the longest interview lasted 68 minutes, with an average of 45 minutes.

[1] For some analyses, only two experience levels were used: low corresponding to 1–5 SRs and high corresponding to 6+ SRs, giving us six relative novices and 7 relatively highly experienced participants.

Marshall processed the raw data (i.e. recordings, field notes) prior to analysis. The field notes were reviewed and full transcriptions of each interview were produced. For this study, transcripts aimed to reflect a straightforward summary of the main ideas, which were presented by a fluently spoken participant. The transcripts did not include any mispronunciations, pauses or word emphases which might have occurred during the interview. In total, the interviews generated approximately 10 hours of audio recordings, each taking between five and six hours to fully transcribe.

4.2.3. Data analysis

Marshall conducted the initial analysis concurrently with data collection, as recommended by Miles et al. [33]. The initial analysis was based on tabulating responses in order to identify:

- Challenges participants faced when doing systematic reviews.
- Tools used by participants.
- Positive and negative experiences of tools.
- Participant opinions of the importance of the features included in the evaluation framework compared with the importance assigned to them.

Kitchenham and Brereton reviewed all the tables for consistency. Initially, comments were tabulated verbatim (as reported in [19]). Subsequently, all three authors reviewed the initial analyses and realized from the biographical data that the actual sample included participants with a range of experiences that would enable additional analyses of the data. This resulted in Kitchenham and Brereton undertaking additional analyses (beyond those reported in [19]) that are reported in this paper and which are described below:

1. A summary of the general problems/issues reported by participants and cross-referenced to the SE literature in order to identify similarities and differences between the SE domain and health care and social services domains.
2. An analysis of the comments by individual participants concerning general systematic review tools and systematic review lifecycle tools. This was intended to give a balanced view of the advantages and disadvantages of automating the SR process.
3. A thematic analysis of the comments related to systematic review lifecycle tool features to provide some quantification of trends. Details of the coding process and an example of how the codes were established is provided in Appendix C.
4. An investigation of whether participants' responses were influenced by their experience of undertaking systematic reviews.
5. An investigation of whether participants' responses were influenced by the type of systematic review they performed.
6. An investigation of the importance of factors related to the usability and ease of installation. This was intended to clarify the features required to represent tool usability.
7. A comparison of our results with other related SE studies. This was intended to highlight similarities and differences between the SE domain and health care and social services domains, particularly in the context of participant experience.

5. Results of the cross-domain study

The details of the participants' roles, research domains and SR experience are given in Table 1. The participants covered a range of disciplines, including nursing, psychology and education in the domains of health care and social sciences, and a variety of roles, including research associate[2], lecturer, senior lecturer[3], information officer/specialist and professor. The term information officer/specialist is used to identify someone whose main role is to provide support for the search process of systematic reviews. This job title confirms the importance of systematic reviews in the health care and social sciences domains.

The group of 13 participants in this study had experience of different types of a system-

[2]Usually a post-doctoral researcher working on a funded project and employed on a fixed-term contract.
[3]An academic position in the UK corresponding to an Associate or Assistant Professor in the USA.

Table 1. Cross domain study participant information

ID	Role	Domain	No. of SRs	Type of SR
P01	Research Associate	Health care (Primary Care)	6–10 (Medium)	Both
P02	Research Associate	Health care	1–5 (Low)	Quantitative
P03	PhD Student	Health care(Physiotherapy)	1–5 (Low)	Qualitative
P04	Senior Lecturer	Health care (Health Psychology)	1–5 (Low)	Qualitative
P05	Information Officer	Health care	11–15 (Medium)	Quantitative
P06	Lecturer	Health care (Nursing)	1–5 (Low)	Quantitative
P07	Lecturer	Social Science (Educational Psychology)	1–5 (Low)	Quantitative
P08	Information Officer	Social Science	> 15 (High)	Both
P09	Professor	Social Science	> 15 (High)	Both
P10	Systematic Reviewer	Social Science (Public Health)	6–10 (Medium)	Both
P11	Research Associate	Social Science (Education Technology)	1–5 (Low)	Both
P12	Professor	Social Science (Education & Child Psychology)	> 15 (High)	Qualitative
P13	Information Specialist	Health care	> 15 (High)	Both

atic review, different levels of experience, and different domains of interest. Specifically:

- In the health care domain, there were seven participants; two concentrated on qualitative reviews, three on quantitative reviews, and two conducted both types of review. Four of the participants were relative novices who had conducted 1–5 reviews, but of the remaining three, one had performed 6–10 reviews, one 11–15 reviews and one > 15 reviews.
- In the social science domain, there were six participants; one concentrated on qualitative reviews, one on quantitative reviews and four conducted both types of reviews. Two of the participants were relative novices (1–5 reviews), one had conducted 6–10 reviews and three had conducted > 15 reviews.

Thus, there was a good coverage of the factors expected to influence the participants' responses in these semi-structured interviews: domain, experience and type of review.

5.1. Issues faced by researchers in other domains

An important issue when evaluating the participants' answers was to determine whether their experiences were relevant to software engineering researchers. In order to investigate this issue the participants were asked about the main challenges and specific problems they had faced when conducting systematic reviews.

Table 2 summarizes the challenges and issues mentioned by the participants. In columns three and four, it was identifies whether these issues had been raised in the SE literature. Column 3 refers to issues that are general problems and identifies whether they are raised in [9] or in [14]. Column 4 refers to process factors discussed in the recent SE related text book which [17] includes an update of guidelines for systematic reviews in software engineering. Column 5 identifies the participants who made a comment and Column 6 specifies their experience.

Table 2 identifies three high level concerns (i.e. those unrelated to specific SR activities) that were mentioned 11 times by six different participants. It is interesting that none of those participants had the highest level of experience. Possibly after doing many SRs, researchers overcome their initial perception of the difficulty of SRs, or, in the case of perceiving SRs to be *Time Consuming*, become inured to the issue.

In the case of the challenges related to specific SR processes, Management issues produced the most comments, both in terms of unique issues raised (of which there were seven), and in terms of the total number of comments (of which there were 13) which were made by eight different participants. It is interesting that the SE literature on SR challenges summarized by

Table 2. Challenges and specific issues reported in interviews

Main Challenges	Interview Specific Issues	Discussed in [9] or [14]	Discussed in [17]	Id	Experience
Search Process	Search String translation	Yes	No	P01	M
	Inconsistency with terminology	Yes	No	P01, P06, P09, P10	M, L, H, M
	Time consuming	Yes	No	P03	L
	Developing the search strategy	No	Yes	P04, P08, P10, P13	L, H, M, H
Time consuming	General	Yes	No	P02, P03, P04, P05, P07, P11	L, L, L, M, L, L
No Standardization	General	Yes	No	P02	L
High Difficulty	General	Yes	No	P02, P03, P07, P11	L, L, L, L
Management	Managing large-scale SRs	No	Yes	P04, P05, P09	L, M, H
	Transparency	No	Yes (reporting)	P05	M
	Handling duplicates	Yes	Yes	P06, P07	L, L
	Collaboration	Yes	Yes	P06, P07, P12, P13	L, L, H, H
	Negotiating with policy makers	No	No	P10	L
	Relationships between studies & papers	No	Yes	P12	H
	Version control	No	No	P12	H
Analysis	Qualitative Analysis	Yes	Yes	P05	H
	Meta-analysis	No	Yes	P06, P10	L, M
Study selection & screening	Resolving disagreements	Yes	Yes	P06	L
	Managing the criteria	Yes	Yes	P12	H
	Criteria consistency across multiple coders	Yes	Yes	P12, P13,	H, H
	General	Yes	Yes	P05, P08	M, H
Quality assessment & critical appraisal	Resolving disagreements	No	Yes	P06	L
	Managing the criteria	Yes	Yes	P12	H
	Criteria consistency over multiple coders	Yes	Yes	P12, P13	H, H
	Assessing quality of study not the paper	Yes	Yes	P12	H
	General	Yes	Yes	P11	L
Protocol Development	Developing research questions	Yes	Yes	P08, P10	H, M
	General	Yes	Yes	P10	M
Producing Report	Formatting references	No	No	P13	H
	General	No	Yes	P10	M
Validation	Knowing when to check for consistency	No	Yes	P12	H

Kitchenham and Brereton [9] did not concentrate on these issues, although they feature more extensively in Hassler et al. [31] and in the latest SR guidelines [17]. This might reflect the greater maturity in the health care and social sciences domains and allows to identify an area which will become more important for SE researchers in the future. Other activities that attracted numerous comments are:

- The search process, with a total of 10 comments about four different issues which were made by eight different participants of all experience levels.
- The study selection and screening process, with a total of six comments consisting of four different issues made by five different participants but including only one comment from a participant with low experience levels.
- The quality assessment and critical appraisal process, with a total of six comments about five different issues made by four participants including two low experience and two high experience participants.

These issues were discussed in the SE literature and the number of high experience participants that mentioned these issues suggests that they remain a challenge irrespective of experience levels.

Three challenges that had no overlap with SE challenges or guidelines are:

1. **Negotiating with policy makers.** Researchers in other domains are often commissioned to do systematic reviews and may, therefore, need to negotiate with the policy makers who commissioned the study. In SE, there are no policy makers who commission systematic reviews, so currently this is not an issue.
2. **Version control.** Systematic reviews in SE are usually considered one-off pieces of research, so are not generally concerned about version control. Researchers in other domains produce reports for policy makers and may need to update those reports periodically, so version control is more important.
3. **Formatting references in the final report.** Although not mentioned as a specific issue in SE papers, it is certainly the case that outputs from different digital libraries are not usually equivalent and can be difficult to integrate, unless converted into an intermediate format compatible with reference manager systems such as EndNote or BibTeX.

These challenges were each mentioned only once.

Overall the results in Table 2 suggest that researchers in other domains face many of the same issues as software engineering researchers. It can be concluded, therefore, that their experiences of tool support for SRs are relevant to those of researchers in software engineering. Furthermore, these results suggest that challenges remain even for highly experienced researchers and, in particular, management issues should be expected to become more important as SE researchers become more experienced. This is likely to happen because as researchers become more experienced with the SR methodology, they will be tempted to take part in more complex and larger scale SRs.

5.2. Tools used in other domains

Table 3 shows the tools that participants reported using to assist their SRs. All but three of the participants (i.e. P10, P11 and P12) reported using reference managers, with RefWorks and EndNote being the most frequently used ones. Six participants used tools that assist analysis including Microsoft Excel, statistical software, meta-analysis tools, and textual analysis tools. Seven participants used SR lifecycle tools: four used RevMan and three used EPPI-Reviewer.

Table 4 reports the positive comments participants made about the tools, other than SRLC tools, they used. Both RefWorks and EndNote attracted a large number of positive comments, seven and nine, respectively. However, the comments were generated by three of the four RefWorks users but only two of the five EndNote users.

