ADVANCING EVIDENCE-BASED PRACTICE IN SYSTEM DEVELOPMENT:
PROVIDING JURIED KNOWLEDGE TO SOFTWARE PROFESSIONALS.

by

EDGAR E. HASSLER III
DAVID P. HALE, COMMITTEE CHAIR
JEFFREY C. CARVER
JOANNE E. HALE
RANDY K. SMITH
UZMA RAJA

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
in the Department of Information Systems, Statistics, and Management Science
in the Manderson Graduate School
in the Culverhouse College of Commerce & Business Administration
in the Graduate School of
The University of Alabama

TUSCALOOSA, ALABAMA

2014
ABSTRACT

The concept of utilizing information derived from carefully crafted scientific research to optimize the efficiency and/or effectiveness of a practice is by no means a new idea. For over two thousand years physicians, scientists, and business professionals have relied on evidence to improve decision making. The advances made over last 50 years with regard to Information Systems (IS), and the proliferation of technology, set the stage for a new paradigm in the use of information in practice commonly referred to as Evidence-Based Practice (EBP). Originating in the field of medicine, the EBP paradigm has been adopted in many of the healthcare domains and spread to other domains such as education, management, and computer science.

In the spirit of Fredrick Taylor’s *The Principles of Scientific Management* (1914), the collection of essays presented in this work endeavor to advance the use of empirically based, juried evidence for decision making in a business context. The specific context selected is the domain of SE – a key component in a technology laden world. Within the SE domain, the essays address three objectives in the advancement of the EBSE paradigm by:

1. Mapping the research completed to date regarding the implementation of EBP in the SE domain to identify gaps and opportunities in the research.
2. Identifying the barriers deemed most important by the members of the SE research community who conduct systematic literature reviews in support of EBSE.
3. Developing the use of algorithmic techniques as a discriminant function in the selection process of the systematic review methodologies.
Together, the collection of essays represent a line of inquiring within a broader research stream concerning the implementation of EBP – a modern version of Taylors work – within the SE domain. The collection of essays provides valuable insights concerning the status of EBSE and its literature, the problems associated with secondary research under the paradigm, and the basis for a discrimination function designed to assist in resolving a key issue for those seeking guidance in academic literature.
DEDICATION

To Georgette, I think I am finally done – but just getting started.
LIST OF ABBREVIATIONS AND SYMBOLS

$D$  Matrix of orthonormal vectors representing documents of a term-document matrix

$D_q$  Scaled pseudo-document vector representing a query

DOI  Digital Object Identifier

EBM  Evidence-Based Medicine

EBP  Evidence-Based Practice

EBSE  Evidence-Based Software Engineering

$GWF$  Global Weight Function

$i$  Row index of a cell in a matrix

$j$  Column index of a cell in a matrix

LSA  Latent Semantic Analysis

$LWF$  Local Weight Function

QGS  Quasi-Gold Standard

$S$  Diagonal matrix of the singular values of a term-document matrix

SE  Software Engineering

SLR  Systematic Literature Review

SMS  Systematic Mapping Study

SVD  Singular Value Decomposition

$T$  Matrix of orthonormal vectors representing terms of a term-document matrix

$t$  The number of dimensions in a semantic space
term Weighted value of a vocabulary term in a semantic space

URL Uniform Resource Locator

VSM Vector Space Model

$X$ A term-document matrix

$X_q$ Pseudo-document vector representing a query
ACKNOWLEDGMENTS

I want to thank everyone who made this journey possible. First and foremost, I want to thank Georgette, my wife, for supporting my decision to go back to school again and for her patience during the entire process. I want to thank my daughters and family for their encouragement. I want to thank Dave and Joanne Hale for the opportunity to return to school mid-career and for all the support and mentoring they so willingly provided along the way. Finally, I want to thank my committee for their guidance and all the faculty members who provided the lessons that made this work possible.

This material is based upon work supported in part by the U.S. National Science Foundation under Grant No. 1305395. The author is grateful to all involved in the study, particularly the workshop participants for their insights and cooperation and the ESEM 2013 conference organizers for providing a meeting room for the workshop.
CONTENTS

ABSTRACT ........................................................................................................................................ ii
DEDICATION ................................................................................................................................ iv
LIST OF ABBREVIATIONS AND SYMBOLS .................................................................................... v
ACKNOWLEDGMENTS ................................................................................................................ vii
LIST OF TABLES ........................................................................................................................ xiv
LIST OF FIGURES ....................................................................................................................... xv
CHAPTER 1 Introduction ........................................................................................................ 1
  1.1 Research Objective ............................................................................................................. 3
  1.2 Foundational Concepts ..................................................................................................... 3
    1.2.1 The Roots of EBP ........................................................................................................ 4
    1.2.2 The Essence of EBP .................................................................................................... 5
    1.2.3 EBP in Software Engineering .................................................................................... 6
  1.3 Organization of the Research ............................................................................................ 7
CHAPTER 2 Evidence-Based Software Engineering: A Mapping Study .................................. 8
  2.1 Introduction ....................................................................................................................... 8
    2.1.1 Research Objectives ................................................................................................. 9
2.2 The Systematic Mapping Process and Research Questions

2.2.1 Search for Relevant Publications

2.2.2 Selection of Publications

2.2.3 Classification Based on Evidence Type

2.2.4 Categorization Based on EBP Models

2.2.5 Documentation of Results

2.3 Analysis of Publications

2.4 Results

2.4.1 Search Results

2.4.2 Selection Results

2.4.3 Classification Results

2.4.4 Categorization Results

2.5 Discussion

2.5.1 Identifiability of EBSE Research

2.5.2 Models and Gaps

2.5.3 Evidence Types and Maturity

2.5.4 The Essence of EBSE

2.5.5 Implications for Research

2.5.6 Lessons Learned Concerning the SMS Protocol
4.3 Conceptual Development ............................................................................................ 115
  4.3.1 Goals in the Selection Process ............................................................................. 115
  4.3.2 Assessment of the Selection Process ................................................................. 115
  4.3.3 Characteristics of the Selection Process ............................................................. 117
  4.3.4 Proposed Classifier ............................................................................................ 118
  4.3.5 Factors Impacting LSA Performance ................................................................ 119

4.4 Classifier Evaluation ............................................................................................... 121
  4.4.1 Dataset .............................................................................................................. 121
  4.4.2 Classification of Citations .................................................................................. 121
  4.4.3 Analysis ............................................................................................................. 122

4.5 Results ...................................................................................................................... 123

4.6 Discussion ............................................................................................................... 126
  4.6.1 Implications ...................................................................................................... 127
  4.6.2 Limitations ....................................................................................................... 129

4.7 Conclusion and Future Work .................................................................................. 130

4.8 References ............................................................................................................. 130

CHAPTER 5 Conclusion ................................................................................................. 133

5.1 The EBSE Landscape .............................................................................................. 134

5.2 Barriers in the SLR Process ................................................................................... 135
5.3 Addressing a Priority Issue of the SLR Process ......................................................... 136

5.4 Continuing and Future Work ...................................................................................... 137

REFERENCES ................................................................................................................ 139

APPENDIX A IRB Certificate ........................................................................................ 142
LIST OF FIGURES

Figure 1. Key Concepts and Relationships ................................................................. 4
Figure 2. Mapping Study Process and Outcomes (adapted from Petersen et al. 2008) .... 10
Figure 3. QGS-Based Systematic Search Process (Adapted from He et al. 2011) ........ 13
Figure 4. Publications by model and type ................................................................. 30
Figure 5. Structural Model of Evidence-Based Practice (Satterfield et al. 2009) ......... 33
Figure 6. Structural model publications ................................................................. 36
Figure 7. Original Process View of Evidence-Based Practice .................................... 37
Figure 8. Alternative Process View of Evidence-Based Practice ............................. 38
Figure 9. EBSE process publications ................................................................. 40
Figure 10. High Level SLR Process Flow .............................................................. 43
Figure 11. Low Level SLR Process Flow .............................................................. 44
Figure 12. SLR/SMS protocol publications .......................................................... 49
Figure 13. The EBP Ecosystem ............................................................................. 50
Figure 14. EBSE paradigm and ecosystem publications ........................................ 52
Figure 15. Publication Counts within the EBSE Paradigm ...................................... 55
Figure 16. High Level SLR Process Flow .............................................................. 74
Figure 17. Low Level SLR Process Flow .............................................................. 75
Figure 18. Idea Sharing Phase .............................................................................. 84
Figure 19. Group Discussion Phase ...................................................................... 85
Figure 20. Voting and Ranking Phase ................................................................................... 86

Figure 21. Classifier versus origin decision outcomes ............................................................ 122
CHAPTER 1

Introduction

“Know where to find information and how to use it. That’s the secret to success.”

- Albert Einstein (1879 – 1955)

The concept of utilizing information derived from carefully crafted scientific research to optimize the efficiency and/or effectiveness of a practice is by no means a new idea. The ideals underlying the concept can arguably be traced back to the earliest physicians and scientists more than two thousand years ago (Plato 2009). More recently, one hundred years ago Frederick Taylor describes the use of scientific experimentation to study and optimize the performance of workers carrying out various tasks within a business context in his bestseller The Principles of Scientific Management (1914). Indeed, the discovery of the well-known Hawthorne Effect – people changing their behavior when they believe they are being observed – was the result of the Western Electric Company’s effort to gather empirical data regarding productivity increases due to lighting adjustments beginning in 1924 (Gale 2004).

Sackett and Rosenberg (1995b) suggested the need for EBM in order to incorporate the most current research findings into clinical practice. They propose three essential strategies for adopting EBM:

1. “Become life-long, self-directed learners of EBM”
2. Seek and utilize summaries – especially those reviews that are systematically executed to alleviate bias and
3. Accept evidence-based protocols to overcome the issue of becoming outdated

They note several additional benefits of the use of EBM such as the continuing education of practitioners (Shin, Haynes, and Johnston 1993), improved efficiency and performance (Davis et al. 1992), improved communication through the development of a common language (Sackett and Rosenberg 1995b), improved consistency in the education of clinicians (Bennett et al. 1987), and a reduction in the burden of both learning and teaching within the domain (General Medical Council 1993).

The ideals embodied in the concept along with the infrastructure needed to support EBM have come to be known generically as evidence-based practice (EBP) (Ilott 2012, Reynolds and Trinder 2008, Gray, Plath, and Webb 2009, Charters et al. 2009) and have served as a template for other healthcare related fields (Kitson, Harvey, and McCormack 1998, Rousseau 2006, Satterfield et al. 2009), such as nursing (Lusardi 2012) and public health (Brownson, Gurney, and Land 1999), as well as education (Davies 1999), social work (Gray, Plath, and Webb 2009), software engineering (Kitchenham, Dyba, and Jorgensen 2004a), and management (Pfeffer and Sutton 2006). Thus, the current instantiation of EBP and its associated literature represents the use of evidence to inform practice as demonstrated by its utilization as the foundation for other fields seeking to implement the use of EBP.
The current literature on EBP suggests best practice to be an extension of Taylor’s use of scientific evidence for decision making by broadening the sources from which to draw evidence (Pawson 2006) and increasing the volume of evidence (Sackett et al. 1996). Whereas Taylor primarily relied on evidence generated through experimentation within his own company (Taylor 1914), today’s practitioners have numerous sources, both internal and external to the organization (Pfeffer and Sutton 2006), from which to draw evidence for decision making purposes.

1.1 Research Objective

Over the past decade, within the Software Engineering (SE) discipline, work has begun to adopt the principles of EBP to form the Evidence-Based Software Engineering (EBSE) paradigm (see Kitchenham, Dyba, and Jorgensen 2004b). The objective of this research is to:

4. Map the research completed to date regarding the implementation of EBP in the SE domain to identify gaps and opportunities in the research.

5. Identify the barriers deemed most important by the members of the SE research community who conduct systematic literature reviews in support of EBSE.

6. Develop the use of algorithmic techniques as a discriminant function in the selection process of the systematic review methodologies.

1.2 Foundational Concepts

In order to form a baseline to assess the status of EBSE, it is necessary to first understand the overall concept of EBP and its intentions. Figure 1 depicts the evolution of EBSE as a form of EBP, which in-turn has evolved from Evidence-Based Medicine.
The remainder of this section provides an overview of the roots of the EBP paradigm, the core intentions that guide its implementation, and the introduction of EBP into the domain of software engineering.

1.2.1 The Roots of EBP

Evidence based practice has been in the medical literature for over 50 years (see Zientz et al. 1960), only recently has it become widespread. In the early nineteen seventies, an initiative was established that has grown into the Cochrane Collaboration, which brings together over 28,000 health career providers in more than 100 countries (The Cochrane Collaboration 2013a). The collaboration was originally formed as the Cochrane Centre to facilitate systematic reviews in the domain of healthcare for randomized controlled trials (The Cochrane Collaboration 2013b). Since inception, it has become “an international, non-profit, independent organization [sic], established to ensure that up-to-date, accurate information about the effects of healthcare interventions is readily available worldwide” (The Cochrane Collaboration 2013d). The goal of the collaboration is “the protection and preservation of public health through the preparation, maintenance and dissemination of systematic reviews of the effects of health care, for the public benefit” (The Cochrane Collaboration 2013c).

With the increased linkage of electronic medical research and medical data bases with accessibility via the internet, the foundation was set for Sackett and Rosenberg (1995b) to reiterate the need for EBM in order to incorporate the most current research findings into clinical practice.
Sackett and Rosenberg (1995b) noted several additional benefits of the use of EBM such as the continuing education of practitioners, improved efficiency and performance, improved communication through the development of a common language, improved consistency in the education of clinicians, and a reduction in the burden of both learning and teaching within the domain.

The success of EBM has been a primary driver in the renewed interest for utilization of scientific evidence to inform domains of practice within the realm of the other healthcare domains. The field of modern medicine is one of the first domains to successfully institutionalize the use of evidence-based practices (Rousseau 2006). This instantiation of evidence-based practices and its associated literature represents a modern interpretation of the use of evidence to inform practice as demonstrated by its use as the foundation for other fields seeking to implement EBP.

1.2.2 The Essence of EBP

The kernel of EBP is the integration of the best available evidence from systematic research with the expertise of the individual practitioner for the purpose of improved decision-making (Sackett et al. 1996). The use of the most current information, tempered by expertise, to provide the best possible decisions is a goal for practitioners and their associates in the business world.

Sackett et al. (1996) provides a definition of EBM most commonly cited when discussing EBP in general:

Evidence-based medicine is the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients.

Examining this definition for its essence, one can see that central to EBP is the improvement of decision-making with regard to patient care (Sackett and Rosenberg 1995b). The definition emphasizes the care with which decision-making should be approached (conscientious and
explicit), the use of evidence that is relevant to the situation (judicious) and representative of the most recent advances in knowledge (current best evidence) (Straus et al. 2011). It also recognizes the individual expertise of the clinicians and their ability to integrate their experiences with other evidence available to them (Sackett et al. 1996). In the case of EBP, the best available evidence is often considered to be evidence derived from systematic research (Sackett 2000). The emphasis on systematic research is due, in part, to ethical considerations in the practice of medicine and the clinician's desire to “do no harm“. Before treatments are applied to patients in practice, extensive research into both the intended and unintended consequences must be undertaken. Only after the efficacy of a treatment has been established and the associated risks considered to be outweighed by the benefits is a treatment employed in practice (Sackett and Rosenberg 1995a).

As EBP has been ported to other domains, the strict adherence to the use of quantitative data has been relaxed and adapted to better suit the domain of implementation (Gray, Plath, and Webb 2009, Satterfield et al. 2009). Establishing what constitutes best evidence in a given domain is a fundamental point in regard to implementing EBP in the domain of interest1 (Hjørland 2011).

1.2.3 EBP in Software Engineering

In 2004, Kitchenham, Dyba, and Jorgensen proposed the concept of EBSE as potentially beneficial to both SE practice and research (Kitchenham, Dyba, and Jorgensen 2004b). As defined by Kitchenham, Dyba, and Jorgensen (2004b) the goal of EBSE is “to provide the means by which current best evidence from research can be integrated with practical experience and

---
1 Due to the nature of the discussion concerning what constitutes evidence and how it should be evaluated, a generalized description is beyond the scope of the current study. Interested readers are referred to Pawson (2006), Reynolds and Trinder (2008), and Straus et al. (2011) for additional background regarding definitions of evidence.
human values in the decision making process regarding the development and maintenance of software”. In the same year, Kitchenham proposed the first guidelines for the production of a systematic literature review (SLR) – a specific type of secondary study considered a critical research method utilized to find and synthesize primary studies in SE – in support of the EBSE paradigm (Kitchenham 2004).

1.3 Organization of the Research

Each of the chapters that follow address one of the objectives previously outlined and provides a progressive step on which subsequent chapters build. In the first essay (Chapter 2), a systematic mapping methodology is utilized to identify and map the EBSE literature. From the map, gaps and opportunities for further research are identified.

Based on the substantial body of literature regarding the EBSE subdomain of systematic reviews uncovered in the first essay, the second essay (Chapter 3) seeks to identify the barriers deemed most important by the members of the SE research community that conduct systematic literature reviews in support of EBSE.

The third essay (Chapter 4) investigates one of the high priority barriers identified in the second essay – the time and resources needed to select a study for inclusion in a review. Two potential algorithms for automating the selection process are identified and then evaluated utilizing the set of studies uncovered during the mapping portion of the research.

Together, the three essays represent a cohesive line of inquiry in a larger research stream to facilitate the advancement of EBSE in both practice and research.
CHAPTER 2
Evidence-Based Software Engineering: A Mapping Study

2.1 INTRODUCTION

In 2004, Kitchenham, Dyba, and Jorgensen proposed the concept of evidence-based software engineering (EBSE) as potentially beneficial to both software engineering (SE) practice and research (Kitchenham et al. 2004). The EBSE paradigm is an adaptation of evidence-based practice (EBP) as exemplified by the healthcare domains. As defined by Kitchenham et al. (2004) the goal of EBSE is “to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software”. In the same year, Kitchenham proposed the first guidelines for the production of a systematic literature review (SLR) – a specific type of secondary study considered a critical research method utilized to find and synthesize primary studies in SE – in support of the EBSE paradigm (Kitchenham 2004).

In the decade following the introduction of the EBSE paradigm, there has been considerable growth in the body of literature concerning it. Indeed, more than 290 SLRs have been published demonstrating its growth and acceptance. Furthermore, the use of the systematic mapping study (SMS) has been adopted to assist researchers in identifying areas in need of synthesis and areas in which gaps exist or current research is inadequate (Petersen et al. 2008).

The SLR process has been utilized in the examination of the SLR process itself (Kitchenham and Brereton 2013), researchers’ experiences conducting SLRs (Imtiaz et al. 2013), the SLRs
being produced (Cruzes and Dyb 2011, Kitchenham et al. 2010), the impact of SLRs (Kitchenham et al. 2009), and the automation of the SLR process (Hamad and Salim 2014). The SMS process has also been directed at the SLR process for examination of visual data mining techniques in support of conducting SLRs (Felizardo et al. 2012).

While the SLR process continues to be the focus of summarizing research, it is only one component of the EBSE ecosystem. The implementation of EBSE in the domain of practice and the infrastructure needed to support EBSE as a paradigm are additional components of the EBSE ecosystem (Kitchenham et al. 2004). Despite a decade of research, these components have not been examined in either SLR or SMS research. Additionally, there has been no holistic examination of the EBSE ecosystem itself.

Thus there is a need to examine the EBSE paradigm from a holistic viewpoint in order to determine gaps in the research literature and opportunities for the summarization of existing research regarding the EBSE paradigm and all of its components.

2.1.1 Research Objectives

This study maps the research completed to date regarding the implementation of EBP in the SE domain. Specifically, through the lens of EBP as exemplified in other domains, this study seeks to systematically map the areas of the EBP paradigm in need of development within the SE domain.

The remainder of this chapter is organized as follows. Section 2.2 presents the SMS protocol utilized in conducting the study. A description of the analysis process utilized is presented in Section 2.3. In Section 2.4 the detailed results of the research are presented, followed by discussion and limitations of the work in Section 2.5. Finally, Section 2.6 proposes future directs and concludes the work.
2.2 The Systematic Mapping Process and Research Questions

A Systematic Mapping Study (SMS) methodology was selected to investigate and assess the research completed to date regarding EBSE, and the SLR and SMS protocols in SE. A mapping study provides a broad overview of an area of research (Kitchenham and Charters 2007). It is utilized to uncover the quantity, extent, nature, and structure of literature in a given domain (Petersen et al. 2008, Budgen, Turner, et al. 2008). Mapping studies also provided a basis for further research (Kitchenham et al. 2011, Petersen et al. 2008, Budgen, Turner, et al. 2008).