On the negative side, as shown in Table 5, RefWorks was criticised for its 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 are a lot of things compacted onto one screen."). The criticism of EndNote was about whether it could effectively handle large numbers of papers/studies ("people

Table 3. Use of SR lifecycle tools and other tools

ID	SR Lifecyle tool	Other tools
P01	RevMan	RefWorks
P02	RevMan	RefMan, STATA, Microsoft Word
P03	None	RefWorks
P04	None	EndNote, NVivo, Microsoft Word
P05	RevMan	RefWorks, Endnote
P06	None	RefWorks, Federated Search Tool
P07	RevMan	Mendeley, Microsoft Excel, Mplus, NVivo, Custom Web-based coding tool, MetaEasy, MetaLight, SPSS
P08	EPPI-Reviewer	EndNote, RIS conversion tool
P09	EPPI-Reviewer	EndNote, ProCite, Microsoft Word
P10	EPPI-Reviewer	None
P11	None	None
P12	None	Microsoft Excel, NVivo, Altal.ti, Mendeley
P13	None	EndNote, Mendeley, PubReMiner, RefMan

are concerned that it doesn't have the capacity to deal with the huge numbers of references.").

Table 6 reports the positive comments about the SRLC tools. The version of EPPI-Reviewer current when the interviews took place was EPPI-Reviewer 4. It was a comprehensive single or multi-user web-based system for managing systematic reviews across health care and social science domains. During the interviews, the participants were very positive about the variety of ways in which the tool can support the systematic review process (see Table 6). For example, EPPI-Reviewer's support for study selection uses text mining to prioritise the most relevant studies, so those are viewed first. It allows the review team to start the full data extraction of the studies before finishing the screening. Its support for thematic analysis uses visualisation techniques to depict the relationships between concepts.

On the negative side, as shown in Table 7, the participants felt EPPI-Reviewer had a steep learning curve ("It's not something you can just pick up and use instantly.") and that it "takes a while to learn all of the different things." In addition, two participants felt that training could be improved.

RevMan primarily supports the preparation and maintenance of Cochrane Reviews, although, it can be used to support other reviews. As can be seen in Table 6, the participants appreciated its good support for statistical analysis techniques, in particular meta-analysis and its support for protocol development.

However, on the negative side some users felt restricted by the tool at times, 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.") (see Table 7). Other users also felt confused by the tool and felt it was all a bit too complicated.

Both tools exhibit features of particular relevance to the domain they were developed for, i.e. EPPI-Reviewer was developed by social scientists and, therefore, provides good support for qualitative analysis. In contrast, RevMan was developed by the Cochrane group primarily to support reviews of randomised controlled trials (RCTs), which are formal medical experiments where experimental subjects are real patients suffering from a specific illness. The reason why RevMan is able to provide support for protocol development is that primary studies should all follow a similar RCT process. Similarly, most RCTs are capable of being synthesized quantitatively, which explains the support for formal meta-analysis.

These results, together with those reported in Table 3, suggest that the users of RevMan may also need to use Reference Manager tools and ad-

Table 4. Participants comments on tools – positives

Tool	Comment	Participant
RefWorks	Okay (Better than doing them by hand)	P01
	Helped manage the search process	P03
	Removes duplicates	P03, P06
	Useful for managing study selection	P03, P06
	Useful for traceability	P03
	Helped share the work load between multiple reviewers	P03
	Useful for handling large numbers of studies	P03
	Able to classify studies using folder	P06
EndNote and EndNote Web	Helps manage the search process	P04, P05
	Links with several databases	P04
	Web-based allowing remote access	P04
	No financial payment required (for EndNote Web)	P04
	Can be used, unconventionally, to support study selection	P04
	Easier to use than RefWorks	P05
	Handles duplicates effectively	P05
	Creates individual databases for each SR project	P05
	Help with search strategy	P05
RefMan	It was OK	P02
Mendeley	Supports collaboration	P07
	Good support for version control	P12
	No financial payment required	P13
Federated search tool	Searches multiple sources	P06
	Useful for piloting search	P06
PubReMiner	Useful for developing protocol	P13
	Helps identify key journals	P13
Custom web-based tool	Supports multiple users (collaboration)	P07
	Exports data into other formats	P07
	Supports role management	P07
STATA	Good usability	P02
	Easier to use than RevMan	P02
NVivo	Helps find themes & trends across papers	P04
MetaEasy	Calculates effect sizes for individual studies	P07
Microsoft Excel	Clear presentation of data	P07
Microsoft Word	Supports protocol development	P02, P04, P09

vanced analysis tools. Although two of the users of EPPI-Reviewer reported using other tools, neither reported to need other advanced analysis tools. Furthermore, one user of EPPI-Reviewer did not report using any other tool. Thus, it seems that EPPI-Reviewer offers more complete support for the systematic review lifecycle than RevMan.

Of the two SRLC tools, EPPI-Reviewer is likely to be the most promising one for adoption by software engineers. However, it is possible that it is too much oriented to the requirements of the social sciences domain to be readily usable by software engineering researchers.

5.3. Importance of different features for SRLC tools

Finally, the participants were presented with a list of the features which had included in the

Table 5. Participants comments on tools – negatives

Tool	Comment	Participant
RefWorks	Problems with importing search results	P01
	Managing paper-study relationships is confusing	P01
	Not an ideal tool	P03
	Difficult for new users	P03, P06
	Poor usability, user interface	P03, P06
	Lost work	P03, P06
	Difficult to set up	P03
	One database for all reviews – so messy	P05
	Handles duplicates poorly	P05
	Less useful as number of papers increases	P05
	Poor export facility	P05
	Problems formatting references	P06
	Frequent major updates to user interface	P06
	Problems with search engine and database compatibility	P06
EndNote and EndNote Web	Not compatible with all databases	P04
	Extraction can be a bit clunky	P04
	Less useful as number of references increases	P05, P13
	Poor export facility	P05
	Trust issues (Web version is online and free)	P13
RefMan	Unnecessary for small numbers of papers	P02
	Problems formatting references	P13
	Problems with maintenance and support	P13
	Not very effective	P13
	Poor support for collaboration	P13
Mendeley	No version control	P07
	Copyright concerns	P13
Federated search tool	Searches multiple sources	P06
MetaEasy	Poor tool integration	P07
MetaLight	Difficult to use	P07
Microsoft Excel	Not that useful	P07
	No support for version control	P12
	Problems with interface	P12
	Doesn't support complex SR tasks	P12
	Too generic	P12

evaluation framework for SRLC tools. The participants were asked to rate the features on a five point ordinal scale:

1. Mandatory – meaning that the feature was essential in any tool aiming to support the SR lifecycle.
2. Highly desirable – meaning that although not mandatory, such a feature is extremely important in a SRLC tool.
3. Desirable – meaning that the feature would be useful for most researchers.
4. Nice-to-have – meaning the feature might be useful, but its omission would not seriously affect the tool's value to its users.
5. Not needed – meaning the feature is unnecessary and there is a danger that the feature would increase the complexity of the tool without adding any useful facilities.

The participants were also asked to identify any important features which had been overlooked.

The counts of the importance ratings of the features given by the 13 participants are pre-

Table 6. Participants comments on SR lifecycle tools – positives

Tool	Comment	Participant
RevMan	Good support for statistics & meta-analysis	P01, P05
	Support for protocol development	P01
	Nice chart generation	P05
EPPI-Reviewer	Supports the whole process	P08
	Good support for study selection	P08, P09
	Supports qualitative analysis (thematic analysis)	P09
	Helps manage the search process	P08, P09
	Generates tables and charts to be used in the report	P08
	Flexible coding system	P09
	Allows data extraction in tandem with study selection	P09
	Exports data into other formats	P09
	Supports basic meta-analysis	P09
	Supports role management	P09
	Customisable interfaces	P09
	Supports re-use of data from past SRs	P09
	Good support for "tedious" bits of SR process	P10
	Good support for document management	P10
	Supports inter-rater reliability	P10
	Easy to use	P10

sented in Table 8, where the bold number is the modal response rating for the feature.

The points raised by the participants during the discussion of the features are summarized below. The features relating to the same overall concern are grouped together.

5.3.1. Support for SR tasks

SRLC tool features related to the tasks needed to be performed in a systematic review are labelled SRT1 to SRT11 in Table 8.

Protocol management

Table 9 identifies the main issues participants raised when discussing protocol development and validation. The column labelled "Participants" identifies the number of participants who made comments related to each of them and the column labelled Experience identifies the experience level of the participants. This table includes the issue referred to a *Viability* which was only mentioned by one person in the context of protocol development and validation. It was included here because it referred to the concern that the feature might not be capable of implementation, which was mentioned by many other participants during discussions of other SR support tools.

With respect to support for developing the review protocol, participants' views differed (see Table 8 row SRT1). Four participants thought it would be used particularly for version control, while two felt it would be useful for complex projects (i.e. large teams). Three participants, however, were unsure of its usefulness since they simply used Microsoft Word to track changes. Another participant pointed out that the Cochrane Handbook assisted with protocol development.

Participants' views also differed with respect to the value of tool support for protocol validation (see Table 8 row SRT2). The two modal responses were Desirable (five participants) and Not Needed (five participants). Two participants thought it would help avoid missing anything. However, two other participants felt that introducing automation might be over-complicating the process. In addition, two participants mentioned problems with existing approaches to protocol validation that enforced protocol standards in the context of registering Cochrane reviews and submitting proposals to professional bodies.