After defining the research questions to be answered, a protocol for conducting the mapping study is prepared (Kitchenham and Charters 2007). The protocol details the strategy and process to be utilized in conducting the four primary stages of the mapping study consisting of searching, selecting, quality assessing, and analyzing existing literature. (Budgen, Turner, et al. 2008)

The product of systematically mapping the EBSE literature results in defining the quantity, extent, nature, and structure of literature regarding EBSE, SLR and SMS research within the context of an EBP ecosystem. Petersen et al. (2008) provides the basis for the systematic mapping process utilized in this study as shown in Figure 2.

![Figure 2. Mapping Study Process and Outcomes (adapted from Petersen et al. 2008)](image-url)
The process proceeds through six distinct steps. Each step produces an outcome that is then utilized in the execution of a following step in the process. The culmination of this process is the systematic map produced as the outcome of the final step in the process.

The initial step of the process is to define the research being undertaken. This includes the research questions, scope of the research, and any constraints. The outcome of this initial step is the protocol utilized in executing the study, and for the current study is presented in the remainder of this section and its subsections. The study then proceeds through five stages:

1. A systematic search to identify relevant publications (detailed in § 2.2.1).
2. The selection of publications for inclusion in the study (detailed in § 2.2.2).
3. The classification of publications based on evidence type detailed in (§ 2.2.3).
4. The categorization of publication based on EBP models (detailed in § 2.2.4).
5. The analysis of publications to answer the research questions (detailed in § 2.3).

After completion of the stages, a report detailing the results and summarizing the findings is prepared. The goal of the report is to provide a transparent, repeatable account of the study and, via the findings, direct the future efforts of primary studies (Petersen et al. 2008, Kitchenham and Charters 2007).

The content-based research questions addressed by this study are as follows:

*RQ1:* Is the population of EBSE (as a whole) and the components of the EBSE ecosystem literature uniquely identifiable?

*RQ2:* Does the population of the literature represent the conceptual models that describe EBSE and its ecosystem?

*RQ3:* What type of evidence does the population of literature represent?
RQ4: What are the gaps between the conceptual models used as the topologies and the population of EBSE and EBSE ecosystem literature?

2.2.1 Search for Relevant Publications

The intent of the search strategy is to identify and collect all of the relevant literature as defined by the inclusion and exclusion criteria (see Section 2.2.2.1 and 2.2.2.2). To accomplish this task, a systematic search approach (He et al. 2011) is implemented. The results of the overall search assist in answering the first research question (RQ1).

The complete set of relevant publications within a search universe is defined as the Gold Standard (He et al. 2011). The sensitivity of a search may be established by examining the portion of relevant papers retrieved. Similarly, the precision of a search may be established by examining the portion of publications retrieved that are relevant (He et al. 2011).

\[
Sensitivity = \frac{\text{Relevant publications retrieved}}{\text{Total number of relevant publications}} \times 100
\]

\[
Precision = \frac{\text{Relevant publications retrieved}}{\text{Number of publications retrieved}} \times 100
\]

As the goal of the search in a mapping study is to retrieve the largest portion of the relevant papers as possible, the sensitivity of the search is more important than the precision. This creates a problem however, as the true number of relevant studies cannot be known. As an alternative measure, He et al. (2011) recommend the use of a quasi-sensitivity measure, defined as the portion of a known set of relevant paper retrieved. For this measure, a known set of relevant publications (referred to as the Quasi-Gold Standard) serves as a proxy for the unknown gold standard set of relevant publications.
It is termed acceptable when the quasi-sensitivity of the search meets or exceeds a minimum threshold of 70% (He et al. 2011).

As illustrated in Figure 3, the Quasi-Gold Standard (QGS) based approach (He et al. 2011) first establishes a standard set of publications based on a manual search of known venues. The standard set of papers is then utilized to evaluate the results of an automated search process via the quasi-sensitivity measure. An acceptability threshold of 90% was established to ensure high sensitivity (He et al. 2011) in the search process of the current study.
2.2.1.1 *Search: Manual Process.* For the search at hand, two venues were selected to establish the QGS set of publications:


The venues were selected as each is known to publish state-of-practice work regarding EBSE and SLR methodology. The results of the manual search of each venue are further supplemented by the addition of 63 papers known by the author to be relevant.

2.2.1.2 *Search: Automated Process.* For the automated search, a number of search engines are utilized. With the exception of Google Scholar and the ACM Digital Library, there is very little overlap among the major search engines servicing the computer science and information systems disciplines (Bailey et al. 2007). However, Google Scholar limits the availability of results to first 1000 records returned (Google Inc. 2014) and the ACM Digital Library lacks the ability to provide bulk downloading of search results. Therefore, a variety of sources must be examined to ensure coverage of a topic. The following resources were searched in order to identify and collect relevant publications:

- IEEEExplore
- Science Direct
- Web of Science
- Compendex/INSPEC
- Scopus
- ACM Digital Library

Due to the limitations of the ACM Digital library engine, the other five engines will be searched first. After completion of this initial search, the selection process (see Section 2.2.2) is completed. With the initial results in hand, the search of the ACM Digital Library is undertaken.
After entering queries, the results were reviewed online with relevant publications not found in
the existing set being downloaded and added to the set of relevant results.

For each of the selected search engines the following universal search string was utilized:

```
(((systematic AND (review or reviews))
OR (mapping AND (study OR studies))
OR (evidence AND based))
AND ‘software engineering’)
OR EBSE
```

Due to the nuances of each search engine, the search string was decomposed into substrings
and customized in order to be executed on each of the selected search engines. This was
accomplished by division along the OR logical operators while keeping the overall logical AND
operations intact.

### 2.2.2 Selection of Publications

After completion of the automated search, the following process and criteria are utilized in
the selection of publications for inclusion in the study. Each publication found during the search
process is reviewed in multiple rounds for inclusion in, or exclusion from, the study. The initial
round is based on titles. The second round is based abstracts. Finally, a full reading of each
article is completed. In each round, a random sample consisting of a minimum of 20% of the
articles identified for exclusion are reviewed by a second author for confirmation of exclusions.

#### 2.2.2.1 Selection: Inclusion Criteria.

The inclusion criteria are as follows:

- Publications that discuss EBSE
- Publications that report on the process of SLR or SMS
- Publications that report on the use of SLR or SMS
- Publications that report on tools supporting the SLR or SMS process
2.2.2.2 Selection: Exclusion Criteria. The exclusion criteria are as follows:

- Publications not meeting the inclusion criteria
- Publications prior to 2004
- Duplicate reports of the same study (when several reports of a study exist in different venues the most recent version of the study will be included in the study)
- Non-English language publications

The decision to exclude publications prior to the year 2004 is based on the first appearance of EBSE in the literature during that year.

2.2.3 Classification Based on Evidence Type

As previously noted, the objective of the study is to assess the quantity, extent, nature, and structure of literature regarding EBSE, SLR and SMS research within the context of an EBP ecosystem. Thus a classification scheme is needed to classify each publication based on the type of evidence it offers.

Wieringa et al. (2006) have proposed a classification scheme for requirements engineering articles that is based on the activities found in the engineering cycle. While useful from the perspective of investigating solutions provided to practice, the classification scheme concentrates primarily on applied research while ignoring basic and theoretical research. Additionally, the classification scheme provides no indication as to how the results were produced: A key indicator when determining validity and generalizability.

An alternative classification scheme proposed by Montesi and Lago (2008) focuses on the genre of a given publication and includes specific methodologies found in existing literature. However, the classification descriptions are not mutually exclusive, thus a publication may be placed in several genres simultaneously. For example, a case study that is published following
enhancements made based on feedback from a conference presentation would properly be classified under three genres: Publishable Paper; Empirical Research reports – Case Studies; and Extended Versions of Conference Papers. Additionally, the methodologies selected for inclusion do not include several basic methodologies such as interviews, focus groups, or simulations.

Evaluation of the type of evidence present in each publication requires a classification system that provides mutually exclusive classification and a distinction among rigorous scientific research, theoretical work, opinion pieces and general announcements used to inform the community. The existing classification systems are found lacking in at least one aspect of the requirements. Thus, a classification system founded in research design and capable of accommodating non-research design publications is created. The results of this classification will be utilized in answering the third research question (RQ3).

### 2.2.4 Categorization Based on EBP Models

In order to answer the second and forth research questions (RQ2, RQ4), the literature must be related to each of the known EBP models through a coding process. Since there is no known categorization system for this purpose, a system of categories based on the existing EBP models is developed. The four models identified for examination include:

1. The structural model of EBP (Satterfield et al. 2009).
2. The process model of EBP (Sackett and Rosenberg 1995).
3. The systematic literature review process (Kitchenham and Charters 2007).
4. The EBP ecosystem model.

For each model, an examination of the model and its components is completed to establish an initial set on categorizations along with a description of each category. The aggregation of the categories from the individual model serves as the initial categorization system of the study (see
Section 2.2.5). As all possible deviations from the models due to the domain of implementation cannot be foreseen, the system is subject to further augmentation during the coding process in the analysis phase of the study.

2.2.5 Documentation of Results

Results of the search are recorded into a Microsoft Excel spreadsheet for tracking purposes. The spreadsheet includes the following items:

- Author
- Paper Title
- Year of Publication
- Publication Venue
- URL
- DOI
- Database Searched
- Indicator of Inclusion/Exclusion
- Reason for Inclusion/Exclusion

For publications that are to be included in the study, the following additional items are also recorded:

- Indicator for inclusion in the quasi-gold standard
- Classification of research type
- Primary area of EBSE ecosystem addressed
- Indicators of subareas addressed within the major area
2.3 Analysis of Publications

According to Corbin and Strauss (2008), literature can provide concepts and relationships that are relevant to the emerging theory; thus, these concepts and relationships may be validated within the context of the study. Literature may also be used to provide insight or new ways of approaching the data. Literature may provide a starting point for data collection, to stimulate questions to ask of the data. After the theory has been developed (or at least development has paused to allow for the purpose of writing), the literature may be used as supplemental validation of, or contrast to, the findings.

Further, literature may be used as sources of primary data, to be analyzed as one will analyze interview or observational data. "Every book, every magazine article, represents at least one person who is equivalent to the anthropologist's informant or the sociologist's interviewee" (Glaser and Strauss 1967, 163). The use of published literature as data has several advantages: they are accessible, in consideration of the effort, cost, and speed of data gathering; and they provide an excellent range of comparison groups that otherwise may be unattainable (Charmaz 2006). In contrast, there are limitations to the usefulness of literature: data are often offered only in summarized, interpreted form, preventing the researcher from accessing the richer raw data; in some cases, the literature may be inaccurate or misleading; and the researcher is denied the richness of his or her own experience (Corbin and Strauss 2008). Therefore, caution must be used in relying on another's interpretation or summary of incidents, ideas, and events (Corbin and Strauss 2008).

The literature found in the previous search and selection process is thus analyzed as the data on the EBSE ecosystem bearing in mind the noted cautions. The analysis is carried out first from an ecosystem perspective with the objective of determining which of the models are addressed
by the publication. After determining models addressed by the publication, the components within the models are determined.

Within each model, coding is utilized to identify the major themes found in each publication. Coding is the process of breaking down the data (Corbin and Strauss 2008) for the purpose of discovering theoretical concepts, their properties, and the relationships among concepts. Data may be examined at many levels; the analyst may examine the data elements of words, phrases, sentences, paragraphs, or entire documents (Corbin and Strauss 2008). Further, a researcher may move among the levels to examine the data from multiple perspectives. Coding entails the examination of a data element and attaching to it a meaningful conceptual label (Charmaz 2006, Corbin and Strauss 2008, Strauss 1987). Coding often takes place in the form of margin notes. However, it could also be done on note cards, or in the form of data base elements. A critical element is that coding should not be a mere restatement of the data. Theoretical elements must be at a higher conceptual level than the data it is derived from (Strauss 1987, Charmaz 2006). Through iterative rounds of refinement, a set of subtopics is created that address components found within all major areas. The previously derived categories are utilized as a starting point for the coding with new categories emerging as needed in the analysis.

As a result of this process, each publication is assigned one classification to indicate the type of evidence present, and one or more category codes indicating the model(s) addressed along with the sub-components identified.

Primary document analysis is accomplished by the author. A random sample consisting of 25% of the publications was checked by a second doctoral researcher for accuracy. Any discrepancies were discussed until agreement was reached.
2.4 RESULTS

2.4.1 Search Results

In reference to the first research question (RQ1): Is the population of EBSE and EBSE ecosystem literature uniquely identifiable? This study finds in the affirmative. It is clear that the population of EBSE and EBSE ecosystem literature can be uniquely identified given a sufficiently broad search and clear inclusion/exclusion criteria. This conclusion is supported by the results of the search process as described in this section and the results of the selection process in the next section.

The manual search of the ESEM and EASE conference proceedings when combined with the known set of publications yielded a QGS consisting of 115 publications. Of the origin 63 known publications, a total of 22 publications overlapped with the results of the conference venue searches. The detailed results of the manual search are summarized in Table 1.

<table>
<thead>
<tr>
<th>Venue</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>ESEM</td>
<td>40</td>
</tr>
<tr>
<td>EASE</td>
<td>43</td>
</tr>
<tr>
<td>Other Known</td>
<td>32</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>115</strong></td>
</tr>
</tbody>
</table>

The automated search utilizing the selected search engines returned a total of 12,429 publications. As shown in Table 2, the Compendex/INSPEC search engine returned the largest number of results followed closely by the ACM Digital Library. After removal of duplicates, a final total of 6,362 publications remain for processing in the selection phase of the study.
Table 2. Automated search results

<table>
<thead>
<tr>
<th>Database</th>
<th>Publications</th>
</tr>
</thead>
<tbody>
<tr>
<td>IEEExplore</td>
<td>1661</td>
</tr>
<tr>
<td>Science Direct</td>
<td>226</td>
</tr>
<tr>
<td>Compendex/INSPEC</td>
<td>4820</td>
</tr>
<tr>
<td>WOS</td>
<td>5</td>
</tr>
<tr>
<td>Scopus</td>
<td>1066</td>
</tr>
<tr>
<td>ADM Digital Library</td>
<td>4651</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>12429</strong></td>
</tr>
</tbody>
</table>

2.4.2 Selection Results

Examining the titles of publications found by the automated search eliminated 5,709 of the search results. A random sample of 1427 (25%) of the eliminated titles was examined by a second evaluator for agreement. The second evaluator questioned the exclusion of two of the publications. This disagreement was resolved by including the publications for further consideration.

Further examination of the publication abstracts eliminated an additional 87 publications. The entire list of proposed eliminated was examined by the second evaluator with no disagreements noted.

In the final round, the full text of each of the remaining papers was examined. As a result, 24 of the publication were found to be not applicable and proposed for elimination. The second evaluator examined the entire list of publications proposed for elimination and found no disagreements. Table 3 summarizes the elimination of publications.
Table 3. Publication selection results

<table>
<thead>
<tr>
<th>Process Step</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Publications Found</td>
<td>6362</td>
</tr>
<tr>
<td>Eliminated by Title</td>
<td>-5709</td>
</tr>
<tr>
<td>Eliminated by Abstract</td>
<td>-87</td>
</tr>
<tr>
<td>Eliminated by Reading</td>
<td>-24</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>542</strong></td>
</tr>
</tbody>
</table>

After completion of the selection process, the remaining 542 publications were compared to the QGS set of publications. A total of 107 publications from QGS list were found in the automated search and selection process results. Thus a quasi-sensitivity of 93.04% was achieved by the search and selection process. Only eight publications from the QGS set were not found during the search. Closer examinations of the eight missing publications from the QGS list reveals they are not indexed by conventional search engines. The eight publications were added to the automated search and selection results to form a final set of 550 selected publications.

Of the 550 publications selected for inclusion, a total of 383 publications were found to be instances of either the SLR protocol (291) or SMS protocol (92) that addressed a topic outside the scope of the EBSE. While many of the 383 studies provide insight into the production of SLRs through accounts of the protocol and lessons learned, they have been removed from the analysis process as the study seeks to examine publications specifically targeting EBSE and its processes. Future research specifically targeting each of the secondary protocols should include such studies in the analysis.

Similarly, six of the publications recovered are instances of the use of EBSE principals in practice. As these publications do not directly address the EBSE ecosystem or the development of a major area of interest, they have been removed from further analysis. Additionally, there are four publications found by the search process for which the full text of the article could not be
retrieved. As the titles and abstracts suggest they meet inclusion, without the full text of each 
they cannot be conclusively excluded from consideration. Regardless of this fact, they are also 
excluded from the analysis with the provision that the exclusion represents a potential limitation 
in the study. Thus a total of 157 publications are included in the analysis that follows.

### 2.4.3 Classification Results

As the existing classification systems are found lacking for the purposes of this study, a 
classification system founded in research design and capable of accommodating non-research 
design publications is created and utilized to identify the type of evidence present in each 
publication. In this section, the development of the classification scheme is presented followed 
by the results of its application to the selected publications.

#### 2.4.3.1 Classification Scheme Development

The categorization scheme utilized in this study seeks to overcome the limitations of the previous classification schemes. It 

begins with the premise that research serves four purposes: Describe phenomena 
and its basic components.

1. Describe phenomena and its basic components.
2. Explain relationships among phenomena or its components.
3. Predict the occurrence or behavior of phenomena in a given context.
4. Prescribe a treatment in order to induce the occurrence or behavior of phenomena 
in a given context.

The description of phenomena implies the use of observational methods were thick, rich 
accounts of phenomena are produced. Such methods are typically considered qualitative 
methods and include four primary designs (Trochim 2006):
1. **Participant Observation** – Study in which the researcher is a participant in the culture or context under observation

2. **Direct Observation** – Study in which the researcher unobtrusively as possible observes the culture or context of interest

3. **Interview** – Study in which the researcher directly interacts with a respondent or group to obtain information via discussion

4. **Case Study** – Study in which the researcher undertakes an intensive study of a specific individual or context and may include a combination of methods

Qualitative methods may also assist in the construction of theory that explains a phenomenon and its relationships (Eisenhardt 1989). However it is common to compliment such research with quantitative investigations (Cozby 2008, p108). Quantitative methods can be described coarsely by four classifications (Schwab 2005):

1. **Experiment** – Study in which a probability or convenience based sample is randomly assigned to cases and the values of independent variables are controlled by the researcher

2. **Quasi-Experiment** – Study in which a probability or convenience based sample is non-randomly assigned to cases and the values of independent variables are either controlled by the researcher or an exogenous event

3. **Survey** – Study in which a probability based sample is non-randomly assigned to cases and the values of independent variables are inherent to each case

4. **Field Study** – Study in which a convenience based sample is non-randomly assigned to cases and the values of independent variables are inherent to each case
The use of multiple methods in research designs in which the selected methods are intended to compensate for weaknesses in the other selected methods is known as *triangulation* (Schwab 2005, p208) and forms an addition classification to distinguish such research. Only after the description and explanation of phenomena has occurred can predictive and prescriptive studies be accomplished.

While the traditional social science classifications provide coverage for the majority of methods utilized in primary research, they do not emphasize the following specialized types of evidence collection:

1. **Archival** – Study based on information extracted from archival records (Cozby 2008, p116)
2. **Simulation** – Study based on imitation of a phenomena via a model of the phenomena (Cozby 2008, p48)
3. **Illustration** – Study based on demonstration, “proof by example”, and other such foundational techniques
4. **Secondary** – Study based on the synthesis of results from primary studies

Finally, to account for viewpoint types of publications noted in previous studies (Wieringa et al. 2006, Montesi and Lago 2008), four additional classifications are necessary:

1. **Tutorial** – Publications that provide instruction on the use of a technique or tool in practice
2. **Experience** – Publications emphasizing the personal experiences of the author regarding the application of a technique or tool in practice
3. **Opinion** – Publications containing the author's opinion about what is good or wrong about something
4. *Announcement* – Publications which announce events to the community

Table 4 summarizes the classifications and provides a brief definition of each.