Table 7. Participants comments on SR lifecycle tools – negatives

Tool	Comment	Participant
RevMan	Most features locked out if not doing a Cochrane review	P01,
	Not flexible enough	P02, P07
	Doesn't support many important aspects of SRs	P05
	Limited support for reporting phase	P05
	Confusing	P07
	Over restrictive conceptual model	P07
	Expensive	P07
	Limited support for developing the protocol	P07
	Not nicely integrated	P07
EPPI-Reviewer	Problems importing search results	P08
	No support for searching	P08
	Difficult to learn	P09
	Limited training support for novices	P09, P10
	No support for protocol development	P09
	No support for network meta-analysis	P09
	Limited information about updates	P10

Table 8. Importance of features

ID	Feature	Mandatory	Highly Desirable	Desirable	Nice	Not needed	Our Assessment
SRT1	Protocol Development	2	**4**	2	3	2	Desirable
SRT2	Protocol Validation	1	1	**5**	1	**5**	Desirable
SRT3	Search Process	3	**4**	3	3	0	Highly Des
SRT4	Study Selection	5	**6**	2	0	0	Highly Des
SRT5	Quality Assessment	5	**7**	1	0	0	Highly Des
SRT6	Data Extraction	**7**	5	1	0	0	Highly Des
SRT7	Data Synthesis	5	**7**	1	0	0	Highly Des
SRT8	Text Analysis	0	3	2	**5**	3	Nice
SRT9	Meta-analysis	4	**5**	2	2	0	Nice
SRT10	Reporting	0	2	**7**	4	0	Nice
SRT11	Report Validation	0	3	3	3	**4**	Nice
SRM1	Multiple Users	**9**	2	2	0	0	Mandatory
SRM2	Document Management	**6**	4	2	1	0	Mandatory
SRM3	Security	**6**	2	1	3	1	Desirable
SRM4	Role Management	3	3	2	**4**	1	Highly Des
SRM5	Reuse of past data	3	**7**	3	0	0	N/A
IS1	Ease of Setup	**6**	5	1	1	0	Highly Des
IS2	Installation Guide	4	**5**	1	3	0	Highly Des
IS3	Tutorial	**4**	**4**	3	2	0	Highly Des
IS4	Self-contained	0	**6**	**6**	0	1	Highly Des
E1	Free	0	**5**	3	1	4	Highly Des
E2	Maintained	6	**7**	0	0	0	Highly Des

Table 9. Comments about Protocol Development & Validation

ID	Feature	Theme	Participants	Experience
SRT1	Protocol Development	Helps track changes	2	L(1), H(1)
		Helps version control	4	L(1), M(1), H(2)
		Existing tools	4	L(3), H(1)
		Viability	1	L(1)
		For complex projects	2	H(2)
SRT2	Protocol Validation	Bad experiences	2	L(1), H(1)
		Over-complicating things	2	L(1), H(1)
		Useful checklist	2	L(1), H(1)

Table 10. Comments about Search & Selection

ID	Feature	Theme	Participants	Experience
SRT3	Search Process	Time Saving	3	L(2), H(1)
		Viability	5	L(2), M(1), H(2)
		Help Search Strategy	2	M(1), H(1)
SRT4	Study Selection	Time Saving	3	L(1), M(1) L(1)
		Managing Disagreements	3	L(2), H(1)
		Additional checking	2	H(2)

Search and study selection

Table 10 displays the main themes related to Search and Study selection. Although none of the participants felt that automated support for the search process was Not Needed (see Table 8 row SRT3), the opinions about its importance were divided among all the other importance levels. Three participants commented that such support would save them a lot of time. However, five participants were concerned that it would be difficult to develop trustworthy automated support (e.g. "It would be highly difficult to automate all that."). Two also mentioned the need for support to help develop the search strategy (e.g. "The bit where our time is most valuable is developing the search strategy in the first place.").

All participants felt that tool support for study selection was useful (see Table 8 row SRT4), with five participants regarding it as Mandatory and six as Highly Desirable. Three participants mentioned the potential for saving time. Three thought the facility would be useful for resolving disagreements and two mentioned the opportunity to check that things had not been missed. However, one participant felt that a lot of what the feature was targeting could be solved with a "quick conversation" between the members of the review team.

Quality Assessment
and Data Extraction

Table 11 shows the main themes related to Quality Assessment and Data Extraction. Concerning tool support for quality assessment (see Table 8 row SRT5), the majority of participants felt this would be another useful feature since "all these things otherwise require meetings and organisation". Participants also suggested specific features they would like to see:

- The ability to tailor quality criteria.
- The ability to link the quality assessment to data analysis.
- The ability to compare independent assessments and look for disagreements.

With regards to tool support for data extraction (see Table 8 row SRT6), all participants felt that tool support would be useful, with seven participants regarding it as Mandatory and five as Highly Desirable. In the context of an end-to-end tool, one participant said it would make extracted data ready to go "straight into the analysis". Four participants, however, were not sure how such

Table 11. Comments about quality assessment & Data Extraction

ID	Feature	Theme	Participants	Experience
SRT5	Quality Assessment	Viability	2	L(1), H(1)
		Managing Disagreements	1	H(1)
SRT6	Data Extraction	Viability	4	L(3), M(1)

Table 12. Comments about data analysis & Synthesis

ID	Feature	Theme	Participants	Experience
SRT7	Data Synthesis	Viability	2	L(1), H(1)
		Time Saving	3	L(2), H(1)
SRT8	Text Analysis	Viability	2	L(2)
		Time Saving	1	M(1)
		Managing consistency	1	H(1)
SRT9	Meta-analysis	Not always necessary	4	L(2), M(1), H(1)

a tool could work particularly when handling qualitative data.

Data analysis and synthesis

Table 12 shows the main themes related to Data Analysis and Synthesis. Concerning automated support for data synthesis (see Table 8 row SRT7), all participants felt this would be useful, with five suggesting such a feature should be Mandatory and seven suggesting it was Highly Desirable. Three participants mentioned potential time saving. One participant felt that "less experienced reviewers would find [this feature] particularly useful". However, two participants mentioned factors that might make such a feature difficult to implement (i.e. many different types of analysis and new analysis methods being ahead of tool support).

Overall support for a text analysis feature was muted (see Table 8 row SRT8); the modal value was Nice-to-have (five participants). Two participants mentioned difficulties implementing such a tool (i.e. missing things and false positives). However, one participant felt that text analysis would become "increasingly more important as the complexity of the literature increases", while another mentioned that the technology was now getting to the stage where such a feature was viable. In terms of possible benefits, one participant thought that it would save time, another that it could be used to check the consistency of reviewers extractions.

The participants felt that tool support for meta-analysis (see Table 8 row SRT9) was either Mandatory (four participants) or Highly Desirable (five), although four participants noted that not all SRs require meta-analysis. One participant thought it would be useful for novices as, "for a lot of people undertaking a SR for the first time, meta-analysis is their biggest fear".

Report writing and validation

Table 13 shows the main themes related to report writing and report validation. With a modal value of Desirable, most participants felt that tool support for writing the report was not very important (see Table 8 row SRT10). Three positive comments were that it would give reviewers a starting point. In contrast to this, four participants noted that there are many different formats required by journals, meaning that full support might be unrealistic. Two participants also mention other existing tools (i.e. RevMan for Cochrane reviews and Google Documents).

With regards to tool support for report validation (see Table 8 row SRT11), the modal value was Not Needed and the other responses were spread across all the other levels excluding the Mandatory level. Two participants mentioned that there were other existing tools (i.e. Word with track changes and PRISMA).

Table 13. Comments about report writing & validation

ID	Feature	Theme	Participants	Experience
SRT10	Report Writing	Time Saving	1	H(1)
		Viability	4	L(3), H(1)
		Starting point	3	L(2), H(1)
		Existing tools	2	L(1), H(1)
SRT11	Report Validation	Existing tools	2	L(1), M(1)

Table 14. Comments about SR process management

ID	Feature	Theme	Participants	Experience
SRM1	Multiple Users	Multiple-user process	5	L(1), M(2), H(3)
		For complex projects	3	L(2), H1
SRM2	Document Management	Document integration	3	M(2), H(1)
SRM3	Security	Already done	2	L(2)
		Proprietary data	5	L(1), M(1), H(3)
SRM4	Role Management	Over-complicating things	1	L(1)
		For complex projects	3	L(2), H(1)
		For overseeing	2	M(1), L(1)
SRM5	Re-use	For updates	2	L(1), H(1)
		Use previous work	3	L(1), M(1), H(1)

5.3.2. SR process management

SRLC tool features related to the management of the SR process are labelled SRM1 to SRM5 in Table 8.

Table 14 shows the major themes concerning SR process management. The majority of participants felt support for multiple users within a tool was really important with nine participants considering it Mandatory (see Table 8 row SRM1). Five participants noted that people do not write systematic reviews on their own, so such a facility is mandatory. Three participants mentioned it was appropriate for complex projects: one participant thought "It should do for large projects", another "If I was working with people internationally", and another mentioned the SRs are generally "team collaboration type projects".

Most participants felt that tool support for document management would be a useful feature (see Table 8row SRM2), with six participants regarding it as Mandatory and four as Highly Desirable. In particular, three participants mentioned the importance of being able to manage links between primary studies and one mentioned "Going from a reference manager to a study-based system".

Most participants felt the feature which supports security, should be included in a tool (see Table 8 row SRM3). Six participants regarded it as Mandatory and two as Highly Desirable. Five participants (including one novice) mentioned security was needed to address problems associated with confidential information and intellectual property rights. Two novice participants argued, however, that since SRs deal with published studies, security wouldn't be necessary. It is possible that systematic reviewers with more experience are more likely to have come across reviews where confidentiality was important.

The participants were divided as to the importance of tool support for role management (see Table 8 row SRM4). Although three participants regarded role management as Mandatory, the modal value for this feature was Nice-to-have which was the assessment made by four participants. Three participants felt it was important for complex projects (large teams). Two other participants thought that it would help to get an overview of the whole team, one of them pointing

Table 15. Comments about ease of use

ID	Feature	Theme	Participants	Experience
IS1	Ease of Setup	Depends on tool	2	H(1), H(1)
		Poor installation frustrates	2	M(1), H(1)
		Job for IT Staff	2	M(1), H(1)
IS2	Installation Guide	Job for IT Staff	1	L(1)
IS3	Tutorial	None	n/a	n/a
IS4	Self-Contained	Depends on tool	3	L(1), H(2)

out that it was particularly important for the first author. Another participant, pointed out that "it does not necessarily mean that you don't trust people to do a good job, it would just cut down the chances of a mistake". One novice researcher mentioned that it might over-complicate the process.

It is possible that systematic reviewers without software engineering experience would not appreciate it that in order to produce a software tool that supports independent quality assessment and data extraction of documents by two or more researchers, it identifies disagreements among their extractions and facilitates the production of a final mediated extraction, a certain kind of role management is essential.

All participants felt that tool support for re-using data from past SRs would be useful (see Table 8 row SRM5). Two participants mentioned it was important for updating existing reviews. Other participants mentioned possible uses of such a feature:

- When using primary studies that were used in a previous SR, the quality assessment could be reused.
- The references for primary studies used in previous SRs would be available.
- Using the search terms, you could automatically identify papers that were used in previous SRs.

5.3.3. Ease of use

Features related to the setup of a SRLC tool are labelled IS1 to IS4 in Table 8.

Most participants were in favour of tools that were easy to setup (see Table 8 row IS1), and included an installation guide (see Table 8 row IS2) and a tutorial (see Table 8 row IS3). They also felt having a self-contained tool[4] was either Highly Desirable (six participants) or Desirable (six participants) (see Table 8 row IS4).

Table 15 identifies the main discussion themes for ease of use features, identifying issues that were mentioned more than once. With respect to a simple setup accompanied by an installation guide, three participants mention IT staff were available to handle installation issues. Two participants felt that without a simple installation process, users would become frustrated with a tool. Two participants, however, felt that "if the tool is good enough", then, "some people are prepared to give [the difficult setup] a go". These features are discussed further in Section 5.6.