<table>
<thead>
<tr>
<th>Source of Evidence</th>
<th>Publication Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative Studies</td>
<td>Experiment</td>
<td>Study in which a probability or convenience based sample is randomly assigned to cases and the values of independent variables are controlled by the researcher.</td>
</tr>
<tr>
<td>(Schwab 2005)</td>
<td>Quasi-Experiment</td>
<td>Study in which a probability or convenience based sample is non-randomly assigned to cases and the values of independent variables are either controlled by the researcher or an exogenous event.</td>
</tr>
<tr>
<td></td>
<td>Survey</td>
<td>Study in which a probability based sample is non-randomly assigned to cases and the values of independent variables are inherent to each case.</td>
</tr>
<tr>
<td></td>
<td>Field Study</td>
<td>Study in which a convenience based sample is non-randomly assigned to cases and the values of independent variables are inherent to each case.</td>
</tr>
<tr>
<td>Qualitative Studies</td>
<td>Participant Observation</td>
<td>Study in which the researcher is a participant in the culture or context under observation.</td>
</tr>
<tr>
<td>(Trochim 2006)</td>
<td>Direct Observation</td>
<td>Study in which the researcher unobtrusively as possible observes the culture or context of interest.</td>
</tr>
<tr>
<td></td>
<td>Interview</td>
<td>Study in which the researcher directly interacts with a respondent or group to obtain information via discussion.</td>
</tr>
<tr>
<td></td>
<td>Case Study</td>
<td>Study in which the researcher undertakes an intensive study of a specific individual or context.</td>
</tr>
<tr>
<td>Specialized Studies</td>
<td>Triangulation</td>
<td>Study in which multiple methods are employed in an effort to compensate for weaknesses of each selected method.</td>
</tr>
<tr>
<td></td>
<td>Archival</td>
<td>Study based on information extracted from archival records.</td>
</tr>
<tr>
<td></td>
<td>Simulation</td>
<td>Study based on imitation of phenomena via a model of the phenomena.</td>
</tr>
<tr>
<td></td>
<td>Illustration</td>
<td>Study based on demonstration, “proof by example”, and other such foundational techniques</td>
</tr>
<tr>
<td>Viewpoint Publications</td>
<td>Tutorial</td>
<td>Publications that provide instruction on the use of a technique or tool in practice.</td>
</tr>
</tbody>
</table>
Experience: Publications emphasizing the personal experiences of the author regarding the application of a technique or tool in practice.

Opinion: Publications containing the author's opinion about what is good or wrong about something.

Announcement: Publications which announce events to the community.

2.4.3.2 Classification of Publications. Regarding research question three (RQ3): What type of evidence does the population of literature represent? This study finds there is empirical support regarding for the EBSE ecosystem and SLR process, but very limited empirical support for the EBP structural model and EBP process model. Within the areas most closely associated with practice – the structural and process models – the dominant form of publication is illustration.

As shown in Figure 4, overall a wide variety of publication types are found in the selected publications. Only the direct observation method is absent among the selected publications. The illustration type study is dominant with more than twice the number of publications than any other type. Secondary type research publications, which include SLR and SMS studies targeted within the ecosystem, are also well represented. Also, there are a considerable number of opinion and experience reports found among the publications. The least utilized types are simulation and archival research publications.

Examining the types of primary research, qualitative based publications (48 articles) are the lead form of primary research followed by quantitative publications (31 articles). However, the non-research type publications (91 articles) – which include tutorial, experience report, opinion, and announcement – outweigh qualitative and quantitative combined.

Inspecting the number of publications found in each of the major models of the EBSE ecosystem, the SLR/SMS protocols (173 publications) are clearly a focal point in the literature.
Publications directly addressing the structural model of practice and its associated EBSE process view are among the rarest publications found.
Figure 4. Publications by model and type
2.4.4 Categorization Results

In response to the second research question (RQ2): Does the population of the literature represent the conceptual models that describe EBSE and the EBSE ecosystem? The population of literature supports the use of the common EBP models to represent the EBSE paradigm and the EBSE ecosystem based on the identification of the models or their components within the population of EBSE literature.

In reply to the final research question (RQ4): What are the gaps between the conceptual models used as the topologies and the population of EBSE and EBSE ecosystem literature? The distinct lack of publications in several of the model components suggests a number of gaps exist. The remainder of this section examines each of the models and supporting evidence to facilitate an understanding of the gaps identified.

In order to assess the coverage of topics relating to EBP, the publications must be related to each of the known EBP models through a coding process. As there are no known systems for categorization of publications based on the models of EBP, a categorization system is created and utilized to identify the models and components present in each publication. The four models identified for examination include:

1. The structural model of EBP (Satterfield et al. 2009).
2. The process model of EBP (Sackett and Rosenberg 1995).
3. The systematic literature review process (Kitchenham and Charters 2007).
4. The EBP ecosystem model.

For each model, the discussion begins with an examination of the model and its components. Next, a list of category codes associated with the model is identified along with a description of
each category. Finally, the result of the application of the developed scheme to the selected publications is presented.

The aggregation of the conceptually developed categories from the individual models serves as the initial categorization system of the study. As all possible deviations from the models due to the domain of implementation cannot be foreseen, the system was subject to modification during the coding process in the analysis phase of the study. Before proceeding with the gap identification it is necessary to note additions to the categorization system based on the coding process. The qualitative analysis of the documents for coding purposes revealed the need of four additional codes. Table 5 provides a list of supplemental categories recovered and a description of the items addressed by each category.

Table 5. Supplemental categories revealed during coding

<table>
<thead>
<tr>
<th>Category</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teaching</td>
<td>The training of software engineers in the philosophy, tools, and techniques associated with the EBSE paradigm.</td>
</tr>
<tr>
<td>Issues</td>
<td>Reports of problems or gaps in the current instantiation of EBSE.</td>
</tr>
<tr>
<td>Steering</td>
<td>Suggested directions for the further development of EBSE.</td>
</tr>
<tr>
<td>Status</td>
<td>Reports regarding accomplishments in implementing EBSE or the general perceived acceptance of EBSE within the SE domain.</td>
</tr>
</tbody>
</table>

The category of Teaching is added to each of the models as a component in order to assess which, if any, of the models is being utilized in the training of new software engineers. The remaining three categories (Issues, Steering, and Status) are counted only in the EBSE ecosystem model in order to provide a more holistic picture of progress regarding the overall paradigm.

In coding each of the publications selected for the study, one or more categories are assigned to each. Thus a given publication may address several categories. The remainder of the analysis
examines the categories found for each of the models in conjunction with the publication types found for each category.

2.4.4.1 The Structural Model of EBP. Initial insights into EBP can be gathered from the examination of a structural model of EBP. Satterfield et al. (2009) provide a generalized structural model of EBP based on the examination and synthesis of the structural models that have emerged for the fields of medicine, nursing, psychology, social work, and public health (see Figure 5).

Through a structural lens one can see that the primary components consist of evidence from research; the client with associated characteristics, state, needs, values and preferences; and
resources, including the practitioner’s experience. These primary components are situated within a given environmental or organizational context which may influence their interaction and the decision-making process. Decision making is situated at the center of the model and provides the contextualization of evidence and its conversion into evidence-based practices (Satterfield et al. 2009).

Translating this model into the EBSE context, the following adaptations are made:

- **Stakeholder** replaces client – maintaining the existence of attributes.
- **Best available research evidence** is restated as **practical evidence** – this allows for alternative forms of evidence that may be preferred by practitioners (Shull 2007), and permits a differentiation between evidence in practice versus evidence in research.
- **Environment and organizational context** is restated as **practice environment** to permit differentiation among the environments of the four models being considered.
- **Resources** are restated as **practitioner** to refocus attention on the software engineer in the field in the same manner that the medical specific model focuses on the clinician (Satterfield et al. 2009).

One can anticipate that this structural model and each of its major components would be a topic of investigation or discussion within the literature. Therefore, each is taken as a category as described in Table 6 for inclusion in the categorization system.
Table 6. Structural Categories found in EBSE

<table>
<thead>
<tr>
<th>Category</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structural Model</td>
<td>The existence or overall form of the structural model of EBSE.</td>
</tr>
<tr>
<td>Practice Environment</td>
<td>Conditions or factors within the practice environment that influence or impact the EBSE paradigm and its implementation.</td>
</tr>
<tr>
<td>Practice Evidence</td>
<td>Discussions regarding what does, or does not, constitute evidence for the purpose of decision making in practice.</td>
</tr>
<tr>
<td>Stakeholders</td>
<td>Stakeholder characteristics, state, needs, values, or preferences that impact or influence decision making in practice.</td>
</tr>
<tr>
<td>Practitioner</td>
<td>The characteristics of practicing Software Engineers that impact decision making, along with tools or techniques utilized in support of EBSE.</td>
</tr>
<tr>
<td>Decision Making</td>
<td>Decision making in the context of practice and how it is influenced by the primary components found in practice.</td>
</tr>
</tbody>
</table>

The category of *Decision Making in Practice* is intended to capture the different overall concept of decision making and its interaction with the other components. The actual details of the process are examined separately in the next section.

As shown in Figure 6, publications addressing the structural model of EBSE in practice or its components primarily focus on the topic of evidence in practice. Absent from the literature are descriptions of the structural model as it applies to EBSE and the teaching of the model or its components. Only 7 of the possible 17 publication types were found with illustrative publications being the dominant type of publication. Additionally, no quantitative type publications were noted among the selected studies.
Figure 6. Structural model publications
2.4.4.2 The Process Model of EBP. The process undertaken in the practice of EBP (Sackett and Rosenberg 1995) includes five steps (see Figure 7). As a first step, the practitioner converts their information needs into answerable questions. Next, a search is undertaken to find the best available evidence with which to answer the questions posed. This is followed by a critical appraisal of the discovered evidence for both its validity and usefulness. The results of the appraisal are then applied in practice. Finally, a reflective process is undertaken to evaluate one's performance during the preceding steps. This fundamental process has subsequently been ported to the fields of nursing, psychology, social work, public health, education, management, and software engineering.

![Figure 7. Original Process View of Evidence-Based Practice.](image)

Implicit in this model is the idea that the evidence being sought and utilized is primarily external. Discussions among researchers have recognized that evidence both internal and
external to the practitioner plays a role in the overall process and that a synthesis across both forms of evidence is required for successful decision-making (Porzsolt et al. 2003, Sackett et al. 1996). Porzsolt et al. (2003) posit a modification of the process to reflect these ideals and assist in teaching of the process as illustrated in Figure 8.

While both the original and alternative views of the process of evidence-based practice provide parsimonious guidelines to practitioners regarding workflow, both are incomplete from a research standpoint. As previously noted, the original model does not specifically recognize that evidence may be either internal or external to the practitioner, nor does it specifically recognize the synthesis of evidence that must occur. The alternative model attempts to correct for this loss of fidelity, however neither model distinguishes the decision process in which a course of action is chosen based on the synthesis of evidence.
For the categorical analysis of this study, one can anticipate discussions of the overall process form and each of the steps identified within the process. Table 7 provides a list of the derived categories and a description of each.

Table 7. EBSE Process Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EBSE Process</td>
<td>The overall form or steps of the process to be utilized in the decision making process.</td>
</tr>
<tr>
<td>Question Formulation</td>
<td>The process - or factors influencing the process - of formulating answerable questions in practice.</td>
</tr>
<tr>
<td>Evidence Search</td>
<td>Sources or techniques for acquiring evidence.</td>
</tr>
<tr>
<td>Evidence Assessment</td>
<td>Criteria or techniques for assessing the strength or applicability of evidence in support of EBSE.</td>
</tr>
<tr>
<td>Applying Results</td>
<td>The selection or application of evidence in support of the decision making process.</td>
</tr>
<tr>
<td>Performance Evaluation</td>
<td>Factors or techniques utilized in the evaluation of current performance, or improvement of future performance within the EBSE paradigm.</td>
</tr>
</tbody>
</table>

As shown in Figure 9, publications addressing the EBSE process in practice are limited. While discussions are evenly spread across the components, there is a distinct lack of empirical work regarding the components. Once again, illustrative publications are the dominant type of publication.
Figure 9. EBSE process publications
2.4.4.3 The SLR Process Model. Literature reviews are a key component to advancing scientific research. The initial step of a new research endeavor is a review of previous work to properly ground the new research. The most common method of examining previous work is via literature review. A literature review may also have other goals, such as summarizing the current state of knowledge about an area as a service to the community.

Regardless of the goal of a literature review, these reviews have traditionally been performed in an *ad hoc* manner. A researcher can perform the review in the traditional fashion, that is, by conducting database searches and following references or the researcher can use a more systematic method. To provide structure to the literature review process, medical researchers defined the SLR process. An SLR is a formal, repeatable process by which researchers can identify, evaluate, and interpret the available research about a question or topic area.

As researchers conduct more empirical studies of various phenomena, SLRs will become an even more important tool for gathering and synthesizing the body of research about a topic (Petticrew and Roberts 2006). SLRs are important for researchers because they provide a snapshot of the current state of research and help identify gaps in need of more work. They are also important for practitioners who want to understand the best practices (similar to the way doctors use the results of medical SLRs to improve their practice of medicine). While the results of a literature review, especially an SLR, can be quite useful, and highly cited in the literature, they are quite time consuming.

The primary difference between an SLR and an *ad hoc* review is the level of advanced planning in an SLR. Prior to conducting the review, the researchers develop a protocol that documents: the research question(s) to guide the review, the search strategy (including specific databases and keywords), the criteria for choosing appropriate papers, a quality assessment
method for the papers, the specific information to be extracted from each paper, and a plan for synthesizing the information from the set of papers to draw a conclusion. By using this systematic process, researchers are much less likely to accidentally omit important papers from the literature review.

An additional noteworthy aspect of the SLR process is the formal use of multiple researchers as an additional guard against the introduction of bias into the results of the study. Agreement between the researchers can be quantified at multiple points in the process through the use of the Cohen Kappa statistic. While the use of multiple researchers is the preferred approach to performing an SLR, the use of an expert panel, or advisor in the case of Ph.D. students, permit a single researcher to execute the SLR process and achieve unbiased results.

The SLR process is not without its criticisms. It has been noted that it corresponds to a novice approach of searching for and synthesizing research (Hjørland 2011) based on the descriptions of experts and novices offered by Dreyfus and Dreyfus (2005). In addition, the SLR process is subject to the same pitfalls and limitations of any meta-analysis. This includes:

- Sampling bias due to a reliance on primarily empirical studies which can be combined statistically.
- Publication bias that results in the exclusion of non-significant results.
- Propagation of bias from poorly designed primary studies (King and He 2005, Miller 2000).
- Bias due to the combining of “studies with incommensurable research goals, measures, and procedures” (King and He 2005).
- Type I errors due to small sample sizes when ample studies are not available (King and He 2005, Miller 2000).
While the systematic review process has its roots in the medical field, it was ported to the domain of software engineering by Kitchenham (2004). This influential work was later updated (Kitchenham and Charters 2007) to further refine the process and incorporate lessons learned during the execution of the methodology. The process, as shown in Figure 10, consists of three primary phases: review planning, conducting the review, and documenting the review.

![Figure 10. High Level SLR Process Flow](image)

The goal of the planning phase is to produce a protocol that will be utilized as the execution plan when conducting the review. The protocol is a complete guide to conducting the review and should be validated through a review process to ensure it is complete and free of bias. The review process ideally consists of a review of the protocols components for efficacy and internal consistency by a panel of independent experts with recommendations from the panel incorporated into the protocol, followed by a pilot of the protocol to assist in discovering any deficiencies (Kitchenham and Charters 2007). After conducting the review, the data gathered during the process and a synthesis of that data, along with information about the execution of the process, are documented. If adjustments to the protocol are required during the conduction or documentation of the review, the researchers should return to the planning stage, adjust the protocol, and revalidate it before proceeding. While most of the irregularities are resolved during piloting of the protocol before a final, formal review at the end of the planning phase, this adjustment process provides a means of iterating over portions of, or the entire, SLR process.
Further decomposing the primary phases of the SLR process into process steps is depicted in Figure 11. Examination of the individual steps that comprise each of the phases reveals the systematic, unbiased, transparent nature of the process which makes it repeatable.

**Figure 11. Low Level SLR Process Flow**

### 2.4.4.3.1 SLR categorization: Planning phase.

During the planning phase, the research questions to be addressed are specified and the researcher defines a protocol that guides SLR execution (Kitchenham and Charters 2007). The goal of the protocol is to reduce researcher bias and provide a repeatable, transparent process for conducting the SLR. The protocol should contain, at a minimum, the following information (summarized from Kitchenham and Charters 2007):
• Background information providing the rationale for the study.
• All research questions to be addressed by the study.
• Strategy for the identification of primary sources of studies.
• Strategy for the identification of primary studies to be included in the study.
• Quality assessment criteria for identified studies.
• Data Extraction strategy.
• Procedure for data synthesis
• Project timeline.

Once assembled the protocol is reviewed, preferably by an independent panel, for completeness and concerns of validity. If at any point during the execution of the study the researchers must change the protocol, the panel should re-review the revised protocol (Kitchenham and Charters 2007).

2.4.4.3.2 SLR categorization: Conducting phase. When conducting the review, the researchers proceed through five general steps (Kitchenham and Charters 2007):

1) Identify relevant research by executing the defined search strategy.
2) Select primary studies by applying the inclusion and exclusion criteria.
3) Assess study quality using the quality assessment criteria.
4) Extract required data into data extraction forms.
5) Synthesize data to draw conclusions.

In order to identify relevant research, previously identified search terms are utilized to query multiple electronic databases and other sources such as library resources and grey literature. The results of the search process are then iteratively culled utilizing the inclusion and exclusion
criteria defined in the study protocol. The results are reviewed first based on titles, then by review of the abstracts, and finally a review of the full text. During each of the iterations, prospective studies are eliminated only when the researchers are confident that a given study has no bearing on the present work as defined by the inclusion and exclusion criteria of the study protocol. To reduce researcher bias in the process, researchers perform iterations independently and meet to review the results of the iteration and resolve any conflicts that arise.

Once the primary studies have been selected, a quality assessment of each study is completed by each of the researchers in accordance with the criteria set forth in the study protocol. The independent results are then compared for inter-rater agreement and reliability. The cumulative results may then be used to weight the results of the identified studies during the synthesis of extracted data.

Next, the data extraction phase is completed by each researcher independently and the results compared for inter-rater agreement and reliability. As in previous steps, any disagreements are discussed for resolution and agreement. The resulting data set then forms the basis for the synthesis process as defined by the study protocol.

2.4.4.3.3 SLR categorization: Documentation phase. Lastly, during the documentation phase the researchers use all of the information described in the protocol, along with the results of the execution of the protocol to document the review in some type of publication. This phase consists of the following steps:

1) Specify dissemination strategy
2) Write the review report
3) Evaluate the report
To identify categories within this methodology, the overall view as presented in Figure 11 is utilized. Each of the prescribed steps is extracted with one exception: The development and validation of the protocol are collapsed into a single category of **Protocol Specification**. This collapsing of categories is based on previous research (Carver et al. 2013) which indicates validation is often not utilized as described, and that alternative means for assembling a protocol are utilized by some researchers. Table 8 provides the list of categories, and descriptions of each, for the SLR Process.

Table 8. SLR Process Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SLR/SMS Process</td>
<td>The overall form or steps of the SLR process.</td>
</tr>
<tr>
<td>Motivation</td>
<td>Motivation or need for completing a SLR.</td>
</tr>
<tr>
<td>Commissioning</td>
<td>Resource Estimation and contract specification.</td>
</tr>
<tr>
<td>Research Questions</td>
<td>The formulation of research questions to be answered in an SLR.</td>
</tr>
<tr>
<td>Protocol Specification</td>
<td>The creation, modification, or validation of a SLR protocol.</td>
</tr>
<tr>
<td>Study Search</td>
<td>The tools or techniques utilized in the identical of primary studies.</td>
</tr>
<tr>
<td>Study Selection</td>
<td>The tools or techniques utilized in the selection of primary studies.</td>
</tr>
<tr>
<td>Study Assessment</td>
<td>The tools or techniques utilized in the assessment of study quality.</td>
</tr>
<tr>
<td>Data Extraction</td>
<td>The tools or techniques utilized in the quality assessment of studies.</td>
</tr>
<tr>
<td>Data Synthesis</td>
<td>The tools or techniques utilized in the synthesis of primary studies.</td>
</tr>
<tr>
<td>Dissemination</td>
<td>Dissemination outlets or the selection of an outlet.</td>
</tr>
<tr>
<td>Results Presentation</td>
<td>The presentation of SLR results.</td>
</tr>
<tr>
<td>Results Evaluation</td>
<td>The evaluation of SLR results.</td>
</tr>
<tr>
<td>Instance</td>
<td>An instantiation of the SLR Protocol.</td>
</tr>
</tbody>
</table>

Publications addressing the SLR/SMS protocol (see Figure 12) are the most prevalent of all the models. While once again the use of conceptual type publications is popular, most of the other forms of publications are also found in the selected studies. Only direct observation and simulation studies are absent from the 17 possible publication types.
It is also noted that the steps found in the conducting phase of the protocol have received considerable attention among the three phases. Within the conducting phase, the most work is found in the components regarding the search for studies and the extraction of data from studies that have been selected for inclusion. One additional point of interest is that publications regarding the teaching of the SLR/SMS protocol target both the undergraduate and graduate levels of education.
Figure 12. SLR/SMS protocol publications
2.4.4.4 The EBP ecosystem. The ecosystem view of EBP provides the overall flow of information between a given form of practice and its supporting research discipline (see Figure 13). Utilizing the conceptual model of EBP as a representation of practice, the best available evidence is drawn from research evidence; specifically SLRs that aggregate or synthesize the results of primary studies. The production of SLRs is a resource intensive process in terms of time and funding, which suggests the need for supporting infrastructure. In the healthcare domains, the aggregation and synthesis is primarily accomplished by a collaboration community such as the Cochrane Collaboration, which will be referred to as infrastructure in this study. Information and identified needs are fed back into the research process in the form of data and/or guidance for research. This forms a continuous information loop between research and practice allowing for a bidirectional transfer of current information and needs.

Figure 13. The EBP Ecosystem.
Based on this description of the EBSE ecosystem, one can anticipate discussions concerning each of the identified components. As the structural view has been examined previously (see Section 2.4.4.1), the additional components found in the research side of the loop provide additional analysis categories. Table 9 provides a list of the additional categories.

Table 9. EBSE Ecosystem Categories

<table>
<thead>
<tr>
<th>Category</th>
<th>Content Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paradigm</td>
<td>The EBSE paradigm or its institutionalization.</td>
</tr>
<tr>
<td>Primary Studies</td>
<td>Characteristics and use of primary studies for EBSE purposes.</td>
</tr>
<tr>
<td>Secondary Studies</td>
<td>Characteristics and use of secondary studies for EBSE purposes.</td>
</tr>
<tr>
<td>Primary Research</td>
<td>Characteristics of primary research impacting EBSE.</td>
</tr>
<tr>
<td>Secondary Research</td>
<td>Characteristics of secondary research impacting EBSE.</td>
</tr>
<tr>
<td>Technology Transfer</td>
<td>The transfer of EBSE results into practice.</td>
</tr>
<tr>
<td>Feedback</td>
<td>The communication of needs and status from practice to research.</td>
</tr>
<tr>
<td>Infrastructure</td>
<td>Resources specifically intended for the support of EBSE.</td>
</tr>
</tbody>
</table>

Publications addressing the overall EBSE paradigm and ecosystem model (see Figure 14) are also prevalent among the selected studies. Within this grouping, many of the components are represented with the only exceptions being the feedback of information from practice and the teaching of a high level of abstraction that is the big picture view of EBSE. Likewise, the types of publications show considerable variation with only the direct observation publication absent from the possible publication types. The most prevalent types of publications in this group are interviews, illustrative work, and opinion pieces.
Figure 14. EBSE paradigm and ecosystem publications
2.5 DISCUSSION

This study examines metadata – data about data – in order to ascertain the extent, nature and structure of the EBSE literature population. Libraries, and others interested in the curation of knowledge, have long utilized metadata to assess the degree of completeness of a collection, to structure collections for retrieval, and to identify and exploit the structural relations between or within collections of documents (Gilliland 2008). Additionally, metadata “provides a range of intellectual access points for an increasingly diverse range of users” (Gilliland 2008) and provides a means to organize and describe a collection. Thus in assessing the EBSE literature population, the use of metadata provides a concise means to ascertain the current state of the collection.

2.5.1 Identifiability of EBSE Research

While it is clear that many components of the models that are unique to the EBSE paradigm can be clearly identified given a sufficiently broad search and clear inclusion/exclusion criteria, examining the individual model results suggests that while discussions of the paradigm, infrastructure, issues, steering and status can clearly be identified, other components represent a fuzzy boundary where EBSE interfaces with general research in the SE domain and SE practice.

When investigating topics related to EBSE and its processes researcher must look beyond the EBSE literature – and even beyond the SE literature. Only then can they make efficient use of limited research resources and be confident in their research results. For example, Carver et al. (2013) note the continuing issues that researchers experience in assessing the quality of primary studies for the purposes of inclusion in SLRs. Despite this observation, there exists guidance regarding standards for both experimental (Jedlitschka and Pfahl 2005) and case study (Runeson et al. 2009) methodologies in the general SE literature.
Several of the publications found by this study address general research topics such as the standards for conducting or reporting case studies (Host and Runeson 2007, Runeson et al. 2009) and experiments (Goulao and Brito E Abreu 2007, Jedlitschka and Pfahl 2005), the development of study abstracts (Kitchenham et al. 2008, Budgen, Kitchenham, et al. 2008, Budgen et al. 2011), and importance of contextualizing the evidence produced in research (Petersen and Wohlin 2009, Dyba 2013). Additionally, many of the publications identified by this study regarding data synthesis for the SLR/SMS protocol are applicable to research in general or began in general research and have value in the SLR/SMS process. Similarly, most of the identified publications regarding the search for, and selection of, primary studies are useful beyond the EBSE paradigm in the construction of literature reviews in general. While this study has not identified any publications regarding literature search activities that are not motivated by the use of EBSE, given the inclusion/exclusion parameters of this study the existence of such publications cannot be ruled out.

Thus the population of literature that specifically targets EBSE and the EBSE ecosystem is uniquely identifiable, but the literature impacting components of the models of EBSE is not uniquely identifiable from the EBSE perspective. When investigating the various components, researchers must approach the topic of interest from an abstracted perspective in order to identify research beyond the EBSE paradigm that may impact the answers to the questions being explored.

The fuzzy boundary that exists between EBSE literature and general SE literature is also present within the EBSE literature among the primary models. In some cases, it is difficult to assign a publication to one primary model. For example, discussions of what constitutes
evidence impact the structural model, the process model, secondary study (SLR, SMS) protocols, and are a defining point in the paradigm. Such difficulties are a clear indication of interactions among the models and components, therefore a more holistic viewpoint is needed when searching for publications, selecting publications, and designing research investigations.

2.5.2 Models and Gaps

The population of literature supports the use of the common EBP models to represent the EBSE paradigm and the EBSE ecosystem as a continuing prototype for investigation and refinement. However, a distinct lack of publications in several of the model components suggests a number of gaps exist. Figure 15 illustrates the issue by placing the number of publications found by this study in parentheses with the associated component of the EBSE ecosystem. Bearing in mind that this is a visual summary of a decade of research, clearly there is a lack of attention to components outside the realm of research and to ways in which the information loop between research and practice can be closed.

![Figure 15. Publication Counts within the EBSE Paradigm](image)
As previously noted, the structural model – which is utilized in training new practitioners in other domains that have adopted the EBP paradigm – has limited support in the existent EBSE literature. All of the components of the model are found in the population of literature, but support is limited to a descriptive nature reinforcing conceptual reasoning. Also, the model itself as expressed in Figure 5 is completely absent. As the literature is incomplete, it is too early to determine if this absence is due to an oversight or evidence that a different model should be utilized. Regardless, based on the presence of its components the introduction of the model into the literature would serve to start the conversation concerning the appropriateness of the common structural model or adaptations need to best represent SE practice.

The introduction of a structural model may also assist in the investigation of teaching EBSE to new software engineers – another structural component vaguely represented in the literature population. To date the publications regarding the teaching of EBSE are focused primarily of the SLR/SMS protocols. While these protocols are meticulous in producing evidence grounded in empirical research, they are also not the parsimonious, actionable models needed for practicing software engineers. The time and resource required to support such protocols are often not available in day-to-day operations. A more parsimonious, actionable protocol is needed.

Similarly, the EBSE process model for decision making is clearly under-investigated in the existing literature population. However as the model is clearly described in many of the conceptual studies present in the literature, it is unclear as to whether the model has simply been adopted “as-is” or ignored by researchers. It is the opinion of the author that, at a minimum, the model should be contextualized to the SE domain in order to provide better direction and support to practice.
Another distinct gap in the existing literature is the absence of work regarding feedback from practice from an EBSE viewpoint. In the decade since the inception of the EBSE paradigm, considerable progress has been made by researchers in their efforts to amalgamate empirically based evidence in support of SE practice. This study suggests that nearly 300 SLRs are available to support decision making in the practice environment. However, the results of this study also suggest that work to provide feedback from practice into research is needed. While it must be conceded that research regarding this relationship to practice may exist and has been excluded due to the search or selection criteria of the study, one would anticipate the relating of such literature to the EBSE paradigm – especially given the applied nature of SE research. Software Engineers in the field, as well as managers and executives, can support the EBSE paradigm by not only providing that feedback, but also by assisting researchers in determining its form. Such efforts will improve the usefulness of SLR reports and provide insights for researchers regarding important problems in need of solutions.

Finally, as new studies continue to emerge in high interest areas, such as primary study searching and selection in the SLR/SMS protocol arenas, there is a need to synthesize the knowledge within the component and with knowledge of a more general nature. Indeed, beginning research in many of the components by first completing a systematic mapping and review of the topic – that is not limited to the EBSE paradigm, or even the domain of SE – may alleviate the need for foundational work.

2.5.3 Evidence Types and Maturity

While there appears to be empirical support for the utilization of the EBSE paradigm, that support is limited – especially within the models most closely associated with practice. This study found no quantitative based studies supporting either the structural model for practice, or
the EBSE process view that is typically associated with practice. The empirical work that exists to date in both models is primarily of a qualitative nature. However, the maturity of the paradigm in the SE domain must be considered. Given that descriptive research must be developed before normative or predictive theory can be build, this finding is consistent with the development of new theory within a domain. Similarly, the continuing emergence of studies to refine the SLR/SMS protocols and the infrastructure, tools and techniques to support EBSE and the EBSE ecosystem are reflective of continuing refinement of descriptive research and the initiation of normative work.

2.5.4 The Essence of EBSE

In one of the founding publications of EBSE, Kitchenham et al. (2004) states the goal of the paradigm should be

“to provide the means by which current best evidence from research can be integrated with practical experience and human values in the decision making process regarding the development and maintenance of software.”

After a decade of effort, it would appear the community has fallen short of this goal. While the production of best evidence from research – in the form of SLRs – has steadily increased, the lack of publications investigating how software engineers actually practice, or should practice, their profession indicates a loss of focus. The essence of EBSE appears to be the production of evidence for input into the decision making process. The disheartening part of the situation is that many of the researchers engaged in the production of evidence based on research are also the professors training the next generation of software engineers – a prime opportunity to influence how tomorrow’s engineers view, and use, evidence. While there is discussion of teaching EBSE at the undergraduate level (Jorgensen et al. 2005) it is based on the assumption that everything needed to practice EBSE (e.g. scientific method, statistics, evidence evaluation, etc.) must be
taught in a single course. Furthermore, there is an underlying assumption of the skills that will
be needed to practice EBSE. These assumptions need to be investigated to validate or refute
what is required. In addition, the possibility of incorporating the facets of EBSE throughout the
curriculum, instead of just one class, needs to be investigated.

2.5.5 Implications for Research

For researchers working in the EBSE arena, the results of this study provide a number of
insights. First, this study identifies a body of scientific literature regarding the EBSE paradigm
that can be built on and extended. The 157 publications included in the analysis of this study
provide a basis for further development of a majority of the models and components that makeup
the EBSE ecosystem.

Viewed as the existing population of EBSE literature, the collection of publications also
provides a roadmap of the investigative work needed. As previously outlined, there are a
number of gaps in the literature with regard to the common EBP models. There is also no
overwhelming evidence supporting most of the models. In most cases this is due to the lack of
confirming studies, or overlapping studies typically used to identify or extend the boundary of
theories. What is needed is additional descriptive work to fully identify all the components and
their relationships so that the process of normative and prescriptive theory building can begin.

Finally, the arena of SE practice can no longer be viewed simply as a target audience in need
of enlightenment. Practitioners need to be seen as partners in progress, for both research
endeavors and advancement of the profession. In order to build such a relationship, the
communication loop between research and practice must be established and reinforced to the
mutual benefit of both parties. One of the first steps in this process for researchers is to better
understand the goals and needs of our fellow practitioners.
Finally, there are gaps remaining in the adaptation of the EBP models for use in EBSE. This study provides an initial roadmap regarding the components that are absent from the literature and additionally which components require further research due to limited supporting evidence. Specifically, there is a distinct lack of research regarding the practice portions of the EBSE paradigm.

2.5.6 Lessons Learned Concerning the SMS Protocol

The execution of this mapping study also provides insight concerning the SMS protocol with regard to barriers to the process, resources required, and benefits of the protocol. These experiences contribute to the general understand of the protocol from each perspective.

One of the barriers encountered during this study is the lack of classification and categorization systems to assist in the analysis stage of the protocol. In each case, a separate line of inquiry is required in order to first derive the system before use. This involves a literature search for relevant input materials, synthesis of the materials, and an iterative process of piloting the system and refining it. This process contributes significantly to the time required to complete the study.

Another resource intensive portion of the SMS protocol is the selection of studies for inclusion. This process requires the acquisition, storage, and examination of thousands of documents in order to arrive at the relevant set of publications for the study. Aside from the computing resources and subscription access required for this stage, researchers need to allow for the time required to complete the reading required in the execution. The required reading is not limited to the final set of selected papers, but includes the reading of all the title, and many of the
abstracts, that are retrieved in the search stage. Selection of studies is a non-trivial step on the critical path for completing a mapping study and must be planned accordingly.

One of the benefits of the resource expenditure involved in a mapping study is the holistic view that it provides of a domain. It allows the researchers to understand what areas are well supported, the area’s in need of more support, and the area’s in which opportunities for novel approaches to bridging a gap exist. Mapping studies provide a means to devise and plan a portfolio of studies in order to advance a domain of interest.

In addition, this researcher has found that such a study allows one to develop a more intimate knowledge of an area. While this obviously includes the existent knowledge in a domain, it also includes the expectations and viewpoints held by other researchers in the domain. This understanding assists in the successful design of acceptable studies and the conveyance of study results back to the community.

2.5.7 Limitations

A limitation of this study that must be acknowledged is that it is not based on a full semantic interpretation of each of the selected publication’s contents. While a synthesis of the full semantic content of all the publications would most certainly reveal nuances of coverage in each of the models, for the purposes of this study it is believed that the course-grained view offered by simple categorization provides sufficient detail to map EBSE in its current status.

The study selection and data extraction processes are inherently prone to bias due to individual interpretations. This study attempts to guard against bias through the use of multiple researchers and transparent processes. At each step in the SMS, inter-rater agreement among the researchers is assessed to minimize the risk of bias. Continuation to the next step occurs only after agreement has been reach.
The ability to replicate the search results is problematic due to the nuances associated with the different search engines. Preliminary testing indicates that the use of advanced searches instead of basic searches for the same search string can produce different results. Additionally, the order of terms/execution appears to impact the results returned, as well as the ranking of the results.

The constraints imposed by the scoping of the study also present limitations. Publications selected for inclusion are limited to the English language. It is possible that additional publications exist in languages other than English. However, barring a substantial body of literature in additional languages, it is believed that the results of this are unlikely to change significantly. Additionally, as the major conferences and journal outlets known to publish EBSE research utilize English as a standard language, the risk of a significant publication, or body of publications, being overlooked is believed to be minimal.

Similarly, the terms utilized in the search for publications and the strict adherence to the inclusion/exclusion criteria limit the publications considered in the analysis for this study. The risk that significant publications have been overlooked due to these factors is believed to be minimized as the manual search of venues for establishment of the quasi-gold standard and initial piloting of the protocol resulted in only minor changes to the research plan. Furthermore, the identification of publications representing a fuzzy boundary in the publication set indicates the criteria are sufficiently broad to facilitate capture of all publications of interest.

Indexing bias – where journals only index select articles electronically – is also problematic. Similarly, the inclusion/exclusion criteria may have restricted the recovery of relevant literature. However, based on the use of a quasi-gold standard founded on a manual search of two conferences and known papers this treat is minimal.
Likewise, publication bias – the publication of primarily positive results – also poses a possible threat to the finding of this study. Unpublished research reflecting negatively on any of the model components could either bolster the conclusions of the study or refute the interpretation of the results.

### 2.6 Conclusion and Future Work

In the decade since the introduction of the EBSE there has been considerable research and discussion concerning it in the SE literature. A majority of the publications focus on the SLR protocol or represent instances of its use to aggregate areas of research. Indeed, even the research regarding the SLR protocol itself has been the subject of several SLR publications (Da Silva et al. 2010, da Silva et al. 2011, Hamad and Salim 2014, Kitchenham et al. 2009, Kitchenham et al. 2010, Kitchenham and Brereton 2013)

However, there is still considerable work to be done to institutionalize the EBSE paradigm. Further research is needed to establish a structural model of SE practice and its associated decision making process to facilitate work into teaching EBSE at the undergraduate level. Also, the relationship between research and practice must be further explored to enhance feedback from practice and the transfer/dissemination of research into practice.

At present, there is little integration between EBSE research and practice. This study empirically confirms that the understanding of SE practice and closure of the information loop between SE research and SE practice is clear suffering from neglect within the EBSE paradigm. However, this study also demonstrates that the existing EBP models serve as an initial foundation for future research regarding EBSE. This study also provides a roadmap to begin such investigations.
Finally, the classification and categorization systems developed for use in this study provide an additional step in the evolution and understanding of metadata regarding academic literature. The further development of such systems assist in accessing the evidence for, or against, a theory or practice: Allowing better decision making for both the researcher and practitioner.

2.7 REFERENCES


Kitchenham, Barbara, and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University.


CHAPTER 3

Identifying and Ranking Barriers to the Systematic Literature Review Process

3.1 INTRODUCTION

Systematic literature review (SLR) is a formal, repeatable process by which a researcher can identify, evaluate, and interpret available research about a question or topic area. SLRs are of critical importance in software engineering, because they allow researchers to bring together disparate evidence to understand the effects of various SE tools, techniques, and methods. Yet, despite the importance of SLRs to the maturation of SE research — as well as its timely adoption by industry — conducting an SLR is a difficult, time-consuming, and largely manual process.

Researchers are in need of tools that reduce the effort required to conduct SLRs. Such tools would encourage researchers to produce more SLRs. Fortunately, tool support for conducting SLRs is an active area of research. According to a recent mapping study (Marshall and Brereton 2013), eleven such tools are described in the SE literature. Most of these tools (7 of 11) target the selection of studies to be include in the review (Felizardo et al. 2012, Felizardo, Salleh, et al. 2011, Sun et al. 2012, Tomassetti et al. 2011), the extraction of data (Felizardo et al. 2010, Sun et al. 2012, Torres et al. 2012), or the synthesis of data (Cruzes et al. 2007a, b, Felizardo et al. 2010, Felizardo, Riaz, et al. 2011). Three of the tools (Bowes et al. 2012, Fernández-Sáez et al. 2010, Hernandes et al. 2012) target the entire SLR process, and the remaining tool (Ghafari et al. 2012) targets the identification of candidate studies.

2 Portions of this essay appear in the proceedings of the 18th International Conference on Evaluation and Assessment in Software Engineering. The work was sponsored by the National Science Foundation under Grant No. 1305395.
Unfortunately, most of the tools identified by the mapping study (Marshall and Brereton 2013) are in early stages of development, and only two of them (Hernandes et al. 2012, Torres et al. 2012) have been evaluated independently (Marshall and Brereton 2013). Moreover, while the goal of each tool is to address one or more barriers to the widespread adoption of the SLR process, the tools do not necessarily address the most important barriers. The goal in this work is to identify the barriers deemed most important by the members of the SE research community who conduct SLRs. These barriers, in turn, can guide the development of additional SLR-support tools.

Previous work (described in more detail in Section 3.2.2) (Carver et al. 2013) identified barriers that hinder the SLR process. Based on the analysis of data from three sources: (1) the authors’ own experiences conducting SLRs, (2) a graduate-level course on SLRs, and (3) and a survey of authors of SLRs published in SE venues, the aspects of the SLR process that were most difficult, most time-consuming, and most in need of tool support are identified. Additionally, the results of the analysis suggest the need for expert review in all phases of the SLR process, as well as tool support for: collaboration throughout the SLR process, federated search of digital libraries, and maintenance, reuse, and evolution of SLR data. These results motivate further investigation of barriers to the SLR process.

To independently identify additional barriers and to rank the identified barriers by priority, a community workshop was organized. This paper describes the methodology and results of that workshop, which was held during the 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM).
3.1.1 Research Objectives

The primary goal of the research is to identify the barriers deemed most important by the members of the SE research community who conduct SLRs. Consistent with this goal, the objectives of this study are to:

1) Elicit information from participants regarding:
   - Barriers to completing an SLR that stem from a lack of adequate techniques or tooling;
   - Major characteristics and features that should be provided by SLR-support tools; and
   - Strategies for SLR community interaction toward addressing these issues.

2) Prioritize the identified barriers based on participate needs.

3.1.2 Research Questions

In support of the outlined objectives, the research questions addressed by this study are:

RQ1: What additional barriers, beyond the previously identified process barriers, exist in completing SLRs?

RQ2: What additional characteristics or features should be provided by SLR-support tools?

RQ3: What priority should be given to each of the known/identified barriers, characteristics, and features?