With respect to whether SR lifecycle tool should be self-contained, three of the participants, felt it was not a really important issue, since they would be quite satisfied to install other packages if the tool "does stuff that nothing else can do".

5.3.4. Economic features

Economic features are labelled E1 and E2 in Table 8. With regards to the cost of a tool, opinions differed (see Table 8 row E1). At the extremes, five participants thought free tools were Highly Desirable whereas four participants thought free tools were not necessary.

Table 16 identifies the main discussion themes for economic features. The discussion of the cost of tools centred around the concern that it was not possible to get good quality, trustworthy tools that provided all required features without

[4] I.e. a tool able to function, primarily, as a stand-alone application.

Table 16. Comments about economic features

ID	Feature	Theme	Participants	Experience
E1	Free	Good tools aren't free	9	L(4), M(1), H(4)
		Different licences for different users	3	L(1), M(1), H(1)
E2	Maintained	Methods evolve	4	L(2), M(1), H(1)
		Need Defect Management	2	L(1), H(1)

payment. Nine participants mentioned that they did not expect good tools to be free.

Three participants mentioned different licenses for different users would be a good idea, allowing free systems for students or for private use.

All participants felt post development maintenance of a tool (see Table 8 row E2) was either Mandatory (six participants) or Highly Desirable (seven participants). The discussion of this feature concerned the need for maintenance, with four participants pointing out that methods evolve and two mentioning that such large, complex systems would probably include defects that would need to be corrected.

Overall trends

Several themes were identified against more than two features:

- Viability (i.e. the concern that the feature would be difficult to automate) was identified against seven different features.
- Time saving (i.e. the potential for a feature to substantially decrease the SR workload) was identified against five features.
- Use other tools (i.e. the availability of other tools to implement the feature requirements) was identified against three features. The specific features were Protocol Development, Reporting and Report Validation.
- For complex projects (i.e. the feature was considered appropriate for projects with large or distributed teams) was identified against three features. The specific features were Protocol Development, Multiple Users and Role Management.

Table 17 shows the number of times participants mention the issues of Viability and Time Saving for each SR process tool feature[5]. This table suggests that participants were most concerned about the viability of support for the search process, data extraction and reporting. In addition, participants identified time saving as likely for search automation, selection and data synthesis processes more often than for other processes.

Table 17. Distribution of general comments against features

Feature	Viability	Time Saving
Protocol Development	1	0
Protocol Validation	0	0
Search Process	5	3
Study Selection	0	3
Quality Assessment	2	0
Data Extraction	4	0
Data Synthesis	2	3
Text Analysis	2	1
Meta-analysis	0	0
Reporting	4	1
Report Validation	0	0

Table 18 shows the distribution of comments concerning Viability and Time Saving against individual participants. It shows the number of times each participant made a comment about each issue. The table shows that concerns about viability of tool support are spread across all but one of the participants. On the other hand, although only one participant with a high level of experience mentioned time saving four times, four out of six participants who mentioned time saving had low levels of experience suggesting the time taken to complete an SR is of more importance to relative novices. This is consistent with the results shown in Table 2, where five out of six participants who mentioned that SRs

[5]Time Saving and Viability were not mentioned against any other feature groups.

were generally time consuming had low levels of experience.

Table 18. Distribution of general comments against participants

Participant	Experience	Viability	Time Saving
P01	M	1	1
P02	L	1	0
P03	L	5	1
P04	L	1	2
P05	M	1	0
P06	L	3	1
P07	L	2	1
P08	H	0	0
P09	H	2	0
P10	L	2	0
P11	L	2	0
P12	H	1	0
P13	H	1	4

5.3.5. Comparison of importance ratings

Table 8 presents the assessment of the importance of the features to SE researchers. No assessment for the importance of reusing results from previous SRs was provided, because the reuse of past project data is seldom performed in SE systematic reviews, so there was possibility of rating the importance of this feature.

A comparison of the assessment results and the study participants' assessments shows that for every feature, the majority of participants agreed that it was important. Thus, the set of all features that should be included in a SRLC tool is quite robust to differences between domains. As it was expected, there were differences in the evaluation of the importance of features among individual participants and among domains. However, there were also similarities.

For ten features, the modal response of participants to the importance of the feature was exactly the same as this assessment. In the case of three other features, there were two modal values for feature importance, and in both cases one of the modal values was the same as ours. In only three of the remaining features, did the modal value of the participants scores differ by more than one level from ours. The three features with substantial disagreement were:

1. Security, regarded as Desirable by the authors, had a modal value of Mandatory among the interview participants.
2. Meta-analysis, which we regarded as Nice-to-have, but which nine of the 13 interview participants rated as Mandatory or Highly desirable.
3. Role management, which was regarded as Highly Desirable, while the modal response of the participants was Nice-to-have. However, it should also be noted that six of the participants rated this feature as Mandatory or Highly Desirable.

These results confirm that the importance of various features is context dependent. For example, meta-analysis is rarely undertaken in SE research but is a normal part of health care research, so it is much less important to SE researchers than health care researchers. Nonetheless, although there are differences, it appears that the importance of features is surprisingly similar across the different domains. It should also be noted that none of the participants suggested any additional features which confirms that the SR methodology is not radically different in different domains.

5.4. The effect of experience on perceptions of feature importance

There has been considerable discussion in SE about the problems facing novice reviewers (see, for example, [12] and [11]). Furthermore, this issue was directly investigated by Hassler et al. [31]. Therefore the main interest was the investigation whether relative novices had different perceptions of the importance of tool features compared with more experienced reviewers.

Table 19 addresses exactly this issue. The column labelled *Total % Score* is the percentage of the maximum importance score obtained for a specific feature across all participants. The score was obtained by mapping the ordinal scale points for importance to numbers (i.e. Mandatory = 4, Highly Desirable = 3, Desirable = 2, Nice to have = 1 and Not Needed = 0). The

Table 19. Relationship between features scores and experience

ID	Feature	Total % Score	Low Exp	High Exp	Diff
SRM1	Multiple Users	88.46	79.17	96.43	17.26
SRT6	Data Extraction	86.54	79.17	92.86	13.69
E2	Maintained	86.54	75.00	96.43	21.43
SRT5	Quality Assessment	82.69	79.17	85.71	6.55
SRT7	Data Synthesis	82.69	70.83	92.86	22.02
SRT4	Study Selection	80.77	70.83	89.29	18.45
IS1	Ease of Setup	80.77	70.83	89.29	18.45
SRM2	Document Management	78.75	70.83	85.71	14.88
SRM5	Reuse of past data	75.00	66.67	82.14	15.48
SRT9	Meta-analysis	71.15	58.33	82.14	23.81
IS2	Installation Guide	69.12	58.33	78.57	20.24
IS3	Tutorial	69.12	58.33	78.57	20.24
SRM3	Security	67.31	50.00	82.14	32.14
SRT3	Search Process	65.38	79.17	53.57	-25.60
IS4	Self-contained	57.69	54.17	60.71	6.55
SRM4	Role Management	55.77	33.33	75.00	41.67
SRT1	Protocol Development	51.92	50.00	53.57	3.57
SRT10	Reporting	46.15	45.83	46.43	0.60
E1	Free	42.31	37.50	46.43	8.93
SRT2	Protocol Validation	34.62	37.50	32.14	-5.36
SRT8	Text Analysis	34.62	29.17	39.29	10.12
SRT11	Report Validation	34.62	37.50	32.14	-5.36

total percentage importance score for a feature was obtained as follows:

$$TotalScore_i = 100\frac{\Sigma_j Importance_{i,j}}{\Sigma_j(4)} \qquad (1)$$

where $TotalScore_i$ is the percentage of the maximum score for feature i, and the maximum score for a feature is $\Sigma_j(4)$, $j = 1, \ldots, 13$ is the number of participants and $Importance_{i,j}$ is the importance score that participant j gave to feature i. The table is ordered on this column.

The column labelled *Low Exp* reports the percentage score for the six participants who had performed between one and five SRs and the column labelled *High Exp* reports the percentage score for the seven participants who had completed more than five SRs. The column labelled *Diff* is the difference between the *High Exp* score and the *Low Exp* score.

Table 19 shows that, in general, participants with high levels of experience rated tool features higher than relative novices, since only three of the 22 features were scored higher by the relative novices than by the experienced participants.

It also seems that the relative importance of tools is quite similar for both groups, since the Pearson correlation between the scores for relative novices and experienced staff was 0.76. There are three features which exhibit extremely anomalous values:

1. Search Process support was scored much lower by experienced participants than by relative novices.
2. Role Management support was scored much higher by experienced participants than by relative novices.
3. Security support was also scored much higher by experienced participants than by relative novices but is not such an extreme anomaly.

Excluding these feature increases the correlation between the scores to 0.95.

5.5. The effect of SR type and domain

The authors hoped to assess whether the type of systematic review researchers performed influenced their perception of the importance of different framework features. For example, the authors expected researchers who primarily un-

Table 20. Experience and importance scores for analysis features

Experience	SR Type	Domain	Meta-analysis	Data Synthesis	Text Analysis
Low	Quant	HC	3	3	2
Low	Qual	HC	3	2	3
Low	Qual	HC	1	3	0
Low	Quant	HC	2	3	0
Low	Quant	SS	4	3	1
Low	Both	SS	1	3	1
High	Both	HC	4	4	0
High	Both	SS	4	4	3
High	Quant	HC	3	4	1
High	Both	SS	3	4	2
High	Both	SS	2	4	3
High	Qual	SS	3	3	1
High	Both	HC	4	3	1

dertook quantitative systematic reviews to emphasise the importance of meta-analysis tools and researchers who primarily undertook qualitative systematic reviews to emphasise the importance of more general data synthesis facilities and text analysis facilities. It was also expected that social science researchers would undertake qualitative systematic reviews and health care researchers would undertake primarily quantitative systematic reviews.

The expectations of the authors were not met. Table 20 shows the systematic review type, Domain type of participants and their importance scores for meta-analysis, Data Synthesis and Text Analysis. Four of the social science participants and two from health care reported performing both quantitative and qualitative systematic reviews. Of the remaining five health care researchers, three concentrated on quantitative systematic reviews and two on qualitative systematic reviews. Of the remaining two social sciences participants, one primarily undertook qualitative studies and the other primarily undertook quantitative studies. The impact of the domain and SR Type are summarized in Table 21. In the case of tool support for meta-analysis and data synthesis, Table 19 shows that more experienced participants tended to regard such a feature to be more important than the less experienced ones, however, Table 21 suggests that there is no domain effect.

With respect to SR type, Table 21 suggests that participants doing qualitative studies may regard support for meta-analysis and data synthesis as less important than other subjects. However, this result may be confounded with experience since only two of the seven subjects who concentrated on a single study type had high levels of experience whereas five of the six subjects who did both types of study had high levels of experience.