The remainder of the chapter is organized as follows. Section 3.2 provides an overview of the SLR process and a summary of previous work. Section 3.3 details the workshop methodology. The analysis of data generated during the workshop is described in Section 3.4. Section 3.5 reports the workshop results, and Section 3.6 discusses the implications and limitations of the results. Finally, Section 3.7 concludes the paper and describes future work.
3.2 BACKGROUND

To facilitate understanding of the work presented in the remainder of this study, this section provides the requisite background regarding the SLR process (Section 3.2.1). Following the explication of the SLR process, the results of previous work – which motivated the present study – are summarized (Section 3.2.2).

3.2.1 SLR Process

Literature reviews are a key component to advancing scientific research. The initial step of a new research endeavor is a review of previous work to properly ground the new research. The most common method of examining previous work is via literature review. A literature review may also have other goals, such as summarizing the current state of knowledge about an area as a service to the community.

Regardless of the goal, literature reviews have traditionally been performed in an ad hoc manner. A researcher can perform the review in the traditional fashion, that is, by conducting database searches and following references or the researcher can use a more systematic method. To provide structure to the literature review process, medical researchers defined the SLR process. An SLR is a formal, repeatable process by which researchers can identify, evaluate, and interpret the available research about a question or topic area.

As researchers conduct more empirical studies of various phenomena, SLRs will become an even more important tool for gathering and synthesizing the body of research about a topic (Petticrew and Roberts 2006). SLRs are important for researchers because they provide a snapshot of the current state of research and help identify gaps in need of more work. They are also important for practitioners who want to understand the best practices (similar to the way doctors use the results of medical SLRs to improve their practice of medicine). While the results
of a literature review, especially an SLR, can be quite useful, and highly cited in the literature, they are quite time consuming and resource intensive.

The primary difference between an SLR and an *ad hoc* review is the level of advanced planning in an SLR. Prior to conducting the review, the researchers develop a protocol that documents: the research question(s) to guide the review, the search strategy (including specific databases and keywords), the criteria for choosing appropriate papers, a quality assessment method for the papers, the specific information to be extracted from each paper, and a plan for synthesizing the information from the set of papers to draw a conclusion. By using this systematic process, researchers are much less likely to accidentally omit important papers from the literature review.

An additional noteworthy aspect of the SLR process is the formal use of multiple researchers as an additional guard against the introduction of bias into the results of the study. Agreement between the researchers can be quantified at multiple points in the process through the use of the Cohen Kappa statistic. While the use of multiple researchers is the preferred approach to performing an SLR, the judicious use of an expert review panel, or advisor in the case of Ph.D. students, permit a single researcher to execute the SLR process and achieve unbiased results.

The SLR process is not without its criticisms. It has been noted that it corresponds to a novice approach of searching for and synthesizing research (Hjørland 2011) based on the descriptions of experts and novices offered by Dreyfus and Dreyfus (2005). In addition, the SLR process is subject to the same pitfalls and limitations of any meta-analysis. This includes:

- Sampling bias due to a reliance on primarily empirical studies which can be combined statistically.
- Publication bias that results in the exclusion of non-significant results.
- Propagation of bias from poorly designed primary studies (King and He 2005, Miller 2000).
- Bias due to the combining of “studies with incommensurable research goals, measures, and procedures” (King and He 2005).
- Type I errors due to small sample sizes when ample studies are not available (King and He 2005, Miller 2000).

While the systematic review process has its roots in the medical field, it was ported to the domain of software engineering by Kitchenham (2004). This influential work was later updated (Kitchenham and Charters 2007) to further refine the process and incorporate lessons learned during the execution of the methodology. The process, as shown in Figure 10, consists of three primary stages: review planning, review execution, and review documentation.

The goal of the planning stage is to produce a protocol that will be utilized as the plan for executing the review. The protocol is a complete guide to executing the review and should be validated through a review process to ensure it is complete and free of bias. The review process ideally consists of a review of the protocols components for efficacy and internal consistency by a panel of independent experts with recommendations from the panel incorporated into the protocol, followed by a pilot of the protocol to assist in discovering any deficiencies (Kitchenham and Charters 2007). After executing the review, the data gathered during the
process and a synthesis of that data, along with information about the execution of the process, are documented. If adjustments to the protocol are required during the execution or documentation of the review, the researchers should return to the planning stage, adjust the protocol, and revalidate it before proceeding. While most of the irregularities are resolved during piloting of the protocol before a final, formal review at the end of the planning stage, this adjustment process provides a means of iterating over portions of, or the entire, SLR process.

Further decomposing the primary stages of the SLR process into process steps is depicted in Figure 11. Examination of the individual steps that comprise each of the stages reveals the systematic, unbiased, transparent nature of the process which makes it repeatable.
3.2.1.1 SLR: Planning stage. During the planning stage, the research questions to be addressed are specified and the researcher defines a protocol that guides SLR execution (Kitchenham and Charters 2007). The goal of the protocol is to reduce researcher bias and provide a repeatable, transparent process for conducting the SLR. The protocol should contain, at a minimum, the following information (summarized from Kitchenham and Charters 2007):

[P1] Motivation for conducting the SLR and background information

[P2] All research questions to be addressed by the study

[P3] Search strategy (including the target databases and search strings)

[P4] Strategy for identification of primary studies (i.e., the inclusion and exclusion criteria)

[P5] Quality assessment criteria for identified studies

[P6] Data Extraction strategy and accompanying form

[P7] Procedure for data synthesis

[P8] Project timeline

Once assembled the protocol is reviewed, preferably by an independent panel, for completeness and concerns of validity. If at any point during the execution of the study the researchers must change the protocol, the panel should re-review the revised protocol (Kitchenham and Charters 2007).

3.2.1.2 SLR: Execution stage. When executing the review, the researchers proceed through five general steps (Kitchenham and Charters 2007):

[E1] Identify relevant research utilizing the search strategy (P3)

[E2] Select primary studies by iteratively applying the inclusion and exclusion criteria (P4) to the title, abstract, and full text of each candidate paper
[E3] Evaluate the selected studies, using the quality assessment criteria (P5) to determine the reliability and importance of results

[E4] Extract the required data using the data extraction form (P6)

[E5] Synthesize the extracted data to draw conclusions

In order to identify relevant research, previously identified search terms are utilized to query multiple electronic databases and other sources such as library resources and grey literature. The results of the search process are then iteratively culled utilizing the inclusion and exclusion criteria defined in the study protocol. The results are reviewed first based on titles, then by review of the abstracts, and finally a review of the full text. During each of the iterations, prospective studies are eliminated only when the researchers are confident that a given study has no bearing on the present work as defined by the inclusion and exclusion criteria of the study protocol. To reduce researcher bias in the process, researchers perform iterations independently and meet to review the results of the iteration and resolve any conflicts that arise.

Once the primary studies have been selected, a quality assessment of each study is completed by each of the researchers in accordance with the criteria set forth in the study protocol. The independent results are then compared for inter-rater agreement and reliability. The cumulative results may then be used to weight the results of the identified studies during the synthesis of extracted data.

Next, the data extraction phase is completed by each researcher independently and the results compared for inter-rater agreement and reliability. As in previous steps, any disagreements are discussed for resolution and agreement. The resulting data set then forms the basis for the synthesis process as defined by the study protocol.
3.2.1.3 **SLR: Documentation stage.** Lastly, during the documentation stage the researchers use all of the information described in the protocol, along with the results of the execution of the protocol to document the review in some type of publication. This phase consists of the following steps:

[D1] Specify dissemination strategy

[D2] Write the review report

[D3] Evaluate the report

3.2.1.4 **SLR: Summary.** SLRs are useful for drawing conclusions about a phenomenon, based upon the current state of published evidence. Medical researchers, practitioners, and policy makers have long relied on SLRs, because they integrate up-to-date, reliable, and critical information that support important decisions. Practitioners benefit in a similar manner from the rigorous collection, filtering, and summarization of the latest developments in the understanding of both practice and artifacts. For researchers, SLRs are of critical importance because they allow researchers to bring together disparate evidence to understand the effects of various tools, techniques and methods: A key component in the maturation of research.

In contrast to traditional meta-analysis reviews, SLR’s make explicit the underlying assumptions of a review through the documenting of search, study selection, data extraction, and synthesis procedures in both the protocol and the final report of a review. This results in increased transparency and replicability. Indeed, a SLR protocol can be viewed as the “lab package” equivalent of experimental methodology.

Any number of barriers may be associated with each step or item enumerated previously. Previous work has identified a number of potential barriers to each of the stages associated with this process. These barriers served as inputs to the workshop planning.
3.2.2 Summary of Previous Work

Previous work (Carver et al. 2013) identified barriers to the SLR process by analyzing data gathered from (1) local experiences conducting SLRs, (2) a graduate-level SLR course and (3) a survey of authors of SLRs published in SE venues. In the graduate-level SLR course, each student defined and conducted an SLR then produced a report describing their experiences, including any specific difficulties encountered during the process. For the SLR author survey, a survey was sent to all known authors of SLRs published in SE venues. The survey asked the authors to describe: the SLR process utilized, difficulties encountered in the process, activities that consumed the most time, and the aspects of the SLR process most in need of tool support.

The survey results contained both quantitative and qualitative data. For the analysis of the qualitative data, a grounded theory approach (Corbin and Strauss 2008) is utilized beginning with open coding – to identify the important content in the data – and continuing with axial coding – to identify themes in the responses. Throughout the analysis, multiple researchers performed the coding independently (to reduce the potential for bias). The analysts compared their results and resolved any discrepancies. Based on the analysis, three important sets of results are identified that motivated the community workshop.

The first set of major results concerns the overall SLR process. The survey respondents identified aspects of the SLR process that they found to be most difficult (extracting data, assessing quality and selecting papers), most time-consuming (extracting data, selecting papers, and searching databases) and most in need of tool support (searching databases, selecting papers and extracting data). The overlap among the identified problems suggested that the community workshop participants were likely to identify barriers associated with particular aspects of the SLR process.
The second set of major results concerns the **SLR protocol**. About 1/3 of the survey respondents stated that their protocol deviated from the original definition provided by Kitchenham (2004). Only about 1/2 of the survey respondents had their protocol reviewed by external experts (as recommended by Kitchenham). Of those that had their protocols reviewed, nearly all incorporated changes based on the expert feedback, indicating that they found value in the expert review. Finally, the students in the graduate-level SLR course reported that discussions with classmates and expert review by the professor were both important to the success and quality of their SLR.

The third set of major results concerns **specific SLR protocol items**. Mirroring the results described earlier in this paper, *searching multiple databases* is a key difficulty of conducting an SLR. Difficulties in searching multiples databases are often due to the need to reformulate the search string to conform to the syntax accepted by a particular database. Further, Boolean logic and advanced search functionalities behave differently across commonly used databases. It was also noted that approximately 1/3 of the survey respondents did not perform quality assessment on their primary studies.

Key findings of the previous work include the identification of the following barriers to the current SLR process:

- Lack of expert review in all stages of the SLR process
- Lack of tool support for:
  - Identifying relevant studies in multiple databases
  - Collaborating on SLRs
  - Storing extracted data for update and reuse
  - Evolving SLRs as new knowledge emerges

These findings seeded the workshop discussions.
3.3 METHODOLOGY

The primary goal of the study is to gather community input on the barriers in the SLR process. Thus a venue which maximizes the potential number of researchers engaged in SLR research being collocated was sought. A two-hour workshop was conducted during the 2013 Empirical Software Engineering International Week (ESEIW ’13), which includes: the 2013 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM ’13), the International Software Engineering Research Network (ISREN) meeting, and the International Conference on Predictive Models in Software Engineering (PROMISE). This venue was chosen due to the likelihood that many SLR authors would already be attending ESEIW’13 events and could participate with little additional expense or effort. Consistent with the goal of the workshop, participants discussed:

- Barriers to completing a SLR that stem from a lack of adequate techniques or tooling
- Major characteristics and features that should be provided by SLR-support tools
- Strategies for SLR community interaction toward addressing these issues

The author and two additional researchers served as workshop facilitators.

3.3.1 Participants

To ensure a lively and substantial discussion, invitations were extended to all authors or co-authors of an SLR published in an SE venue. Due to other events being held concurrently during ESEIW’13, some interested authors were not able to participate in the workshop. The 13 workshop participants were all active researchers who conduct SLRs as part of their work. The participant population included 11 faculty members and 2 doctoral students. Both doctoral students and 10 of the faculty members were from the computer science or software engineering discipline. The other faculty member was a Ph.D./M.D. researcher at a medical school, who was
invited specifically because of his expertise in the use of SLRs in the medical field. The group included representatives from North America (4), South America (3), and four European nations (6). All participants had conducted one or more SLRs and were interested in improving the SLR process by upgrading the available support-infrastructure.

3.3.2 Data Acquisition

Data collection is based on the Nominal Group Technique (Delbecq et al. 1975) for data collection during the workshop because of its usefulness in the type of context present in the workshop (Lethbridge et al. 2005), namely:

- Most participants are new to one another (i.e. do not generally work with one another).
- Some participants may be more vocal than others.
- We anticipated that some group members think better in silence.
- We were not aware a priori of particular issues that could have been controversial or generated heated interactions.

3.3.2.1 Introduction and explanation phase. Following introductions, the group facilitators provided a brief overview of the SLR protocol and process stages, following the description in Section 3.2.1, to ensure that everyone was using the same terminology and conceptual model. The participants were informed that the goal of the workshop was to gather feedback to refine the list of existing community needs identified in previous work and to enumerate any additional barriers to conducting an SLR that were not previously identified. After a brief period for questions to clarify logistics/procedure, the workshop began in earnest.

3.3.2.2 Idea generation phase. The first step was a type of brainstorming in which the participants were encouraged to write down as many ideas as possible related to the topic at
hand. In this case the main topic was barriers to conducting an SLR. Participants performed this task individually and in silence to encourage participation from all attendees and to prevent some of the more vocal participants from dominating the discussion. For each idea identified by a participant, he or she created a Post-it note that contained a name for the idea, along with a brief description of it. The idea generation phase lasted 10 minutes.

3.3.2.3 Idea sharing phase. Based on the potential barriers identified in Section 0, an initial set of SLR discussion categories was created to organize the discussion. An easel was placed in the front of the room for each discussion category. The initial set of discussion categories included:

- Motivation for Conducting SLRs
- Research Question Formation
- Search and Selection Stages
- Quality Assessment Stage
- Data Extraction Stage
- Analysis and Presentation Stages

The group focused on each discussion category, one at a time. In turn, each participant read aloud one of the ideas he or she had identified for that discussion category and then affixed the corresponding Post-it note to the appropriate easel. This process continued until all participants had posted all of their identified ideas related to that discussion category. The procedure was then repeated for each of the remaining discussion categories. During this phase, if a participant determined that their idea was related to multiple discussion categories, the participant duplicated the Post-it note and affixed it to all of the relevant easels. Additionally, participants
were allowed to generate new ideas during this phase (i.e., ideas not recorded during the initial brainstorming) and to affix them to the appropriate discussion categories.

Figure 18 provides an example of one of the SLR discussion categories, at the end of this phase.

3.3.2.4 Group discussion phase. The first activity in this phase was Concurrent Subgroup Deliberation, in which participants randomly divided into subgroups of three to discuss the ideas identified in the previous phase. Each subgroup focused on one of the six discussion categories from the Idea Sharing Phase. The groups completed the four largest discussion categories first. Two groups that completed their discussion categories early were asked to assist with the remaining two discussion categories. The goal of the Concurrent Subgroup Deliberation was for
the participants to consolidate and organize the ideas on the Post-it notes into a set of overriding themes, thereafter referred to as barriers. During the group discussions, any ideas that a group thought belonged with a different discussion category could be offered to the group working with that discussion category for inclusion in their analysis. In this way, ideas could be moved among the groups for optimal placement within the discussion categories. This activity continued for 30 minutes, during which time the facilitators were available to answer questions or provide clarification. Figure 19 shows some of the activity during this phase.

The second activity in this phase was *Subgroup Presentation and Group Discussion*, in which a presenter from each group described the results of the subgroup analysis. After the presentations, we conducted a facilitated group discussion in which the participants discussed the relationships among the different barriers within groupings, within discussion categories, and across discussion categories.

Figure 19. Group Discussion Phase
3.3.2.5 Voting and ranking phase. Finally, to get some idea of the general priority of each of the identified barriers, the participants voted. Each participant was given 12 paste-on dots – 12 represented 33% of the total number of ideas to be voted on. Participants were asked to consider all of the identified barriers (from all discussion categories) and to determine which ones they thought most important (i.e., in need of being addressed with the highest priority). Because the number of votes received by a barrier is an indication of its relative priority, the participants could allocate their 12 dots in any manner they chose. For example, if a participant identified one barrier as the most important and the only one worth considering, he or she could place allocate all 12 dots to that barrier. Conversely, if another participant considered 12 barriers to be of equal importance, he or she could allocate one dot to each of those 12 barriers. To avoid confounding the results across fields of expertise, the medical school researcher did not participate in the voting. Figure 20 provides an example of one of the discussion groups after the voting phase.

Figure 20. Voting and Ranking Phase
3.4 Analysis

The analysis of data generated during the workshop is based primarily in the counting and tabulation of votes cast. The tabulation provides the priority rankings needed to answer the third research question (RQ3). The composite barriers identified are utilized directly in a direct comparison with prior results as outlined in section 3.2.2 to isolate new barriers and features required. The results of this comparison are then utilized to derive answers for the first and second research questions (RQ1 & RQ2).

The ideas generated during the workshop are examined utilizing qualitative techniques to uncover any latent themes. Responses are first analyzed with open coding, providing an initial level of conceptual extraction. Coding is the process of breaking down the data (Corbin and Strauss 2008) for the purpose of discovering theoretical concepts, their properties, and the relationships among concepts. The data may be examined at many levels; the analyst may examine the data elements of words, phrases, sentences, paragraphs, or entire documents (Corbin and Strauss 2008). In addition, a researcher may move among the levels to examine the data from multiple perspectives. Coding entails the examination of a data element and attaching a meaningful conceptual label to it (Corbin and Strauss 2008). Coding often takes place in the form of margin notes. However, coding could also be performed on note cards, or in the form of database elements. A critical element is that coding should not be a mere restatement of the data. Theoretical elements must be at a higher conceptual level than the data from which they are derived.

Following the initial open coding process, axial coding was performed. The axial coding process is similar to open coding, but focuses on making connections between the elements identified in the open coding process. As axial coding concludes, a tentative core variable that
explains the behavior in question is selected. Utilizing the core variable as a guide, researchers then recode data to provide a clearer conceptual model of the underlying phenomenon.

3.5 Results

This section reports the results from each phase of the community workshop.

3.5.1 Idea Generation and Idea Sharing Phases

During the Idea Generation phase, the workshop participants individually recorded 82 unique ideas concerning barriers to the SLR process. Subsequently, during the Idea Sharing phase, the participants increased this number as they posted these ideas to the discussion categories. First, the participants determined that five of the ideas belonged to two discussion categories. They created a duplicate Post-it note for each of these ideas and affixed the Post-it to the discussion category representing the appropriate SLR stage. Second, as the participants were socializing their list of ideas, the group identified 13 additional ideas. They recorded each of these ideas on a Post-it note and affixed the idea to the discussion category representing the appropriate SLR stage. As a result of these two phases, the participants produced a total of 100 ideas that were carried forward to the next phase.

There was one other important result of the Idea Sharing Phase. As the participants were affixing ideas to appropriate discussion categories, they became aware that an important discussion category was missing. It was clear that some of the ideas were related to the underlying epistemology of the SLR process rather than to any one SLR stage. Based on this observation, we created a new discussion category for Overall Protocol along with its own easel upon which participants could affix ideas.
3.5.2 Group Discussion Phase

Prior to the beginning of this phase, the participants determined that they would combine the Motivation for Conducting SLRs and Research Question Formation discussion categories into one discussion category for the remainder of the analysis process. They made this decision due to their belief that the underlying themes and issues expressed in the individual Post-it notes shared sufficient commonality to warrant combining. Therefore, we combined the Post-it notes from the two phases onto one easel for analysis. As a result of the addition of the Overall Protocol category and this combination, there were still six discussion categories.

During the Concurrent Subgroup Deliberation portion of this phase, we noted three instances where groups transferred a Post-it note describing an idea to a different discussion category. In each case, the group that was originally assigned the idea reached a consensus that the idea was not a good fit for the discussion category to which it was initially assigned and that it should be reassigned to a different discussion category. A member of the group to which the idea was initially assigned transported the corresponding Post-it note to the new group and presented the rationale for reassignment. In all cases, the receiving group accepted the idea and incorporated it into their deliberation.

As a result of this phase, the groups consolidated the 100 barriers into 37 composite barriers that moved on to the next phase. Table 10 describes the 37 barriers in detail. Column 1 indicates the SLR stage(s)/sub-stage(s) from the existing literature (Kitchenham 2004, Kitchenham and Charters 2007). Column 2 indicates the Discussion Category of the Group Discussion. Column 3 describes the consolidated barrier. Column 4 provides a count of the number of Post-it notes that were combined into each barrier. Column 5 is discussed in Section 3.5.3.
Table 10. Summary of Group Discussion and Voting Phases

<table>
<thead>
<tr>
<th>SLR Stage</th>
<th>Discussion Category</th>
<th>Barriers to the SLR Process</th>
<th>Ideas</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>P3, P4, E1, E2</td>
<td>Search &amp; Selection Strategy</td>
<td>Inadequate search engines</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Keywords too high-level</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Poor support for automatic search in existing databases</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Report Quality</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Completeness of sources</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Selection Activity</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Which DBs should be used?</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Factoring research question into &quot;executable&quot; strings</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Access to papers</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Making consistent the understanding of incl/excl criteria</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>E5, D1, D2</td>
<td>Analysis &amp; Presentation</td>
<td>Methods to Synthesize Primary Studies</td>
<td>6</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Qualitative Data</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Being More Prescriptive</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How to quantify generalizability</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Best Methods to Present Results</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relating to Industry</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>P5, E3</td>
<td>Quality Assessment</td>
<td>Articles Quality</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Measurement</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Studies Diversity</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Process</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>P6, E4</td>
<td>Data Extraction</td>
<td>Tool Support</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Process Efficacy Assurance</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data Suitability for Purpose</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Primary Papers structure</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Synthesis of qualitative studies</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>P1, P2</td>
<td>Motivation &amp; Research Questions</td>
<td>Connecting with Industry</td>
<td>3</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scoping the SLR</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Structuring Research Questions</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Planning for Relevance</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>*</td>
<td>Overall Protocol</td>
<td>SLR vs Mapping Study</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Decision Tracking &amp; Tractability</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Collaboration</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Standard Terminology</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longitudinal Integration</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SLR Process versus Reality</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Resource Consumption</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tool Integration</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

As previously noted, the results of the **Group Discussion** phase provide an answer for the first and second research questions (RQ1 & RQ2). Regarding the first research question: What additional barriers, beyond the previously identified process barriers, exist in completing SLRs? Comparing the list to previously identified barriers, there are several new barriers identified
among the discussion categories. Across the categories several of the individual barriers identified indicate that the existing primary studies – on which a SLR depends – present additional barriers to the process. The primary factor noted both in the list of composite barriers and a discussion among the participants is the perceived low quality of existing literature. This barrier is present in the *Search and Selection Strategy* category and the *Quality Assessment* category. Also in the *Data Extraction* category, there is a similar barrier with regard to the structure (or lack of structure) of primary study articles that is seen as an impediment to the extraction process. Likewise, in the *Overall Protocol* category, there was discussion of how the lack of standard terminology across the discipline impacts the entire SLR process.

Additional barriers identified by the workshop include:

- Access to primary research articles;
- Methods to synthesize qualitative studies, and;
- Connecting with industry;

Finally, all of the barriers identified in the *Analysis and Presentation* category provide new insights. While previous research (Kitchenham and Brereton 2013) has investigated the synthesis stage, the investigation focused on methods utilized or available for synthesis while ignoring the presentation of the results of the synthesis. The workshop demonstrates there are a number of barriers associated with the *Analysis and Presentation* stages, thus additional research is needed regarding barrier associated with the process stages.

Examining the results with regard to the second research question: What additional characteristics or features should be provided by SLR-support tools? Key findings for this question are that future tools need to provide a means for improving the collaboration among research team members and must contain provisions for tracking decisions made during the
execution of the SLR process. In both cases, the features should be based in an architecture that supports geographically distributed teams.

### 3.5.3 Voting and Raking Phase

After completion of the **Group Discussion** phase, each group presented the results of their deliberation by briefly explaining each barrier. The last column of Table 10 shows the number of votes each barrier received. The relative number of votes among the barriers illustrates the priorities the workshop attendees placed on the barriers. The workshop participants cast more votes (26% of the total cast votes) for the barriers within the **Search and Selection** stage than for any of the other SLR stages (7% more than any other stage). Specifically, the two barriers that received the most votes were *inadequate search engines for the existing databases* and *difficulties associated with keyword selections*. To highlight the highest priority barriers across all SLR phases, Table 11 summarizes the top six barriers.

**Table 11. Top 6 Identified Barriers to the SLR Process**

<table>
<thead>
<tr>
<th>SLR Area</th>
<th>Barriers to the SLR Process</th>
<th>Votes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Extraction</td>
<td>Tool Support</td>
<td>18</td>
</tr>
<tr>
<td>Quality Assessment</td>
<td>Articles Quality</td>
<td>17</td>
</tr>
<tr>
<td>Analysis &amp; Presentation</td>
<td>Methods to Synthesize Primary Studies</td>
<td>14</td>
</tr>
<tr>
<td>Search &amp; Selection Strategy</td>
<td>Inadequate search engines</td>
<td>11</td>
</tr>
<tr>
<td>Analysis &amp; Presentation</td>
<td>Qualitative Data</td>
<td>11</td>
</tr>
<tr>
<td>Motivation &amp; Requirements</td>
<td>Connecting with Industry</td>
<td>9</td>
</tr>
</tbody>
</table>

The results of the **Voting and Ranking** phase provide the insights required to answer the third research question: What priority should be given to each of the known/identified barriers, characteristics, and features? Based upon the voting of the workshop participants, the SLR stages that present the greatest barriers to SLR authors, and thus should receive priority, are the **Search and Selection**, **Data Extraction**, and **Quality Assessment** stages. Taking into account the previous finding noted in section 3.2.2, solutions targeting each of noted areas will:
Reduce the time required to complete a SLR;

The difficult of conducting a SLR; and

Provide tool support for the areas previously identified as in need of it.

3.5.4 Supplementary Notes

The results from this workshop are consistent with the previous findings described in Section 3.2.2. Each of the barriers noted in the previous findings, as well as the key implications of that research, were expressed by participants in this workshop. Even though the previously identified barriers were utilized in the workshop planning, these barriers were not discussed with the participants prior to commencement of, or during, data collection. Therefore, the current results validate the previous findings.

An additional insight expressed throughout the workshop and agreed on by the majority of the participants is that the SLR Process has multiple interactions among the stages. This interactive nature leads to a process that is more iterative in nature when implemented and causes cognitive dissidence for some researchers – most notably novice researchers, which makes the methodology difficult to teach.

Finally, the qualitative analysis of all generated ideas without regard for the SLR stage to uncover any latent themes revealed four primary themes, which are presented and discussed in detail in Section 3.6.

3.6 Discussion

The results of this workshop provided additional evidence about previously identified barriers to the production of SLRs and revealed additional barriers that were previously unidentified. To provide a more global analysis of the SLR ecosystem, this section describes
barriers independent of specific SLR phases. Specifically, the qualitative analysis identified four high-level latent themes among the barriers:

- Barriers Associated with the SLR Process
- Barriers Associated with Primary Studies
- Barriers Associated with the Practitioner Community
- Barriers Associated with Tooling

The remainder of this section discusses the results of the workshop within the context of the identified themes. Following this discussion, a description is provided of the implications of these barriers for:

- SLR Methodology Research
- SLR Authors
- SLR Consumers

This section concludes with a discussion of the limitations of this study.

3.6.1 Barriers Associated with the SLR Process

This subsection presents barriers pertaining to the SLR process itself and to the current methodology for conducting SLRs.

3.6.1.1 The SLR protocol prescribes a sequential process rather than an iterative process.

Even when the final SLR publication provides the sense of a sequential process, the underlying reality is much more iterative and interactive. The SLR protocol, as defined by Kitchenham and Charters (2007), describes a waterfall-style methodology in which authors must perform extensive planning prior to beginning the SLR and in which each SLR step is dependent primarily on the output of its predecessor. Kitchenham and Charters acknowledge that, in
reality, the execution of the protocol is more iterative. Specifically, they suggest the results of a
protocol stage may result in protocol refinements.

Similarly, the workshop participants indicated that, based on their own experiences, the SLR
process is more iterative than this original definition indicates, with interactions among all stages
of the SLR process. For example, one of the ideas recorded on a Post-it note stated:

The question you really have gets manipulated by what evidence is available.

Therefore, it may be valuable to begin describing the SLR process in a more iterative
manner, akin to the approaches used in agile software development, in which each phase is
planned incrementally as the previous phase concludes.

3.6.1.2 Meta-analysis is difficult. For quantitative empirical results, participants lamented the
lack of metrics to describe the similarity between two or more studies — in terms of variables,
underlying conceptual models, experimental operationalization, and data analysis techniques —
which convolutes cross-study meta-analysis. Example ideas recorded on the Post-it notes
included: “how to ‘mix’ evidence from multiple studies” and “how to ‘mix’ evidence across study
designs.” The participants also noted that there is a lack of understanding in how to summarize
qualitative empirical results and that additional meta-methods are needed to aggregate qualitative
studies results. Specifically, ideas recorded on the Post-it notes included “qualitative data rarely
accessible” and “…only the conclusions published.”

3.6.1.3 Researchers lack methods for result interpretation, generalization and framing. In
motivating work (Carver et al. 2013), we noted that novice SLR authors had issues concerning
interpreting results, generalizing results and framing conclusions. The workshop outcomes
reinforce these findings by indicating that more experienced authors also experience
considerable difficulty with a stage that is often overlooked, namely the Analysis and
Presentation stage. Specifically, one of the ideas recorded on a Post-it note was that it is unclear “how to make the evidence actionable.” Combining these two results suggests that the Analysis and Presentation stage has associated barriers for all types of SLR authors.

3.6.1.4 Barriers Associated with Primary Studies. The following quote articulates well a common belief expressed by the workshop participants. This quote came during the Idea Sharing phase of the workshop: SLR’s are only as good as the primary studies they are based on. If the primary studies are of poor quality, the results of the SLR will also be of poor quality. There is a lot of bad research out there and it is critical that we recognize the impact it can have on our results.

It is beyond the scope of this paper to comment on the quality of research; however, the workshop participants identified two sources of cognitive dissidence related to primary studies, which are presented in the following subsections.

3.6.1.5 Titles and abstracts are often misleading or irrelevant. The participants agreed that the titles of the papers describing primary studies are often misleading. Likewise, in many cases the abstract of a published article does not provide the type of information needed to quickly identify the relevance of the primary studies included in that article (e.g. a concise summary of the study results). Both of these issues complicate the Search and Selection SLR stage process.

3.6.1.6 Terminology and its use are not standardized. During a discussion of how keyword issues impede the Search and Selection stage, the group agreed that two of the ideas recorded on Post-it notes best described the issue:

- **Software Engineering terminology** — “there are many synonyms. This implies that research has to be made in several rounds. It is time-consuming.”
• **Lack of unified Software Engineering terms** — “the terminology in Software Engineering is not unified, and many papers are sloppy on defining their concepts.”

The lack of standardized terminology results in added complexity during the *Search and Selection* stage. The group noted that the hierarchical taxonomies utilized by many of the electronic databases do not assist in resolving this problem as many of the terms utilized in the taxonomies are too abstract to be useful. In addition, authors must spend extra time during the study selection process reviewing full articles to understand how the authors are utilizing various terms. This result is consistent with the results of our previous work that found the *Search and Selection* stage to be one of the most time-consuming aspects of the SLR process.

### 3.6.2 Barriers Associated with the Practitioner Community

The participants identified several barriers related to the usefulness of SLRs to the practitioner community. More specifically, these are barriers related to which SLRs should be conducted and how to present those results so they are actionable by practitioners.

#### 3.6.2.1 Researchers have difficulty relating to industry needs

Many workshop participants were disappointed that the results of the SLRs were not utilized more by practitioners. Some specific issues that arose include difficulties with:

- Mapping industrial problems to SLR research questions
- Identifying who should specify the SLR research questions
- Impressing on industry the value of the results of an SLR

The group posited that as a community, we need to better understand how to present the results of an SLR in a manner that is appealing to practitioners. In particular they were unclear on the “*best way to report the data (tables, charts, etc. - which is the most effective.*)”
Drawing an analogy to business markets, researchers require input and feedback from SLR consumers, be it researcher or industry, to understand their needs. Consumers must voice their needs to influence the development of new products and instigate changes to existing products. Only then can SLR producers provide a final product that meets or exceeds expectations and provides some value to the consumer.

3.6.2.2 Researchers have difficulty conveying the value of a structured process. The participants believed that policy and decision makers in industry do not understand the value of a repeatable, SLR process. The participants expressed a common belief that “industry is mostly satisfied with less systematic evidence.”

3.6.3 Barriers Associated with Tooling

Throughout the workshop the participants continually expressed a concern with the amount of time and effort required to complete each stage of the SLR process. The participants agreed that the current set of SLR tools, while helpful in some cases, is inadequate overall. There is a need for a fully integrated set of tools to support the SLR process. As discussed below, everyone expressed concern about publicly available journal databases and managing intermediate results.
3.6.3.1 Electronic database lack adequate search and retrieval facilities. Workshop participants repeatedly pointed out the lack of adequate search engines as a problem and ranked that barrier high during the voting process. Because many of the commonly-used electronic databases provide inconsistent information, the Search and Selection stage produces unnecessary duplication. Some specific concerns raised by the participants include the following:

- Many of the digital libraries provide limited search facilities and poor support for the mass download or culling of primary studies and their titles, abstracts, and other metadata.
- There was great concern that incompatibilities in the search engines often resulted in the need to rewrite search strings for each database.
- Due to the issues with titles and abstracts, authors have to perform searches against the full text of the articles, rather than using the metadata fields provided by a given database.

3.6.3.2 Tools to support data extraction and management are inadequate or nonexistent. The number one barrier identified in the voting process was the lack of tool support in the Data Extraction stage. The participants indicated that the data extraction and management process should be more defined more broadly than just automated extraction of bibliographic details. Rather, it should also include an emphasis on the management of extracted data and traceability of that data back to the original source. The participants also expressed the need for these tools to have bidirectional interfaces to permit the transfer of data to/from existing analytic toolsets.

3.6.4 Implications for SLR Methodology Research

For researchers seeking to improve the SLR research process, the results of this study provide a roadmap of the SLR community needs. Each barrier identified may be viewed as a
point of departure for future research and tool development. Moreover, the voting results provide prioritization for research and tool development that will maximize the benefits of such efforts. Specifically, methods are needed to:

- Reduce the time and effort required to search for and select primary studies
- Improve the assessment of the quality of primary studies
- Synthesize primary studies that differ in design or context

Work in these high-priority areas will permit more timely production of SLR’s, improve the quality of the SLR process, and increase researchers’ abilities to triangulate knowledge and thus to advance understanding of a given domain.

3.6.5 Implications for SLR Authors

The results of the workshop serve as a checklist of the potential barriers that SLR authors may encounter during the execution of the SLR process. With knowledge of these barriers, SLR authors can better mitigate the challenges associated with the process through improved preparation and planning. Specifically, recognition of the time required to complete each phase of the SLR process will allow authors to develop more realistic time-lines, improve goal-setting, and more accurately estimate the resources required during the planning phase of a SLR.

3.6.6 Implications for SLR Consumers

For SLR consumers, the results of the workshop provide insights into the production process and potential interface points that may be used to guide the direction of SLR research, assist in the SLR production process, and improve the usability of SLRs in practice. In addition to the researchers conducting the SLR research process, external subject matter expertise from the practitioner community is needed as input to determine and prioritize the topics or questions to
be studied in future SLRs. Workshop participants indicated that subject matter experts are also needed to assist in the development and review of SLR plans. Collectively, the workshop facilitators and participants concluded that the usability and durability of the information contained in the final SLR product (e.g., the research report or article) would be improved via feedback provided by SLR consumers.

3.6.7 Limitations

While the results of this study provide guidance for future SLR researchers and consumers, some care should be taken concerning the generalizability of the results. The limits on generalizability stem from:

- Selection bias — the self-selected workshop participants may mean that the sample is not representative of the whole population of SLR authors
- Experience/proficiency bias — a majority of the workshop participants were accomplished SLR researchers whose views may not be representative of all SLR authors
- Time-limit bias — the two-hour time-limit may have constrained both the idea generation and group discussions of the participants

Even with these issues, the results of the workshop are still consistent with the findings of our previous work (Carver et al. 2013). However, given additional time, participants may have identified additional barriers to the SLR process. In fact, some participants were still providing input as the workshop was concluding. All participants agreed to participate in additional focus groups as this research continues.
3.7 Conclusion and Future Work

This paper describes the goals, methodology, and results of a community workshop held during ESEIW’13 prior to ESEM’13. The goals of the workshop were to augment a list of barriers to the SLR process that were identified in previous work (Carver et al. 2013) and to rank the previously and newly identified barriers by priority. During the workshop, data was collected using the Nominal Group Technique (Delbecq et al. 1975), and after the workshop the collected data are analyzed. Table 10 and Table 11 summarize the workshop results, and Section 3.6 discusses the identified barriers (in the context of four latent themes reveal through qualitative analysis) and their implications.

The results of the community workshop provide a better understanding of the barriers to the SLR process, including new information about which barriers the community thinks are in urgent need of being addressed. The workshop was an important step toward improving the SLR process, and serves as motivation for future work to better understand how the highest-priority barriers can benefit from tool support. That is, the community will once again be engaged to elicit the requirements that SLR-support tools must satisfy to address each barrier and to prioritize the requirements relative to each barrier.

It is anticipate that the results of the next community workshop will provide concrete guidance to researchers engaged in the development of SLR-support tools and thus will instigate needed improvements to existing SLR-support infrastructure. Such improvements will reduce the effort required to conduct SLRs and encourage the production of more SLRs, which will contribute to the maturation of SE and ultimately lead to better software.
3.8 References


Delbecq, Andre L, Andrew H Van de Ven, and David H Gustafson. 1975. Group techniques for program planning: A guide to nominal group and Delphi processes: Scott, Foresman Glenview, IL.


Felizardo, Katia Romero, EY Nakawaga, Daniel Feitosa, Rosane Minghim, and José Carlos Maldonado. 2010. An approach based on visual text mining to support categorization and classification in the systematic mapping. Paper read at Proc. of EASE.

Felizardo, Katia Romero, Mehwish Riaz, Muhammad Sulayman, Emilia Mendes, Stephen G MacDonell, and José Carlos Maldonado. 2011. Analysing the use of graphs to represent the results of Systematic Reviews in Software Engineering. Paper read at Software Engineering (SBES), 2011 25th Brazilian Symposium on.

Felizardo, Katia Romero, Norsaremah Salleh, Rafael Messias Martins, Emilia Mendes, Stephen G MacDonell, and José Carlos Maldonado. 2011. Using visual text mining to support the


Kitchenham, Barbara, and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University.


CHAPTER 4

Selection by Example in Software Engineering Systematic Reviews

4.1 INTRODUCTION

A Systematic Literature Review (SLR) is a means by which researchers provide an unbiased, repeatable comprehensive evaluation and synthesis of a topic of interest based on primary research studies. Such studies are critical not only to the advancement of research concerning a topic, but are also a key component in providing guidance regarding best practices and empirically grounded evidence to practitioners in the field.

In 2004, the Software Engineering (SE) discipline embarked on the implementation of Evidence-Based Software Engineering (EBSE) and adopted the conduct of SLRs as a key component of the paradigm (Kitchenham et al. 2004, Kitchenham 2004). Over the last decade since its adoption, nearly 300 SLRs have been published covering a wide variety of topics.

When conducting a SLR, researchers initially plan the review in the form of a formal protocol which includes criteria for including and excluding studies from the review. Next, researchers enter the execution phase of the review in which the literature is searched, studies are selected and assessed, data are extracted from the studies, and the data are synthesized. Finally, the results of the investigative effort and findings from the synthesis are documented in a written report. At each stage of the process, multiple researchers work independently and then meet to resolve any differences through discussion (Kitchenham 2004, Biolchini et al. 2005, Kitchenham and Charters 2007).
One of the main barriers in conducting SLRs is the time associated with the selection of studies from the extensive set of results returned as part of the search stage of the protocol (see Chapter 3). During the selection stage of the protocol, researchers first review the title of each article to make an initial determination of the relevance of a study. Studies deemed clearly irrelevant are excluded from further consideration, studies judged as potentially relevant are retained for further examination. This process is repeated in subsequent rounds utilizing the abstracts of the remaining studies, and finally the full text of each study (Kitchenham and Charters 2007).

A common problem in addition to the time required for the selection of studies to be included in the review is the consistent application of the inclusion/exclusion criteria of the review (Carver et al. 2013). While inconsistencies in the application of the inclusion/exclusion criteria are resolved through the discussions among researchers, such inconsistencies result in the expenditure of additional time to resolve the inconsistencies and re-accomplish the selection process.