5.6. Revising the setup and installation features

During the previous validation of the SRLC tool framework [18] it was difficult to distinguish between the three features related to installing and using the SRLC tool and therewas an idea that would be better to integrate the three features into a single feature. The scores given by each participant to each of the three features is shown in Table 22. Across the three features, 10 of the 13 participants gave the same score for all three features. Those that gave different scores, scored the Installation guide and Tutorial lower than Ease of Set up. This result supports the view that only one high-level feature is needed to address the set up and installation.

However, participants' earlier comments relating to the difficulty of using EPPI-Reviewer and RevMan (see Table 7) suggest that usability is a significant issue to users. Therefore, a feature relating to provision of a Tutorial should be included. However, it might be preferable to

Table 21. The Impact of domain and SR type on scores for analysis features

Factor	Type	Participants	Meta-analysis	Data Synthesis	Text Analysis
Domain	HC	7	71.43	78.57	25.00
	SS	6	71.43	71.57	35.71
SR Type	Both	6	75.00	91.67	41.67
	Qual	3	58.33	66.67	33.33
	Quant	4	75.00	81.25	25.00

Table 22. Experience and importance scores for features related to installation and set up

Experience	Ease of set up	Installation Guide	Tutorial
Low	4	1	2
Low	3	3	2
Low	3	3	3
Low	3	3	3
Low	1	1	1
Low	3	3	3
High	4	1	1
High	4	4	4
High	4	4	4
High	4	4	4
High	2	2	2
High	3	3	3
High	4	4	4

generalise the feature and use the term *Ease of Use*, with a tutorial as one way of implementing such a feature.

6. Discussion

In this section the results of this cross-domain study is discussed from the viewpoint of the research goals.

6.1. The relevance of experiences from other domains

The results show that there are some differences between SE reviews and those in health care and social sciences. For example, health care and social science researchers may undertake systematic reviews commissioned by clients, whereas in SE these are normally researchers that undertake systematic reviews to further their own research goals.

There were other differences which the authors believe are likely to be due to the relative immaturity of systematic reviews in software engineering. For example, in Hassler et al.'s study [31] researchers with a high level of experience were defined as those who had performed three or more SRs, whereas in this study the highest experience levels of more than 15 SRs were categorized. In addition, reports from SE researchers summarized in [9] concentrated on technical processes which were emphasized in the first two versions of the SE systematic review guidelines. In contrast comments from the participants of this study identified issues related to review management not only issues related to technical processes. This is consistent with the results of Hassler et al.'s study [31] in which he noted that researchers with higher experience levels voted for features that aided tactical activities, whereas novices voted mainly for tools supporting operational tasks. As researchers in software engineering begin to perform more complicated systematic reviews, both in terms of SRs that involve many distributed researchers, as well as studies that involve large numbers of

candidate primary studies, possibly of different study types, it was expected that SE researchers would experience more problems associated with systematic review management.

Another difference was that there were two additional challenges mentioned by study participants that were not considered in the SE literature: version control and formatting references. Both of these issues seem important in a comprehensive SRLC tool, so need to be considered in any comprehensive evaluation framework for SRLC tools.

We also observed some differences in the ratings of importance of SLRC tool features, compared with our assessment of the importance of such features to SE researchers:

- Support for meta-analysis appeared to be more important to participants than it was assessed to be to SE researchers in this study. This was true even for two of the three participants who primarily undertook qualitative reviews. It appeared that study participants were well aware that meta-analysis tools are essential for some quantitative studies, even if they did not use such tools themselves.
- Support for security was more important in the health care and social science domains than it is in SE. In particular, more experienced participants were very concerned about restricting access to confidential information (only one of the five participants who mentioned this was a relative novice), whereas two relative novices felt that since they were dealing with existing published papers confidentiality was not an issue. In terms of SE researchers, it would certainly be the case that mapping studies were unlikely to have any confidentiality issues.
- There was a lack of strong support for textual analysis tools. Kitchenham and Brereton [9] reported that there were a substantial number of software engineering studies addressing textual analysis for systematic reviews and Marshall and Brereton [9] identified the number of tools to support textual analysis, so more enthusiasm was expected for such a feature. However, the modal response among the 13 participants was that such a feature was only "Nice-to-have". Nonetheless, participants were enthusiastic about other features that could be implemented using textual analysis such as Study Selection (modal response "Highly Desirable") and Data Synthesis (modal response "Highly Desirable") and one user of EPPI-Reviewer pointed out that EPPI-Reviewer used textual analysis to implement a feature that finds the most relevant studies. It was concluded that textual analysis may be necessary in order to implement SRLC tool features, but it may not be needed as a top level feature available directly to tool users.

Overall, it was concluded that there are common challenges among the different domains and the results of this study could be used to to evaluate and refine our evaluation framework. Furthermore, since the domains have similar challenges, it is in the interest of software engineering researchers to remain aware of innovations in the systematic review methodology to avoid the risks of both missing out on new methods or re-inventing the wheel.

6.2. Tools used to support systematic reviews

Participants identified 14 tools that they used while doing systematic reviews. The most commonly used tools were reference manager tools in particular RefWorks and EndNote. In addition, the participants mentioned two SRLC tools: RevMan and EPPI-Reviewer. However some of the tools were general purpose tools such as Microsoft Word and Excel, while others were statistical software tools or bespoke tools. The core set of ten tools that support systematic reviews including reference managers, SRLC tools and meta-analysis tools, together with tools identified in Marshall and Brereton's mapping study [21] and tools identified from other sources (i.e. [28, 34], and the Cochrane Collaboration website) were incorporated into an online tool called SRToolbox [29]. This set of tools has been substantially updated since this research was completed, and the most up-to-date categorized list can be found at the website systematicreview-

tools.com. This website is maintained by Marshall and has replaced the Cochrane Collaboration web pages on tools.

With respect to SRLC tools, EPPI-Reviewer was believed to be relevant to the needs of SE researchers, however, it is unclear to what extent it is tailored specifically to the needs of social scientists, and it is not free.

In the context of features required in SRLC tools, a common discussion point with our participants was whether it was even possible to automate some of the features. Participants of all experience levels feared that advanced tools might be untrustworthy, in particular that they would miss things or make classification errors or be incomplete. Thus, SRLC tool developers need to have a sound rationale for the algorithms they use to implement features, before their tools are likely to be widely accepted. Furthermore potential tool users in SE should appreciate the difficulty of implementing some of the features they might desire.

Another important issue was that most participants did not expect good quality tools to be free. Also the participants agreed that tools needed to be maintained because methods evolve and complex tools usually have residual errors that need to be corrected.

6.3. The impact of participant experience

Generally, more experienced participants rated features of support tools as more important than relatively inexperienced participants. It is likely that the more experienced participants had taken part in some large, complex systematic reviews and have, therefore, experienced the problems that such reviews can cause. Certainly, there is some evidence that more experienced participants undertook more varied SRs. Table 1 shows that five of the six relative novices undertook only one type of SR (either qualitative or quantitative) whereas only two of the seven more experienced researchers performed only one type of study.

The implication for SE researchers is that the need for SE tools in general, and SRLC tools in particular, should be expected to increase as SE researchers become more experienced with the SR process, and attempt larger and more complex systematic reviews. In particular, Table 2 and Table 8 indicate the importance of tools to support SR process management in addition to tools supporting specific SR tasks.

Throughout this study, the participants often mentioned that the importance of tool features depended on the size of the team and the complexity of the SR. Thus, requirements for SRLC tools should probably be elicited from researchers who have experienced the problems of large-scale SRs. In addition, the evaluations of such tools should ideally involve experienced researchers and large-scale SRs.

Also, since novice researchers usually undertake relatively small reviews in small teams, they might be best served by using a variety of tools, including Microsoft Excel and Word and a reference manager system, that they are already familiar with. It is unlikely that novices would benefit from extensive automation if the overheads, such as the required learning time needed to use a tool effectively, are significant.

6.4. Implications for the evaluation framework

One of the main aims of the study was to provide some independent assessment of the SRLC tool evaluation framework [18]. Kitchenham and Brereton had been deeply involved in the adoption of systematic reviews in SE. Originally, the promoted process was developed from the health care domain and the main focus was on on adapting the methodology to the SE domain. After developing the evaluation framework based on SE practice, it was thought that it would be extremely valuable to investigate whether there were more insights to be obtained from other domains.

The discussion about the features of an SRLC tool and the relative importance of such features confirmed that all of the features and the majority of the importance ratings were consistent with the views of the health care and social science researchers. In particular, none of the features was considered completely unnecessary and only

three features had importance ratings very different from the ratings obtained in this study.

However, some changes were made in the evaluation framework as a consequence of the study results:

1. Analysis of the three features related to the ease of installation and setup confirmed the view that it was better to have only one feature labelled *Ease of Setup*, where installation guides are a means by which the feature can be implemented. In addition, since several participants commented that RevMan and EPPI-Reviewer were difficult to use, it was recommended to replace the Tutorial feature by the *Ease of use* feature, with a tutorial as one means of assisting tool users to use the tool effectively. The feature set should be renamed as *Usability*.
2. The discussion about the importance of textual analysis convinced us that it was not really a self-standing feature, but represented a means of supporting various features such as *Data synthesis* and *Study selection*. The evaluation framework includes additional assessment criteria to assist evaluating how well each feature is implemented. Now the textual analysis is included as one of the additional criteria used to assess the support for these features.
3. Three challenges that were mentioned by participants but had not been discussed in the SE literature were identified. One of them was *negotiating with policy makers* which does not appear to be an issue of relevance to software engineering researchers, and indeed, may only be of relevance in the UK to health care and social science researchers. The other two issues were *version control* and *formatting references*. Both of these issues should be of concern to software engineering researchers. Version control was already mentioned in the evaluation framework as an associated assessment criteria for the Protocol Development but it should also be included in the associated evaluation criteria for Report Development. Formatting references should be included in the additional assessment criteria of the Search process.
4. Importance level was not assigned to the *Reuse of Past Project Data*. It was decided to adopt the rating of Highly Desirable which was the modal value of the participants' ratings. However, the users of this evaluation framework are expected to downgrade the importance level if they do not plan to keep their SR results up to date.

The changes have only a limited effect on the evaluation framework. For example, the SLuRp tool [20] would have scored 65% with the framework as it was used before this study. The tool score is the weighted sum of the score for each feature set: where the weight for the SR activity feature set is 4, the weight for the Process Management feature set is 3, the weight for the Usability feature set is 2, and the weight for the Economic feature set is 1:

$$ToolScore = \frac{\Sigma_{i=1,\ldots,4} FSW_i FSS_i}{\Sigma_{i=1,\ldots,4} FSW_i} \qquad (2)$$

where FSS_i is the score for feature set i, and FSW_i is the weight for feature set i.