In order to improve upon the consistency of the selection process and reduce the amount of time researchers must dedicate to the task, some researchers have turned to technology as a substitute for human labor. In the healthcare domain, linear systems (Cohen et al. 2006), factorized complement naïve Bayes classification (Matwin et al. 2010), and Vector Support Machines combined with supervised learning (Wallace et al. 2010) have been investigated as automated classifiers for systematic reviews. However, in each case there is a need for extensive training and tuning of the classifier, a factor that reduces the time saved.

Similarly, Visual Data Mining has been proposed in the fields of healthcare, social science, and software engineering as a time saving measure during the selection of studies for reviews.
(Felizardo, MacDonell, et al. 2012, Romero Felizardo et al. 2013, Felizardo et al. 2011, Malheiros et al. 2007) and for the validation of the selection process (Felizardo, Andery, et al. 2012). While this approach shows promise in assisting researchers in the selection process, it is still time consuming and requires specialized training to be effective with the outcome remaining subject to researcher bias (see Malheiros et al. 2007).

A potential solution to the problem of classification time expenditure and consistency is an automated classifier that does not require extensive training or tuning in order to select studies for inclusion in a review. Ideally, the researcher presents several example studies that exemplify what should be included, and the classifier then selects similar studies. One potential basis for the construction of such a classifier is Latent Semantic Analysis, a member of the VSM family of algorithms, which has been utilized in the extraction of semantic factors from a corpus of text and offers a means by which to compare the extracted factors.

4.1.1 Research Objectives

This study investigates the use of two algorithms, the Vector Space Model (VSM) and Latent Semantic Analysis (LSA), as a means of classifying primary studies in the selection process of systematic review methodologies. The objective of this investigation is to evaluate the recall, precision, and possible work savings of the VSM and LSA algorithms in a classifier capable of selecting studies, based on example, for inclusion in SLR type studies.

4.1.2 Research Questions

In support of the outlined objectives, the research questions addressed by this study are:

RQ1: Can the VSM and LSA algorithms be utilized to classify publications for inclusion/exclusion in the context of secondary SE research?
**RQ2:** How effective are the VSM and LSA techniques with regard to:

- **RQ2a:** recall (an indicator of the number of studies missed)?
- **RQ2b:** precision (an indicator of false positives)?
- **RQ2c:** work savings (an indicator of time saved)?

The remainder of this chapter is organized as follows. Section 4.2 provides an overview of the VSM and LSA models, along with a summary of previous work. Section 4.3 details the conceptual development of the classifiers. The evaluation of the classifiers is described in Section 4.4. Section 4.5 reports the results of the evaluation, and Section 4.6 discusses the implications and limitations of the results. Finally, Section 4.7 concludes the chapter and describes future work.

### 4.2 BACKGROUND

#### 4.2.1 Previous Efforts to Support Selection

Efforts to support the selection of primary studies during the execution of SLRs span several technologies. In the medical domain, linear systems that combine feature values assigned to studies are tested and found to provide a work savings in most cases of between 10% and 30% with a 95% recall rate (Cohen et al. 2006), where recall rate is defined as the portion of studies included that should have been included. The use of factorized complement naïve Bayes classification with weight engineering in systematic drug class reviews have provided an estimated workload reduction of 33.5% on average (Matwin et al. 2010). However each of these cases involve the need for extensive training and tuning of the classifiers, a factor left unconsidered in the estimate of work saving.

Wallace et al. (2010) employed Vector Support Machines in combination with supervised learning and estimate the method reduces the number of studies to be reviewed by 50% with no
loss of recall. Similarly, Visual Data Mining has been utilized in fields of medicine, social science, and software engineering (Felizardo, MacDonell, et al. 2012). In the domain of software engineering, Visual Data Mining is proposed as both a means of supporting the selection of studies during Systematic Literature Reviews (Felizardo, MacDonell, et al. 2012, Romero Felizardo et al. 2013, Felizardo et al. 2011, Malheiros et al. 2007) and for the validation of the selection process (Felizardo, Andery, et al. 2012). While this approach shows promise in assisting researchers in the selection process, it is still time consuming and requires specialized training to be effective with the outcome remaining subject to researcher bias (see Malheiros et al. 2007).

4.2.2 The Vector Space Model

Originally conceived as a method of indexing, the VSM is a representation of a corpus of documents in N-dimensional space (Salton et al. 1975). Each document in the corpus is represented by an N-dimensional vector in which each cell in the vector equates to the frequency count of a term used in the document. The collection of document vectors – in the form of column vectors – are then aggregated to form a matrix in which the columns represent the documents in the corpus and the row vectors represent the frequency of terms utilized in the corpus for each document (Landauer et al. 1998, Berry et al. 1995, Salton et al. 1975). Thus each document is represented in \( t \) dimensions, where \( t \) is the number of unique terms in the entire corpus. This matrix is referred to as the term-document matrix for the corpus.

In the original application, document vector lengths of the term-document matrix are normalized to one, yielding a projection of each document onto the unit sphere (Salton et al. 1975). In the projection, documents with similar terms appear closer together than those with dissimilar indexed terms. To produce optimal clustering of the document projections, term
weighting is applied to each term in the term-document matrix (Salton et al. 1975). Optimal term weighting methods vary based on the corpus being indexed, but generally take the form:

$$term_{i,j} = LWF(i, j) \times GWF(i)$$  \hspace{1cm} (1)

Where $LWF(i, j)$ is a local weight function expressing the weight of term $i$ in document $j$, and $GWF(i)$ is a global weight function expressing the weight of term $i$ across all documents (Nakov et al. 2001). The selection of weighting functions is generally considered a step in tuning the model for optimal retrieval and a-theoretical in that the functions selected are based on what works best for a given corpus with a typical query. For a detailed discussion of term weighting the reader is referred to Nakov et al. (2001).

Queries to retrieve documents based on a set of terms are accomplished by constructing a pseudo-document in the form of a document vector in $t$ dimensions, applying the appropriate weighting, normalizing in relationship to the space, and projecting the vector into the document space (Berry et al. 1995, Salton et al. 1975). The distance between the projected pseudo-document and other document projections of the collection are then utilized to identify documents which are similar to the original query (Salton et al. 1975).

4.2.3 Latent Semantic Analysis

Three fundamental issues in information retrieval are addressing the problems associated with synonymy, polysemy, and compound terms (Deerwester et al. 1990). In the case of synonymy, the search terms entered by a user may be expressed in different terms with similar meanings within the corpus to be search. In the same manner, polysemy – where a term has multiple semantic meanings – generates cases where the user intends one meaning of the term while documents in the corpus utilize alternative meanings. In the case of compound terms, if each word in the compound term is treated in isolation the resulting matches may be due simply
to co-occurrence of the words in a document rather than their use as a compound term in the
document. In all cases, the retrieved results do not meet users’ expectations of the search
(Deerwester et al. 1990).

In response to the issues of synonymy, polysemy and compound terms, Deerwester et al.
(1990) proposed what can be viewed as an extension to VSM known as Latent Semantic
Analysis (LSA). The process begins similarly to VSM with the assembly of a weighted term-
document matrix. Next, in place of the vector length normalization process, Singular Value
Decomposition (SVD) of the term-document matrix is undertaken (Deerwester et al. 1990, Berry
et al. 1995) such that:

\[
X = T_0 S_0 D_0^T
\]

(2)

where \(X\) is the original term document matrix, \(T_0\) is a matrix of orthonormal, singular column
vectors representing the terms in \(X\), \(S_0\) is a diagonal matrix of the singular values of \(X\) arranged in
descending order, and \(D_0\) is a matrix of orthonormal, singular column vectors representing the
documents in \(X\) (Berry et al. 1995).

The SVD of the term-document matrix is believed to characterize the semantic dimensionally
of the documents represented by the term-document matrix (Berry et al. 1995, Landauer et al.
1998, Nakov et al. 2003). By zeroing out singular values of \(S_0\) below a certain threshold, the
amount of semantic noise is reduced (Deerwester et al. 1990, Berry et al. 1995), thus allowing a
more meaningful match to queries. The number of dimensions to retain is a part of the tuning
process of the model with 100 to 400 dimensions typically providing best performance (Nakov et
al. 2003). A method of theoretically determining the precise number of dimensions is an open
question in the literature.
After zeroing out the smaller singular values in $S_0$, the three matrices of the decomposition are multiplied as indicated in the decomposition above to form a least-squares estimate of the original matrix known as the semantic space. A side effect of this process is that the reduction in semantic dimensionality collapses the weighting of synonyms causing each of the terms to be weighted equally (Landauer et al. 1998). Another side effect of the process is the polysemy issue is partially resolved as the weighting of such terms are conditioned by other words in the document (Deerwester et al. 1990). Similarly, compound terms – and the individual terms within them – are weight conditioned by surrounding terms in the document.

When utilized for indexing, queries are constructed by forming a pseudo-document of the query and scaling the resulting vector into the semantic space such that:

$$D_q = X_q^T S^{-1}$$

(3)

where $D_q$ is the resulting scaled pseudo-document, $T$ and $S$ are the matrices from the SVD used to create the semantic space of the corpus, and $X_q$ is the pseudo-document representing the query of interest. The scaled pseudo-document is then compared to the other documents in the corpus by a similarity measure (Deerwester et al. 1990, Landauer et al. 1998). Common similarity measures are (Wild and Stahl 2007):

1. The Cosine of the angle formed by two vectors;
2. Pearson’s Correlation;
3. Spearman’s Rho; and
4. The cross-product of the vectors.

Again, the selection of a similarity measure is a-theoretical and based primarily on what yields the best results for a given situation (Wild and Stahl 2007).
4.2.4 Applications of LSA

In addition to the indexing application previously described, LSA has also been applied in a number of other areas. After training with a large corpus of English text, LSA performance on standardized tests for synonym identification is found to be equivalent to the performance of United States college applicants from non-English speaking countries (Landauer et al. 1998).

Similarly, LSA judgments of word relatedness and sorting, as well as subject matter knowledge testing with multiple choice questions show better than chance performance – and in many cases mimic human performance (Landauer et al. 1998). Additionally, LSA has been utilized to model lexical semantic priming for lexical decisions concerning polysemy (Till et al. 1988).

An application of primary interest to this study is the use of LSA for the assessment of essay content and quality (Kakkonen and Sutinen 2004, Chung and Neil 1997, Landauer et al. 1998). In this application, a set of essays that have previously undergone human scoring are utilized for LSA training. New essays are then processed by transforming them into a document vector scaled into the semantic space and comparing the new document with the training set (Kakkonen and Sutinen 2004, Landauer et al. 1998, Chung and Neil 1997). Similarity of meaning is then assessed by the closest match in the training set based on the Cosine measure (Kakkonen and Sutinen 2004). Similarity of content is assessed either as the sum of similarities between the new essay and all essays in the training set (Kakkonen and Sutinen 2004), or as the length of the vector (Chung and Neil 1997).
4.3 Conceptual Development

4.3.1 Goals in the Selection Process

The goal of the selection process in a SLR is to retain the relevant primary studies found during the search process while excluding irrelevant studies from further analysis. Thus the selection of studies for inclusion in a SLR is a dichotomous decision task where optimally the risk of a Type-I error (rejecting a study that should be included) is minimized. Type-II errors (including a study that should be rejected), are of less concern as it is assumed the error will be discovered at a later point in the study. Thus a recovery mechanism is in place for the correction of Type-II errors, while Type-I errors are unrecoverable due to the process as it is defined.

Furthermore, the goal of automating the classification process is to reduce the amount of time researchers consume in the selection process. Thus correct classifications are the most desirable as Type-II errors invoke the use of additional processing resources, in the form of a researcher’s time, to recover from the error.

Therefore, the goal of a classifier is to make dichotomous decisions with minimal risk of a Type-I error, while minimizing the risk of a Type-II error. In other words, should the classifier make an error in classification it is preferable that the error be a Type-II error.

4.3.2 Assessment of the Selection Process

In order to assess performance of a classifier in relation to the goal, a measure of each type of error and the work savings is required. Traditional proportions of each type of error with respect to the total number of studies classified provide a general indicator of classifier performance, but may be misleading (Cohen et al. 2006). Therefore alternative measures are utilized. Measures from the field of information retrieval provide the indicators of each error type. An indicator of the lack of Type-I errors is defined by the *Recall* statistic (Cohen et al. 2006):
The recall measure is an inverse measure of Type-I errors with respect to the actual number of studies that should be included. Based on the goal of the classifier, a recall measurement of one is the ideal goal with measurements less than one indicating increased Type-I errors.

An indicator of Type-II errors is found in the measure of precision (Cohen et al. 2006):

\[
Precision = \frac{\text{Number of Included Studies Correctly Classified}}{\text{Total Number of Studies Classified as Included}}
\]

The precision measure is, once again, an inverse measure of Type-II errors with respect to the number of studies marked for inclusion. Based on the goal of the classifier, a precision measurement of one is the ideal goal with measurements less than one indicating an increase in the number of Type-II errors.

Turning to the measurement of work savings, only one statistic is currently noted in the literature. Cohen et al. (2006) propose a “measure of the work saved over and above the work saved by simply sampling for a given level of recall.” The issue with this measure is that while it measures future work savings, it does not account for the time invested by the researcher to train the classifier. From a decision making standpoint, a classifier that simply acts a pass-through would have a recall of 1.0, precision of 0.0, and a work savings of 0%: Essentially leaving all the decisions to the researcher. On the other hand, a classifier that requires no training, achieves a recall of 1.0 and precision of 1.0 would provide a work savings of 100%. For each training case presented to the classifier, a previous decision as to the inclusion or exclusion of the case must be made by the researcher, essential reducing the amount of work that may be saved by the use of a classifier. While an argument can be made that the time required to review the article is also a factor and therefore a target for optimization (see Felizardo et al. 2011, Malheiros et al. 2007),
removal of the need for the decision negates the need for the preparation time. Therefore, this study defines the work savings in terms of the number of decisions the researcher must make:

\[
Work\ Savings = \frac{\text{Number of Included Studies Correctly Classified}}{\text{Total Number of Studies in the Collection}} \times 100 \tag{6}
\]

Thus the work savings for the initial selection process is more accurately portrayed as the training of the classifier is taken into account, and future work that utilizes the trained classifier when updating a SLR will reflect an increased savings due to the absence of training – assuming there is no change in the inclusion/exclusion criteria.

4.3.3 Characteristics of the Selection Process

Ideally, a researcher should be able to supply the inclusion and exclusion criteria for the decision process from the SLR protocol and the studies to be marked to the classifier and receive back the set of studies to be included in the review. As the complexity of processing the raw inclusion/exclusion criteria from the protocol is – to the best of the researcher’s knowledge – beyond the scope of current capabilities, the perspective that each study marked by the researcher for the purpose of training the classifier embodies the inclusion/exclusion criteria set forth in the protocol of a study. Thus training by example provides a simple, intuitive means for researchers to train the classifier at the expense of work savings.

Another practical consideration is the representation of the study utilized in the consideration for inclusion or exclusion. Ideally, the full text of the study is utilized in the decision process in order to maximize the probability of a correct decision. This is impractical however when the preceding search process yields several thousand studies to be classified. The time, cost, and storage requirements of downloading the full text of each study is a primary resource concern given the percentage of studies typically retained. A more practical solution is to take a phased
approach similar to that currently used by SLR researchers. First, metadata – including the
abstract – of each study is downloaded and an initial decision regarding the inclusion or
exclusion of each study is made. If there is any doubt as to whether a study is to be included or
excluded, the study is provisionally included pending additional information. After the initial
classification, the full text of each study identified for inclusion is downloaded and examined
further. Utilizing this approach provides for conservation of resources at the risk of an increase
in Type-II errors. Thus in the selection phase of an SLR, the ability to classify studies based on a
minimal amount of information is desirable. Furthermore, in the interest of resource
conservation it is permissible to trade an increase in Type-II errors for a decrease in Type-I
errors.

4.3.4 Proposed Classifier

The successes of LSA in the essay assessment task serves as the inspiration for the proposed
classifier. This study hypothesizes that by providing known examples of studies to be included
and excluded, an LSA based classifier is able to correctly classify previously unseen studies. In
addition, the amount of information needed to determine the classification is limited to a subset
of the metadata commonly found in citation downloads – specifically the title and abstract of
each study.

To construct the LSA classifier, the title and abstract of each study are parsed to form a
document vector consisting of the word frequencies found in the text. Next, term vectors from
studies identified for training and previously marked as included or excluded are aggregated to
form a term-document. The term-document matrix is then submitted to LSA to form the
semantic space for comparisons. After construction of the semantic space, the term vectors of
the studies to be classified are scaled into the semantic space as previously noted. Following the
scaling process, a term vector in need of classification is compared to each term vector in the training set to find the closest match. If the closest match for the term vector to be classified is a term vector associated with a study marked for inclusion, then the study represented by the term vector to be classified is marked for inclusion. If the closest match for the term vector to be classified is a term vector associated with a study marked for exclusion, then the study represented by the term vector to be classified is marked for exclusion. This process continues until all term vectors in need of classification are marked as either included or excluded.

4.3.5 Factors Impacting LSA Performance

As the classifier is based on LSA, there are four known primary factors which impact performance of the classifier (Landauer et al. 1998, Nakov et al. 2001, Nakov et al. 2003, Wild and Stahl 2007):

1. Preprocessing of the text.
2. Weighting functions applied to the term frequencies.
3. The dimensionality selected.
4. The similarity measure employed for comparisons.

Preprocessing factors include stop word filtering, stemming, restrictions based on local or global term frequency counts, and the use of controlled vocabulary (Nakov et al. 2003, Wild and Stahl 2007). Stop word filtering is the exclusion of common words that appear with high frequency in a language such as articles (e.g. In English – a, an, and the). Stemming is the process of reducing a word to its root form. Each of these options influence the terms produced for inclusion in the term vector of a document. Restrictions based on a minimum/maximum appearance of a term within a document, or across the document set, may increase/decrease the number of terms included in the term vector. Likewise, the use of a controlled vocabulary
restricts the terms included in the term vector of each document. No firm guidance is available regarding the use of these options. As a result, this study elects to employ only two of the options: Stop word filtering; and stemming. It is the belief of the researcher that the removal of stop words and the use of stemming will each reduce noise in the dataset.

Weighting functions – the transformation of term frequencies – are shown to significantly influence the performance of LSA (Nakov et al. 2001, Nakov et al. 2003). Nakov et al. (2001) investigate the use – and absence – of local and global weighting functions. They find the use of weighting functions to be beneficial to LSA. Of the 12 combinations tested the combination that consistently resulted in superior performance is the use of the Log function for local weighting and Shannon’s Entropy function for global weighting. Based on their findings, this study elects to use a local log weighting in combination with globally based entropy for term weighting.

With regard to the selection of dimensionality, this researcher has found no conclusive guidance regarding the number of dimensions to be retained in LSA. Lacking guidance, two options are explored:

1. Retain all dimensions from the decomposition. This is equivalent to a simplified VSM treatment of the term vectors and will provide a basis for comparisons.

2. Retain the number of dimensions at which the sum of the dimensions divided by the total of all dimensions meets or exceeds 0.5. This is the default setting for the R statistical package which was selected for the analysis.

As previously noted, there are four primary methods of comparing vectors. For the comparison of vectors in essay assessment tasks, Wild et al. (2007) report that the use of the Cosine formed by the vectors provided the most robust results. Thus, the cosine measure is employed for comparison of study vectors in this study.
4.4 Classifier Evaluation

4.4.1 Dataset

The citations downloaded during the execution of a previous mapping study regarding Evidence-Based Software Engineering (see Chapter 2) are utilized as the data to evaluate the classifier. A total of 5979 unique citations were selected to construct the dataset. The publications selected for inclusion in the mapping study analysis consist of 157 citations in the dataset. The remainder of the dataset, 5822 citations, represents publications returned by the search query of the mapping study, but ultimately excluded from the analysis. All the citations are downloaded into the EndNote® citation manager. For each citation, the title and abstract are exported and placed into a text file identified by the associated record number. No modifications are made to the data exported. The studies represented by the citations span a period of time beginning in January 2004 through the end of June 2014.

4.4.2 Classification of Citations

The proposed classifier algorithm is implemented utilizing version 3.1.1 of the R statistical package and the associated LSA add-in. The LSA add-in provides the parser and associated functionality required to convert individual text files into term vectors and term-document matrices. The add-in also contains convenience functions for scaling and measurement.