The score for each feature set is the sum of the score for the extent to which each feature is supported (taking values 0, 0.5 and 1) multiplied by the score of the importance of each feature. This value is converted to the percentage of the maximum score for the feature set:

$$FSS_i = \frac{100\Sigma_{j=1,\ldots,k} FI_j FS_j}{\Sigma_{j=1,\ldots,k} FI_j} \qquad (3)$$

where FSS_i is score for feature set i, FI_j is the numerical importance for feature j in feature set i and FS_j is the extent to which the feature is supported in the tool being evaluated.

As a result of the changes introduced by this study the score for SLuRp decreased to 63% because:

– The feature Ease of Setup was scored as partly true for SLuRp and was given an implementation value of 0.5, since an installation guide was available.
– The feature Installation Guide was removed as a separate feature in the framework decreasing the number of features in the Usability feature set to four.

- The feature Ease of Use was introduced as a feature (to replace the Tutorial feature) with an importance of Highly Desirable. SLuRp scored the minimum value of zero for the feature since there was no tutorial, nor an online help facility, and the system is very complex.
- The feature Re-use of past data was included in the Process Management feature set, with an importance level of Highly Desirable. SLuRp maintains records of past SRs and their results, so it scored the maximum value of one for this feature.
- Text analysis on which SLuRp scored the maximum value of one was removed as a feature in the SR activity support feature set.

6.5. Comparison with other results

As reported in Section 3, Hassler and his colleagues undertook a series of studies investigating SR tool requirements. In contrast to the results reported in this study, their studies concentrated on the opinions and experiences of the SE community.

Carver et al. [14] investigated barriers to the SR process. Many of the issues they mentioned were discussed in Kitchenham and Brereton's systematic review [9]. However, they also provided a much more detailed discussion of the problems with current SE databases including the necessity to deal with duplicates, which was mentioned by one of the participants in this study. They also mentioned the issue of coordinating the reviewing and selection of papers and associated issues for team management and conflict resolution which were mentioned by the participants of this study.

The participants in Carver et al.'s study ranked the SR processes as most in need of tool support. They ranked Searching Databases as most important followed by Selecting papers and Extracting data. In contrast, this study rated Data Extraction as the most important SR task requiring support, followed by Quality Assessment and Data Synthesis. This difference may be caused by the concentration on mapping studies in SE. Carver et al.'s results suggested relatively little support for issues related to protocol development (i.e. Defining Research Question, Identifying Keywords, and Creating Search Strings), which is consistent with the relatively low importance given by our participants to automated support for protocol development.

It is quite difficult to make detailed comparisons between Hassler et al.'s study to identify barriers to the SR process [30] and [31] this one, because in each study, the terminology was based on the terminology used by the participants. In addition, when the participants of Hassler's studies voted, their votes were constrained. They were given a number of tokens (i.e. votes) and these tokens were shared across all the features being voted on and participants could give multiple tokens to specific features. This process meant that participants were prioritising across all the possible tools. In this study the participants were not asked to make any trade-off when they assessed the importance of individual tool features.

Hassler et al. [30] identified barriers faced by systematic reviewers related to the SR process, primary studies, the practitioner community and tooling. The comparison of the discussion points in Hassler's study with the results of this study is shown in Table 23. Hassler identified the difficulty of meta-analysis as a problem, but looking at his comments it appears that data synthesis rather than statistical meta-analysis was a problem, which is consistent with these results. Barriers related to the practitioner community were not mentioned as a problem in health care or social science where the practitioner community may be more accustomed to the need for systematic reviews. Hassler's participants identified barriers related to tooling in terms of needing improved search and retrieval facilities including addressing the problem of rewriting search engine strings which was mentioned as a challenge by one participant. However, support for the search process did not feature as one of the most important features in Table 19. It is noticeable that support for the search process is considered much more important by relative novices than by experienced researchers, so the difference between our result and Hassler's results may reflect the fact that there are few researchers in SE that have completed more than 5 systematic reviews. Hassler

Table 23. Comparison with barriers discussed in Hassler's study [30]

Category	Issue	This study
SR Process	SR protocol is sequential, but process iterative	Not mentioned
	Meta-analysis is difficult	Need support for data synthesis
	Lack of methods for result interpretation	Not mentioned
Primary Studies	Title and abstracts misleading	Not mentioned
	Terminology not standardized	Mentioned by four participants
Practitioner Community	Difficulty relating to industry needs	Not mentioned
	Difficulty justifying structured process	Not mentioned
Tooling	Electronic databases are inadequate for search and retrieval	Problem with string translation mentioned once
	Need data extraction and management tools	Strong support in this study

discussed the need for support for data extraction and management. Our results strongly align with this result, since support for Data Extraction was the second most highly ranked feature by our participants and features related to Management issues, such as Multiple Users, Document Management, Role Management which were all highly ranked particularly by more experienced researchers.

Hassler et al. undertook a second community workshop to identify SR tool needs [31]. In this workshop they had 16 participants of which 10 were categorized as "experts" because they had completed at least three SRs. They compared their results with those of Marshall et al. [19] In this study this analysis was extended to consider the impact of participant experience as shown in Table 24. This table is ordered on the total score for the features in this study. The order of the total score for equivalent features in Hassler's study is shown in parenthesis after the name of the feature. The experience scores for high and low experience participants were included, however, it is important to note that high experience was equated with completing more than five SRs so the comparisons are not exact. One change was introduced to Hassler et al.'s table, that is the Textual analysis feature was equated to Hassler's Automated Analysis rather than to Statistical Analysis.

The most obvious area of agreement between the study results is that, given that Multiple Users and Collaboration are equivalent, they correspond to the most important feature in this study and the second most important in Hassler's study, with importance rated more highly by more experienced researchers.

However, there are major differences between the ranking of tool features. The correlation between the total scores for this study and for Hassler et al.'s study is 0.44. Furthermore, the correlation between the scores for participants with low experience was 0.24, and between scores for high experience participants was 0.25. In addition, the correlation between the high and low experience participants' votes in Hassler's study was only 0.45.

Differences between Hassler's results and the ones obtained in this study could be due to the specific participants but it could also be caused by domain differences, experience differences or differences in the type of SRs in the SE domain. It is suspected that a major issue is the difference resulting from the prevalence of mapping studies in SE. Mapping studies are often confused with SRs in the SE community. However, they are often published in conferences and journals implying that mapping studies are of value to the SE community. This is not the case in health care or social sciences. Concentrating on mapping studies can lead to SE researchers being more interested in the search and selection processes than researchers in other domains and less concerned about data extraction and quality assessment. Also a mapping study analysis is often concerned with the similarities between large

Table 24. Comparison of features scores and experience for this study and Hassler et al.'s study [31]

Our study				Hassler et al.			
Feature	Total	Low Exp	High Exp	Feature	Total	Low Exp	High Exp
Multiple Users	88.46	79.17	96.43	Collaboration (2)	10.7	4.9	13.3
Data Extraction	86.54	79.17	92.86	Coding (= 8)	3.8	4.9	3.3
Quality Assessment	82.69	79.17	85.71	Quality Assessment (= 5)	5.3	2.4	6.7
Data Synthesis	82.69	70.83	92.86	Automated analysis (= 5)	5.3	9.8	3.3
Study Selection	80.77	70.83	89.29	Study Selection (3)	6.9	9.8	5.6
Document Management	78.75	70.83	85.71	Study storage (= 8)	3.8	2.4	4.4
Reuse of past data	75.00	66.67	82.14	Data Maintenance (4)	6.1	7.3	5.6
Meta-analysis	71.15	58.33	82.14	Statistical analysis (11)	2.3	4.9	1.1
Search Process	65.38	79.17	53.57	Integrated search (1)	11.5	9.8	12.2
Protocol Development	51.92	50.00	53.57	Development & validation (= 12)	0.8	0.0	1.1
Reporting	46.15	45.83	46.43	NA			
Protocol Validation	34.62	37.50	32.14	Development & validation (12)	0.8	0.0	1.1
Text Analysis	34.62	29.17	39.29	Automated analysis (= 5)	5.3	9.8	3.3
Report Validation	34.62	37.50	32.14	Report Validation (10)	3.1	2.4	4.4

numbers of studies which is helped by visual analysis and textual analysis techniques. Thus the relevance of results from other domains may depend on the extent to which systematic review approaches in SE continue to be dominated by mapping studies.

Some differences may be caused by the relatively low levels of experience among SE researchers. The high and low experience participants in Hassler's study are probably closer to the low experience participants in our study. So the differences between high and low studies in Hassler's study are more likely to be chance effects than those in this study.

6.6. Limitations

A major limitation of this cross-domain study is that the use of systematic reviews was discussed, however, mapping studies (or scoping studies as they are often referred to in other domains) were not explicitly discussed. Although the participants did not raise the issue of such studies themselves, it is possible that the assessment of the importance of some SRLC tool features might have changed if we had asked them to consider the implications of the features for scoping studies. A particular issue for software engineering SRLC tools is that textual analysis may well play a more important role in managing the study selection and data extraction for mapping studies than it does for systematic reviews. However, we would still expect textual analysis to be used to implement various features rather than being a tool feature in its own right.

Another important limitation is that there were relatively few participants. Nonetheless, the coverage of the three characteristics thought to have some influence on participants' experience was good: domain, type of SRs they undertake, and their level of experience. This means that the group of participants was heterogeneous, which is often considered the best approach to obtain a theoretical sample for a qualitative study.

All of the study participants were UK-based, so this might introduce some cultural bias into the study. However, all versions of the SE systematic review guidelines were based primarily on UK standards and they were widely adopted among software engineers from many different countries. Thus, our SR practices in software engineering may already have a built-in UK cultural bias.

Yet another limitation of this cross-domain study are those related to the method of semi-structured interviews and the experience of

the interviewer. Since this study was part of Marshall's PhD research, he performed all the reviews himself. However, in general, interview-based studies might be improved by the use of observer triangulation. In addition, semi-structured interviews depend strongly on the communication skills of the interviewer [35]. Marshall attempted to address this issue by undertaking a pilot study. Other risks are associated with the participants' impression of the interviewer. Research suggests that people respond differently depending on how they perceive the interviewer (*the interviewer effect* [36]). 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 [36]. In addition, participants' responses can be influenced by what they think the situation requires [37]. Marshall did all the interviews and made every effort to put the participants at ease and to explain the purpose of the interview. In addition, the fact that he was reasonably knowledgeable about systematic reviews and systematic review tools was found useful in overcoming potential problems due to his relatively junior level. Risks associated to missing relevant questions as the participants lead the flow of the interview were mitigated by using a list of questions and key themes to check the progress of the interview.

7. Conclusions

The results of our cross-domain study suggest that, in the context of systematic reviews, experiences of researchers in other disciplines can be valuable for SE researchers. The implications of this are:

- Standalone tools used by systematic reviewers in other domains may be of value to systematic reviewers in SE. We recommend SE researchers, particularly those supervising junior researchers, to periodically consult the SR Toolbox to keep track of available tools.
- SE researchers producing tools for systematic reviews should also be aware of the currently available tools and their features. In particular, in the context of SRLC tools, the features available in the EPPI-Reviewer tool might be worth studying.
- SE researchers can benefit from keeping abreast of systematic review developments in other disciplines. This is important to avoid a methodological drift. Researchers should not want general scientific methods to start to diverge across different domains. Nonetheless, there are some differences between domains that can impact the adoption of standards or tools, such as the importance of mapping studies, which makes it useful for SE researchers to continue to study SR methodology.

In terms of the impact of the results reported in this paper, we made several changes to our framework for evaluating SRLC tools. The changes were easy to implement and overall it appeared that the framework was quite robust across different domains [38].

We intend to continue refining the evaluation framework's feature set and evaluation criteria to accommodate the selection and assessment of novel tools developed to support systematic reviews. For example, a case study is currently under way to compare and evaluate a selection of tools that support network meta-analysis which uses an expanded version of the evaluation framework. Further refinements to the framework will also be reflected as part of the ongoing development of the Systematic Review Toolbox to classify tools.

Acknowledgements

We would like to thank the participants in the study for sharing their experiences with us and the reviewers for their helpful comments on our manuscript.

References

[1] C. Mulrow, "Rationale for systematic reviews," *British Medical Journal*, Vol. 309, No. 6954, 1994, p. 597.

[2] D. Cook, C. Mulrow, and R. Hayes, "Systematic reviews: synthesis of best evidence for clinical

decisions," *Annals of Internal Medicine*, Vol. 126, No. 5, 1997, pp. 376–380.

[3] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," Keele University and Durham University, Joint Report, 2007.

[4] D.S.W. Rosenberg, J. Gray, R. Hayes, and W. Richardson, "Evidence-based medicine: what it is and what it isn't," *British Medical Journal*, Vol. 312, No. 7023, 1996, p. 71.

[5] J. Higgins, *Cochrane Handbook for Systematic Reviews of Interventions*. Wiley-Blackwell, 2008.

[6] J.P. Ioannidis, "The mass production of redundant, misleading and conflicted systematic reviews and meta-analysis," *The Milbank Quarterly*, Vol. 94, No. 3, 2016, pp. 485–514.

[7] D.S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, Vol. 53, No. 5, 2011, pp. 440–455.

[8] F.Q. da Silva, A.L. Santos, S. Soares, A.C.C. França, C.V. Monteiro, and F.F. Maciel, "Six years of systematic literature reviews in software engineering: An updated tertiary study," *Information and Software Technology*, Vol. 53, No. 9, 2011, pp. 899–913.

[9] B. Kitchenham and P. Brereton, "A systematic review of systematic review process research in software engineering," *Information and Software Technology*, Vol. 55, No. 12, 2013, pp. 2049–2075.

[10] P. Brereton, B.A. Kitchenham, D. Budgen, M. Turner, and M. Khalil, "Lessons from applying the systematic literature review process within the software engineering domain," *Journal of Systems and Software*, Vol. 80, No. 4, 2007, pp. 571–583.

[11] M. Babar and H. Zhang, "Systematic literature reviews in software engineering: preliminary results from interview with researchers," 2014, pp. 346–355.

[12] M. Riaz, M. Sulayman, N. Salled, and E. Mendes, "Experiences conducting systematic reviews from novices' perspective," in *Proceedings of the 2010 International Conference on Evaluation and Assessment in Software Engineering*, 2010.

[13] S. Imitiaz, M. Bano, N. Ikram, and M. Niazi, "A tertiary study: Experiences of conducting systematic literature reviews in software engineering," in *In Proceedings of the 2013 International Conference on Evaluation and Assessment in Software Engineering*, 2013, pp. 177–182.

[14] J. Carver, E. Hassler, E. Hernandes, and N. Kraft, "Identifying barriers to the systematic literature review process," in *Proceedings of the 13th International Symposium on Empirical Software Engineering and Measurement*, 2013.

[15] M. Staples and M. Niazi, "Experience using systematic review guidelines," *Journal of Systems and Software*, Vol. 80, No. 9, 2007, pp. 1425–1437.

[16] H. Ramampiaro, D. Cruzes, R. Conradi, and M. Mendona, "Supporting evidence-based software engineering with collaborative information retreival," in *Proceedings of the 2010 International Conference on Collective Computing: Networking Applications and WorkSharing*, 2010, pp. 1–5.

[17] B.A. Kitchenham, D. Budgen, and P. Brereton, *Evidence-Based Software Engineering and Systematic Reviews*. CRC Press, 2015.

[18] C. Marshall, O.P. Brereton, and B.A. Kitchenham, "Tools to support systematic literature reviews in software engineering: A feature analysis," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE'14)*. ACM Press, 2014, pp. 13:1–13:10.

[19] C. Marshall, O.P. Brereton, and B.A. Kitchenham, "Tools to support systematic literature reviews in software engineering: A cross-domain survey using structured interviews," in *Proceedings of the 19th International Conference on Evaluation and Assessment in Software Engineering (EASE'15)*. ACM Press, 2015, pp. 26–31.

[20] D. Bowes, T. Hall, and S. Beecham, "SLuRp: A tools to help large complex systematic literature reviews," in *Proceedings of the 2012 International Workshop on Evidential Assessment of Software Technologies*, 2012, pp. 33–36.

[21] C. Marshall and P. Brereton, "Tools to support systematic literature reviews in software engineering: A mapping study," in *Proceedings of ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*. IEEE Computer Society Press, 2013, pp. 296–299.

[22] J. Molleri and F. Benitti, "SEAR: A web-based automated tool to support the systematic literature review process," in *Proceedings of the 2015 International Conference on Evaluation and Assessment*, 2015, pp. 24–33.

[23] B. Kitchenham, "Evaluating methods and tool Part 1: The evaluation context and methods," *ACM SIGSOFT Notes*, Vol. 21, No. 1, 1996, pp. 11–14.

[24] B. Kitchenham, T. Dybå, and M. Jørgensen, "Evidence-based software engineering," in *Pro-*

ceedings of ICSE 2004. IEEE Computer Society Press, 2004, pp. 273–281.

[25] B. Kitchenham, "Procedures for undertaking systematic reviews," Keele and Durham Universities, Joint Technical Report, 2004.

[26] M. Petticrew and H. Roberts, *Systematic Reviews in the Social Sciences: A Practical Guide*. Blackwell Publishing, 2006.

[27] J.A.S. Torres, D.S. Cruzes, and L. Salvador, "Automatic results identification in software engineering papers. Is it possible?" in *Proceedings of the 12th International Conference on Computer Science and Its Applications*, 2012.

[28] K.R. Felizardo, S. MacDonell, E. Mendes, and J. Maldonado, "A systematic mapping on the use of visual data mining to support the conduct of systematic literature reviews," *Journal of Systems and Software*, Vol. 7, No. 2, 2012, pp. 450–461.

[29] C. Marshall and O.P. Brereton, "Systematic review toolbox: a catalogue of tools to support systematic review," in *Proceedings of 19th International Conference on Evaluation and Assessment in Software Engineering (EASE'15)*. ACM Press, 2015, pp. 26–31.

[30] E. Hassler, J.C. Carver, N.A. Kraft, and D. Hale, "Outcomes of a community workshop to identify and rank barriers to the systematic literatire review process," in *Proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering*, 2014.

[31] E. Hassler, J.C. Carver, D. Hale, and A. Al-Zubidyb, "Identification of SLR tool needs–results of a community workshop," *Information and Software Technology*, Vol. 70, 2016, pp. 122–129.

[32] D. Remenyi, *Grounded Theory: A reader for Researchers, Student, Faculty and Others*, 2nd ed. Academic Conferences and Publishing International Limited, 2014.

[33] M.B. Miles, A.M. Huberman, and J. Saldāna, *Qualitative Data Analysis: A Methods Sourcebook*, 3rd ed. Sage Publications Inc., 2014.

[34] G. Tsafnat, P. Glasziou, M. Choong, A. Dunn, F. Galgani, and E. Coiera, "Systematic review automation technologies," *Systematic Reviews*, Vol. 3, No. 1, 2014, p. 74.

[35] P. Clough and C. Nutbrown, *A student's guide to methodology*, 3rd ed. SAGE, 2012.

[36] M. Denscombe, *The good research guide: for small-scale social research*. Open University Press, 2014.

[37] R. Gomm, *Social Research methodology*. Palgrave MacMillan, 2004.

[38] C. Marshall, "Tool support for systematic reviews in software engineering," Ph.D. dissertation, School of Computer Science and Mathematics, Keele University, 2016.

Appendix A. Interview guide

The interview guide was intended to help structure the interview and ensure that all relevant points were covered. Since these interviews were semi-structured, it might be the case that not all questions were required. Similarly, supplementary questions, not recorded in this guide, could be asked, depending on the individual circumstances of each interview. Questions have been classified into four groups; namely, Group 1: Subject Context, Group 2: Personal Experience with Systematic Reviews, Group 3: Experience with Tools to Support Systematic Reviews and Group 4: Features for a Tool to Support Systematic Reviews.

A.1. Introduction

Welcome the participant and ensure they are suitably comfortable, etc. Explain the purpose of the interview again so as to gather information about tools to support systematic reviews.

A.2. Group 1: G1 subject context

Questions in Group 1 will be asked about the participants' discipline. In particular, we are interested in discovering how SRs are used within the domain, the infrastructure provided when undertaking a SR and any tools that are available to support the process. Four questions will be asked.

G1-Q01. Could you tell me about your discipline?

G1-Q02. How do systematic reviews play a role within your discipline?

G1-Q03. What infrastructure does your discipline provide to support reviewers when performing an SR? (e.g. guidelines)

G1-Q04. What tools to support SRs are available within your discipline?

A.3. Group 2: G2 personal experience with systematic reviews

Questions in Group 2 will be asked about the participants' personal experience when performing an SR. In particular, we are interested to learn the extent of their experience, their thoughts on the usefulness of SRs, what they believe to be the main challenges and which aspects they feel are most in need of support.

G2-Q01. How many SRs have you performed?

G2-Q02. Do you find SRs useful?

G2-Q03. What, in your opinion, are the main challenges when undertaking a SR?

G2-Q04. In your experience, what are the key aspects of the SR process that you feel are most in need of automated tool support?

A.4. Group 3: G3 experience with tools to support systematic reviews

The questions asked in Group 3 will depend on whether or not the participant has experience using a tool to support them whilst undertaking an SR. If the experience exists, the participant will be asked about their experience using the tool(s). If the participant has not used a tool before, they will be asked why they haven't and whether they might consider using one in the future. In addition, question G3-Q09 initiates the snowballing sampling technique.

G3-Q01. Generally, do you feel the SR process could benefit from automated support?

G3-Q02. Have you used a tool (or multiple tools) to support yourself whilst undertaking a SR?

If the participant has experience using a tool, ask questions G3-Q003 to G3-Q06. If they have no experience using a tool, advance to question G3-Q07.

G3-Q03. What is the tool called? (This might have already been identified by question G1-Q04.)

G3-Q04. In your opinion, what were the main strengths of the tool?

G3-Q05. What were its key weaknesses?

G3-Q05. Overall, did you feel that using the tool was useful? (i.e. did you feel sufficiently supported?)

G3-Q06. Would you use the tool again?

G3-Q07. Is there a particular reason why you haven't used one? (e.g. don't know enough

about them, don't feel they are necessary, etc.)

G3-Q08. Would you consider using one in the future?

G3-Q09. Do you know someone who has used one? (Snowball sampling.)

A.5. Group 4: G4 features for a tool to support systematic reviews

Questions in Group 4 involve a data collection exercise. The interviewer will explain that a set of features for a tool to support the overall SR process has been developed. In their opinion, and in the context of the SR process within their discipline, the participant will be asked to determine whether each feature is considered "Mandatory (M)", "Highly-desirable (HD)", "Desirable (D)" or "Nice-to-have (N)". Alternatively, the participant can decide that a feature is "Not necessary (NN)". The interviewer will record the ratings made by each participant using a form with a row for each feature and a column for each rating level.

A.5.1. Feature Set 1 (F1): economic G4-F1

Questions relating to this feature set concern economic factors relating to the initial cost of the tool and the subsequent support for maintaining (or upgrading) the tool. Three questions will be asked.

G4-F1-Q01. How important is it that a tool should not require financial payment to be used?

G4-F1-Q02. How important is a well and freely maintained tool?

G4-F1-Q03. Are there any features you can think of that you might add to this feature set?

A.5.2. Feature Set 2 (F2): ease of introduction and setup G4-F2

Questions relating to this feature set focus on the level of difficulty inherent in setting up and using the tool for the first time. Five questions will be asked.

G4-F2-Q01. How important is a simple installation and setup procedure?

G4-F2-Q02. How important is the presence of an installation guide?

G4-F2-Q03. How important is the presence of a tutorial?

G4-F2-Q04. How important is it that the tool is as self-contained as possible? (i.e. able to function as a stand-alone application with minimal requirements from other external technologies.)

G4-F2-Q05. Are there any features you can think of that you might add to this feature set?

A.5.3. Feature Set 3 (F3): SR activity support G4-F3

Questions relating to this feature set relate to how well the tool supports each of the three main phases of an SR and the steps (or activities) within these phases. Here 12 questions will be asked. G4-F3-Q01 and G4-F3-Q02 concern features that support the planning phase of a SR. G4-F3-Q03 to G4-F3-Q09 relate to features supporting the conduct phase. G3-F3-Q10 and G3-F3-Q11 concern features that support the report phase.

G4-F3-Q01. How important is a feature that supports the development of a review protocol? (e.g. the tool provides support for collaboration using a template and control of versions to keep track of any changes to the protocol during its development.)

G4-F3-Q02. How important is a feature that supports protocol validation? (e.g. enabling evaluation checklists to be distributed to and completed by members of the review team.)

G4-F3-Q03. How important is a feature that provides support for the search process? (e.g. performing an automated search from within the tool which identifies duplicate papers and handles them accordingly.)

G4-F3-Q04. How important is a feature that provides support for study selection and validation? (e.g. the tool provides support for a multi-stage selection process, for multiple users to apply the inclusion/exclusion crite-

ria independently and a facility to resolve disagreements.)

G4-F3-Q05. How important is a feature that provides support for quality assessment and validation? (e.g. the tool enables the use of a suitable quality assessment criteria, allows multiple users to perform the scoring independently and provides a facility to resolve conflicts.)

G4-F3-Q06. How important is a feature that provides support for data extraction? (e.g. the tool provides support for the extraction and storage of qualitative data using classification and mapping techniques and, in addition, the extraction of quantitative data, which manages the specific numerical data reported in a study, should also be supported.)

G4-F3-Q07. How important is a feature that provides support for data synthesis? (e.g. the tool provides automated analysis on extraction data such as table/chart generation.)

G4-F3-Q08. How important is a feature that provides text analysis?

G4-F3-Q09. How important is a feature that provides meta-analysis?

G4-F3-Q10. How important is a feature that supports the report phase of a SR? (e.g. the tool provides a template to assist the report write-up.)

G4-F3-Q11. How important is a feature that supports report validation? (e.g. automated evaluation checklists similar to the example given for protocol validation).

G4-F3-Q12. Are there any features you can think of that you might add to this feature set?

A.5.4. Feature Set 4: (F4) process management G4-F4

Questions relating to this feature set relate to the management of an SR. Six questions will be asked.

G4-F4-Q01. How important is allowing multiple users to work on a single review?

G4-F4-Q02. How important are document management facilities? (e.g. in particular, managing large collections of papers, studies and the relationships between them.)

G4-F4-Q03. How important are security features? (e.g. log-in or a similar system.)

G4-F4-Q04. How important is the feature that provides support for role management? (e.g. state which users will perform certain activities, such as study selection, quality assessment, data extraction etc., and allocate papers accordingly.)

G4-F4-Q05. Is it important that the tool supports multiple projects? (i.e. the user can perform multiple SR projects using the tool.)

G4-F4-Q06. Are there any features you can think of that you might add to this feature set?

Appendix B. Interview Preparation Form

Each participant received the following information, sent on Keele University headed paper:

Study Title Tool Support for Systematic Reviews in Software Engineering

Aims of the Research The aim of this interview is to gather information about the availability, use, potential and effectiveness of automated tools which provide support for systematic reviews.

How long will the interview take? The interview should take no more than one hour to complete.

What will I be asked about? The interview will focus on discussing your thoughts and experience using tools to support the conduct of a systematic review. However, we are also interested in learning about the systematic review process particularly within your discipline. Questions will be asked in the following topics: The role of systematic reviews within your discipline. Known tools that are used to support the conduct of systematic reviews within your domain. Your personal experience undertaking systematic reviews (with/without the help of tools.)

How will information about me be used? The data collected will contribute towards the development of a refined framework for an overall tool to support SRs.

Who will have access to the information about me? The only people who will have access to the data collected are the members of the research team conducting this study. This include Christopher Marshall (PhD Researcher), Prof Pearl Brereton (Lead Supervisor) and Prof Barbara Kitchenham (Second Supervisor). All data will be made anonymous during the analysis process for future reports and research projects. Notes taken during the interview process will be stored on a password protected computer. Audio recordings (providing you have agreed for the interview to be recorded) will be stored in a locked filing cabinet.

Who is funding the research? This research is partly supported by Keele University's Environmental, Physical Sciences and Applied Mathematics (EPSAM) Research Institute.

Appendix C. Coding participants comments about the lifecycle tool features

The mechanism used for coding the participants comments about specific features was to tabulate the comments each participant made about the feature. Then participant comments that addressed a general issue were highlighted, including comments:

- Identified benefits that the feature would deliver (Inc1).
- Identified possible problems or limitations associated with the feature (Inc2).

but excluding comments that:

- Restated or emphasized the participant's rating of the importance of the feature (Exc1).
- Discussed how the feature would work (Exc2).
- Restated some comment about the feature that had already been coded for that participant (Exc3).

The highlighted comments were read and the topics that addressed the same issue were identified and given a short description. The 22 features were coded one feature at a time. However, the use of codes was checked, so that if any similar comments occurred in subsequent features, the same terms were used. After the initial coding of features was completed, we reviewed single comments in each feature to investigate whether such comments occurred for different features.

The coding process was performed by Kitchenham using the comments tabulated by Marshall and then validated by Brereton.

For example, for the comment for the Search Process were as follows:

- P01
  - No comments.
- P02
  - That would be absolutely fantastic. (Comment ignored Exc1 – restated participants' rating of feature.)
- P03
  - That would *save a lot of time.* (Comment Inc1 coded as *Time Saving.*)
  - As long as the *process is done thoroughly and you're not missing anything.* (Comment coded Inc1 as *Viability* defined as 'will the feature work'?)
- P04
  - That would be brilliant. (Comment ignored Exc1.)
  - That would be *time saving.* (Comment coded as *Time Saving.*)
  - I'm not going to say anything is Mandatory I think, because I do them [SRs] without [the features]. (Comment ignored Exc1.)
- P05
  - *I can see there might be problems with that.* (Comment Inc2 coded as *Viability.*)
  - What might be good instead would be to help build this search strategy. (Comment Inc1 coded as *Help Search Strategy.*)
- P06
  - I mean it sounds highly desirable, but it sounds like quite a task. (Comment ignored Exc1.)
  - I think that as a reviewer, you'd probably want to see how they'd *actually confirmed that [that the feature worked].* (Comment Inc2 coded as *Viability.*)
  - I think if that was shown to be highly reliable it would be highly desirable. (Comment ignored Exc3 – restated previously coded comment.)
  - These search engines are updated regularly, These search engines are updated

regularly, constantly update it [the feature]. (Comment ignored Exc3.)
  - It's a big ask. (Comment ignored Exc3.)
- P07
  - That's clearly mandatory in my book. That would be amazing. (Comment ignored Exc1.)
- P08
  - I think it could be useful to give an idea of the number of hits from each database. (Comment ignored Exc2 – discussing how the feature would work.)
  - It would help for a pilot search. (Comment ignored Exc2.)
  - I wouldn't want it to replace searching each individual database. (Comment ignored Exc2.)
  - I'm a bit against one search across all of the databases, because you are not actually searching the databases properly; you are not getting the best out of the databases. (Comment ignored Exc2.)
  - You would use *different strategies for different databases for good reasons.* (Comment Inc2 coded as *Viability.*)
- P09
  - *Well if it did it reliably.* (Comment Inc2 coded as *Viability.*)
  - The problem is you've got different controlled vocabularies in different databases. (Comment ignored Exc3.)
  - It would be highly difficult to automate all that. (Comment ignored Exc3.)
  - I think there are too many things in the way at the moment to be able to implement it. (Comment ignored Exc3.)
- P10
  - No comments.
- P11
  - No comments.
- P12
  - I would say highly desirable *but I don't trust you'd do it. I think there would be stuff missing.* (Comment Inc2 coded as *Viability.*)
- P13
  - Particularly about translating the search strategy. (Code ignored Exc1.)
  - I'd say that's highly desirable because *it's the thing that is time consuming.* (Comment Inc2 coded as *Time Saving.*)
  - The bit where our time is valuable is *most valuable is developing the search strategy in the first place. That sort of translating bit is very time consuming but it does not actually have to use that much expertise really.* (Comment Inc1 coded as *Help Search Strategy.*)