In order to gain a better understanding of the performance of the classifier, 11 independent trials for each of the dimensional selection methods are accomplished. For the initial trial of each dimensional selection method, the title-abstract combinations for publications in the year 2004 are used as a training set. In each subsequent trial, the title-abstract combinations of publications from one additional year are added to the training set. For the final trial, a scenario which leverages the quasi-gold standard (He et al. 2011) set of publications – a known set of
publications utilized to assess the search stage of the mapping study – to improve recall is employed. This provides a scenario in which the maximum amount of previous effort is leveraged in the training of the classifier. For each trial, the remainder of the dataset is utilized as a holdout sample to be classified. The output of each trial consists of an identifier of the title-abstract combination, an identifier of the matched title-abstract combination, the Cosine of the angle between the vectors, and a code indicating the inclusion/exclusion decision.

4.4.3 Analysis

Following the completion of each trial, the include/exclude decision of the classifier is compared to the original decision made in the mapping study for each of the publications. The result of this comparison presents four possible outcomes as shown in Figure 21.

![Figure 21. Classifier versus origin decision outcomes](image)

For each trial, the percentage of correct classifications (e.g. the sum of correct inclusions and correct exclusions), Type-I errors, and Type-II errors are tracked as indicators of the classifiers general performance. In addition, the recall, precision, and worked savings for each trial is calculated to provide a more detailed assessment of classifier performance.
4.5 Results

Table 12 summarizes the division of the publication records over the course of the 11 independent trials of the classifier. For each trial, the timeframe covered by the subset of the overall dataset is indicated, along with the number of studies originally marked for inclusion and exclusion. For the training set within each trial, the number of unique terms extracted by the parser from the training set is indicated by the vocabulary column. Extraction of terms in the holdout set of each trial is restricted to the vocabulary of the training set to facilitate comparisons.

Table 12. Training and holdout dataset compositions

<table>
<thead>
<tr>
<th>Run</th>
<th>Timeframe</th>
<th>Include</th>
<th>Exclude</th>
<th>Vocabulary</th>
<th>Timeframe</th>
<th>Include</th>
<th>Excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2004</td>
<td>8</td>
<td>308</td>
<td>4085</td>
<td>2005-2014</td>
<td>149</td>
<td>5514</td>
</tr>
<tr>
<td>2</td>
<td>2004-2005</td>
<td>22</td>
<td>671</td>
<td>6198</td>
<td>2006-2014</td>
<td>135</td>
<td>5151</td>
</tr>
<tr>
<td>3</td>
<td>2004-2006</td>
<td>28</td>
<td>1086</td>
<td>8203</td>
<td>2007-2014</td>
<td>129</td>
<td>4736</td>
</tr>
<tr>
<td>5</td>
<td>2004-2008</td>
<td>69</td>
<td>2055</td>
<td>11608</td>
<td>2009-2014</td>
<td>88</td>
<td>3767</td>
</tr>
<tr>
<td>6</td>
<td>2004-2009</td>
<td>85</td>
<td>2694</td>
<td>13344</td>
<td>2010-2014</td>
<td>72</td>
<td>3128</td>
</tr>
<tr>
<td>7</td>
<td>2004-2010</td>
<td>100</td>
<td>3396</td>
<td>15150</td>
<td>2011-2014</td>
<td>57</td>
<td>2426</td>
</tr>
<tr>
<td>8</td>
<td>2004-2011</td>
<td>120</td>
<td>4197</td>
<td>17000</td>
<td>2012-2014</td>
<td>37</td>
<td>1625</td>
</tr>
<tr>
<td>9</td>
<td>2004-2012</td>
<td>136</td>
<td>5014</td>
<td>18869</td>
<td>2013-2014</td>
<td>21</td>
<td>808</td>
</tr>
<tr>
<td>10</td>
<td>2004-2013</td>
<td>156</td>
<td>5749</td>
<td>20904</td>
<td>2014</td>
<td>1</td>
<td>73</td>
</tr>
<tr>
<td>11</td>
<td>2004+QGS</td>
<td>78</td>
<td>308</td>
<td>4304</td>
<td>2005-2014</td>
<td>79</td>
<td>5514</td>
</tr>
</tbody>
</table>

Table 13. VSM classification trial results

<table>
<thead>
<tr>
<th>Trial</th>
<th>Correctly Excluded</th>
<th>Correctly Included</th>
<th>Type-I Errors</th>
<th>Type-II Errors</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5271</td>
<td>100</td>
<td>49</td>
<td>243</td>
</tr>
<tr>
<td>2</td>
<td>4900</td>
<td>87</td>
<td>48</td>
<td>251</td>
</tr>
<tr>
<td>3</td>
<td>4530</td>
<td>93</td>
<td>36</td>
<td>206</td>
</tr>
<tr>
<td>4</td>
<td>4114</td>
<td>78</td>
<td>24</td>
<td>166</td>
</tr>
<tr>
<td>5</td>
<td>3617</td>
<td>67</td>
<td>21</td>
<td>150</td>
</tr>
<tr>
<td>6</td>
<td>3021</td>
<td>63</td>
<td>9</td>
<td>107</td>
</tr>
<tr>
<td>7</td>
<td>2348</td>
<td>51</td>
<td>6</td>
<td>78</td>
</tr>
<tr>
<td>8</td>
<td>1578</td>
<td>31</td>
<td>6</td>
<td>47</td>
</tr>
<tr>
<td>9</td>
<td>785</td>
<td>17</td>
<td>4</td>
<td>23</td>
</tr>
<tr>
<td>10</td>
<td>71</td>
<td>1</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>4636</td>
<td>72</td>
<td>7</td>
<td>878</td>
</tr>
</tbody>
</table>
The raw results of the VSM and LSA based classifiers are presented in Table 13 and Table 14 respectively. These raw counts are utilized to produce the performance assessment statistics for the VSM classifier shown in Table 15, and the LSA classifier in Table 16.

As one would anticipate, as the number of publications used to the train of the classifier increased there is a general improvement in the number of correctly classified publications and an accompanying decrease in work savings. The VSM classifier displayed a mean recall of 0.81 with a standard deviation of 0.11 indicating that, on average, it correctly identified 81% of
publications that should have been included. The mean precision of the classifier is found to be 0.32 with a standard deviation of 0.09 indicating that only 32% of the publications marked for inclusion by the classifier are actually relevant. While the precision of the classifier appears low, the potential work savings of the classifier are considerable. The work savings are above 40% in all but three of the trials with a maximum work savings of 89.8%. Pearson’s correlation coefficient reveals a strong relationships between work savings and recall (-0.70), as well as between work savings and precision (-0.61).

Table 16. LSA classifier performance

<table>
<thead>
<tr>
<th>Trial</th>
<th>Correct %</th>
<th>Type-I %</th>
<th>Type-II %</th>
<th>Precision</th>
<th>Recall</th>
<th>Savings %</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>94.3</td>
<td>0.8</td>
<td>4.9</td>
<td>0.27</td>
<td>0.69</td>
<td>89.30</td>
</tr>
<tr>
<td>2</td>
<td>94.3</td>
<td>1.2</td>
<td>4.6</td>
<td>0.23</td>
<td>0.55</td>
<td>83.36</td>
</tr>
<tr>
<td>3</td>
<td>95.4</td>
<td>0.8</td>
<td>3.8</td>
<td>0.33</td>
<td>0.71</td>
<td>77.66</td>
</tr>
<tr>
<td>4</td>
<td>96.3</td>
<td>0.9</td>
<td>2.8</td>
<td>0.34</td>
<td>0.63</td>
<td>70.60</td>
</tr>
<tr>
<td>5</td>
<td>96.5</td>
<td>0.7</td>
<td>2.8</td>
<td>0.36</td>
<td>0.70</td>
<td>62.22</td>
</tr>
<tr>
<td>6</td>
<td>96.7</td>
<td>0.4</td>
<td>2.9</td>
<td>0.38</td>
<td>0.81</td>
<td>51.73</td>
</tr>
<tr>
<td>7</td>
<td>97.0</td>
<td>0.4</td>
<td>2.6</td>
<td>0.42</td>
<td>0.81</td>
<td>40.27</td>
</tr>
<tr>
<td>8</td>
<td>97.0</td>
<td>0.5</td>
<td>2.5</td>
<td>0.41</td>
<td>0.76</td>
<td>26.96</td>
</tr>
<tr>
<td>9</td>
<td>97.0</td>
<td>0.5</td>
<td>2.5</td>
<td>0.45</td>
<td>0.81</td>
<td>13.45</td>
</tr>
<tr>
<td>10</td>
<td>95.9</td>
<td>0.0</td>
<td>4.1</td>
<td>0.25</td>
<td>1.00</td>
<td>1.19</td>
</tr>
<tr>
<td>11</td>
<td>87.7</td>
<td>0.1</td>
<td>12.2</td>
<td>0.10</td>
<td>0.95</td>
<td>82.04</td>
</tr>
</tbody>
</table>

The LSA classifier shows similar characteristics. As the number of publications used to the train of the classifier increases there is a general improvement in the number of correctly classified publications. The LSA classifier displayed a mean recall of 0.76 with a standard deviation of 0.13 indicating that, on average, it correctly identified 76% of publications that should have been included. The mean precision of the classifier is found to be 0.32 with a standard deviation of 0.10 indicating that only 32% of the publications marked for inclusion by the classifier are actually relevant. The work savings are above 40% in all but three of the trials with a maximum work savings of 89.3%. Pearson’s correlation coefficient reveals modest
relationships between work savings and recall (-0.57), as well as between work savings and precision (-0.48).

Comparing the results from each classifier over the trials, two-tailed paired $t$-Tests show no significant difference in the precision ($p = 0.63$) or work savings ($p = 0.21$) of the classifiers with 95% confidence. However, there is a significant difference in the recall ($p = 0.03$, $\alpha = 0.05$) of the classifiers. Specifically, the VSM based classifier outperforms the LSA based classifier in this context and configuration.

Both the VSM and LSA based versions of the classifier correctly classified publications better than chance, thus this study finds in the affirmative with regard to the first research question (RQ1): The VSM family of algorithms can be utilized to classify publications for inclusion/exclusion in the context of secondary SE research.

In response to the second research question components, this study finds the following:

RQ2a: Recall of minimally tuned VSM family classifier is approximately 0.79.

RQ2b: Precision of a minimally tuned VSM family classifier is approximately 0.32.

RQ2c: Work savings of a minimally tuned VSM family classifier can be as high as 90%.

4.6 DISCUSSION

The superior performance of the basic VSM model in this study was unexpected based on the literature. It had been anticipated that the LSA classifier would demonstrate clear advantages over the basic VSM classifier. As the LSA algorithm is a derivative of the VSM algorithm, it is natural to anticipate the performance of the LSA algorithm is at least equal to the VSM algorithm. It is clear, based on this investigation, that the tuning of the LSA algorithm is critical to its performance and may actually be detrimental to the performance of the classifier. As the
understanding of the LSA algorithm is in its infancy, addition research to understand the impact of the various tunable points (preprocessing, weighting, dimensionality, and measure) is needed.

The need for better understanding of what influences the algorithm can also be said of the basic VSM algorithm. While the VSM algorithm lacks the SVD procedure and thus the issue of dimensional reduction, the impact of preprocessing, weighting, and selection of measurement method remains largely unexplored to the best of the author’s knowledge. These three items along with constraints across the larger dataset have significant impact on the content of the individual vectors that represent the documents of a corpus. Thus they must also influence the comparison of the vectors and the outcome of the comparison.

The use of a VSM or LSA based classifier also opens new avenues of approach to continued maintenance of an SLR. It provides a means by which publications can be scanned in real-time to identify the need for updating based on either a given mass of publications accumulating or a disruptive shift in research findings. With this capability however, come new questions. Will the domain of interest remain confined within the original constraints, or will it slowly expand as documents are added to the collection? Is a periodic “intervention” needed to keep the selection of studies confined to criteria defined by the study, or should it be allowed to grow organically? Further research is needed to understand how entropy and scope-creep may factor into the evolution of a publication set over time, especially when the classifier immediately adds studies identified for inclusion into its training set.

4.6.1 Implications

The findings of this study carry several implications for the SE community. This study demonstrates the feasibility of utilizing the VSM and LSA algorithms for the purpose of screening academic publications for content based on known examples of interest. The outcome
of such ability is considerable time savings and a more consistent application of criteria regarding what is of interest and what is not. This will reduce the time required of researchers producing new SLRs for the community and reduce the information overload for practitioners seeking guidance from academic literature.

Likewise, the application of the VSM or LSA classifier will ease the maintenance and updating of SLRs moving forward. By leveraging the inclusion/exclusion decisions made in previous versions of an SLR, the time required for selection of relevant articles from the continuous stream of academic publications is drastically reduced. Thus researchers can produce updated SLRs more frequently with less effort.

In the context of a mapping study such as the one from which the testing dataset for this study is drawn, the results of this study are remarkable with regard to the work savings versus accuracy tradeoff. As mapping studies are designed to map out a domain of research and are therefore more tolerant of lower recall statistics, the performance of the VSM classifier is acceptable. One must consider what level of recall is truly necessary in a given situation. If the risk of injury or loss is a major consequence, addition tuning to achieve higher levels of recall may be required.

The timeline for the production of a review may similarly be a consideration in the recall of the classifier. Given the short timespans often allotted for due-diligence in a business environment, the ability to automate the search and selection portion of the review process offers researchers the opportunity to participate more fully in the business decision as the time allotted can be utilized for analysis and synthesis instead of information search and culling. The timeline for reviews can be reduced from months or years to days or weeks while maintaining an acceptable level of risk with regard to the information utilized in the decision process.
4.6.2 Limitations

Generalization of the results of this study must be approached with caution due to several factors. Specifically, the dataset utilized in the testing and the current understanding of the LSA algorithm present obstacles to generalizations.

The dataset utilized in testing is from a very broad mapping study in which there are several disjoint, but interacting primary topics of interest. Thus the classifier is seeking evidence of one or more topics from among a group of topics of interest in order to include a study. This variety of content may contribute to the classifier matching a study that may have been excluded without the presence of a syntactically similar, yet semantically dissimilar topic. Based on what is known of the LSA model and the comparison measure utilized in the study, one can also speculate that as the number of semantic categories of interest is reduced, the recall and precision performance of the classifiers will improve. Further testing with additional datasets from a variety of topic areas is needed to understand and address this concern.

Similarly, bias in the original classification of publications for the mapping study may have impacted the articles included and thus the training sets and results of this study. It is believed that this is a minimal risk based of the protocol of the mapping study and its execution, however one must cognizant of the risk regardless.

The other primary factor to be considered in the interpretation is the lack of a theoretical basis for the LSA algorithm. Exactly how and why the algorithm displays the characteristic that it does in text processing is still open to debate in the literature and largely a-theoretical. Most of the factors believed to influence LSA performance are based on ad-hoc experiences of researchers attempting to apply the algorithm in a given domain for the purpose of textual analysis. The same is true for the VSM algorithm. While one would anticipate the VSM
algorithm to have foundations in the field of discourse analysis, literature searches have thus far failed to uncover such a study. Until such time that a theoretical basis for this family of algorithms is established, the application of the results of this study in new domains of interest must be approached on a case-by-case basis.

4.7 Conclusion and Future Work

This study demonstrates the feasibility of utilizing VSM family algorithms to classify academic literature for inclusion or exclusion in the context of secondary research in the SE domain. The use of this family of algorithms can produce work savings of as much as 90% with an average recall 0.79 and precision of 0.32. This allows for considerable time and resource savings not only in terms of researchers’ time, but also the effort needed to select studies based on the consistent application of criteria. Future research will focus on improving the recall and precision of the classifier through a better understanding of the algorithms in the VSM family.

Additional research is also planned to expand upon the generalizability of the results through the testing of the classifiers in additional SMS and SLR study contexts. In the spirit of grounded theory building, this will assist in development of theory as to how and why the VSM family of algorithms work, the boundaries in which it works, and allow for empirically founded tuning of the classifier algorithm.

4.8 References


Kitchenham, Barbara, and Stuart Charters. 2007. Guidelines for performing systematic literature reviews in software engineering. Keele University and Durham University.


CHAPTER 5

Conclusion

In the spirit of Fredrick Taylor’s *The Principles of Scientific Management* (1914), the collection of essays presented in this work endeavor to advance the use of empirically based, juried evidence for decision making in a business context. The specific context selected is the domain of SE – a key component in a technology laden world. Within the SE domain, the essays address three objectives in the advancement of the EBSE paradigm by:

7. Mapping the research completed to date regarding the implementation of EBP in the SE domain to identify gaps and opportunities in the research.

8. Identifying the barriers deemed most important by the members of the SE research community who conduct systematic literature reviews in support of EBSE.

9. Developing the use of algorithmic techniques as a discriminant function in the selection process of the systematic review methodologies.

Together, the collection of essays represent a line of inquiring within a broader research stream concerning the implementation of EBP – a modern version of Taylors work – within the SE domain. The collection of essays provides valuable insights concerning the status of EBSE and its literature, the problems associated with secondary research under the paradigm, and the basis for a discrimination function designed to assist in resolving a key issue for those seeking guidance in academic literature.
5.1 The EBSE Landscape

The systematic mapping study of the EBSE literature shows that the core literature concerning EBSE is uniquely identifiable, but surrounding this literature core is a fuzzy boundary where EBSE literature begins to blend with other areas of research. Researchers must be cognizant of this fact in order to capitalize on all available research.

A number of avenues for additional research are also identified by the gaps noted in the mapping study. Two of the most under-researched areas of the EBSE paradigm are the structure of the practice environment and the process utilized by practitioners to acquire, vet, and employ juried evidence.

The structural model – which is utilized in training new practitioners in other domains that have adopted the EBP paradigm – has limited support in the existent EBSE literature. Support is limited to a descriptive nature reinforcing conceptual reasoning. The model itself is completely absent. Additional research is needed to determine if this absence is due to an oversight in the research or if a different model should be utilized. Regardless, based on the presence of its components the introduction of the model into the literature would serve to start the conversation concerning the appropriateness it.

The introduction of a structural model may also assist in the investigation of teaching EBSE to new software. To date the publications regarding the teaching of EBSE are focused primarily of the SLR/SMS protocols. While these protocols are meticulous in producing evidence grounded in empirical research, they are also not the parsimonious, actionable models needed for practicing software engineers. The resources required to support such protocols are often not available in day-to-day operations, thus a more parsimonious, actionable process is needed.

Similarly, the EBSE process model for decision making is under-investigated in the existing literature population. As the model is described in many of the conceptual studies present in the
literature, it is unclear as to whether the model has simply been adopted “as-is” or ignored by researchers.

Another distinct gap in the existing literature is the absence of work regarding feedback from practice from an EBSE viewpoint. In the decade since the inception of the EBSE paradigm, considerable progress has been made by researchers in their efforts to amalgamate empirically based evidence in support of SE practice. Nearly 300 SLRs are available to support decision making in the practice environment. However, this study also suggests that work to provide feedback from practice into research is needed.

While there appears to be empirical support for the utilization of the EBSE paradigm, that support is limited – especially within the models most closely associated with practice. No quantitative based studies supporting either the structural model for practice or the EBSE process view were found. The empirical work that does exist is primarily of a qualitative nature. Given that descriptive research must be developed before normative or predictive theory can be built, this finding is consistent with the development of new theory within a domain. Similarly, the continuing emergence of studies to refine the SLR/SMS protocols and the infrastructure, tools and techniques to support EBSE and the EBSE ecosystem are reflective of continuing refinement of descriptive research and the initiation of normative work.

5.2 Barriers in the SLR Process

Based on the substantial body of literature regarding the EBSE subdomain of systematic reviews uncover in the mapping study, the investigation proceeded to the examination of problems associated with the production of research based evidence for use in practice. Specially, identifying the barriers deemed most important by the members of the SE research community that conduct SLRs.
Workshop discussions provided additional evidence about the barriers in the production process of SLRs which were identified in the literature uncovered during the mapping study. Additionally, new barriers were identified that had previously been unknown. A global analysis, independent of specific SLR phases, of the barriers generated by participants revealed four high-level latent themes among the barriers:

- Barriers Associated with the SLR Process
- Barriers Associated with Primary Studies
- Barriers Associated with the Practitioner Community
- Barriers Associated with Tooling

The voting results of the workshop provide prioritization for research and tool development that will maximize the benefits of such efforts. Specifically, high priority items are to:

- Reduce the time and effort required to search for and select primary studies
- Improve the assessment of the quality of primary studies
- Synthesize primary studies that differ in design or context

Work in these high-priority areas will permit more timely production of SLR’s, improve the quality of the SLR process, and increase researchers’ abilities to triangulate knowledge and thus to advance understanding of a given domain.

5.3 Addressing a Priority Issue of the SLR Process

In the final portion of this investigation, one of the high priority barriers previously identified – the time and resources needed to select study for inclusion in a review – is addressed. Two potential algorithms for automating the selection process are identified and then evaluated utilizing the set of studies uncovered during the mapping portion of the research.
The feasibility of utilizing the VSM and LSA algorithms to discriminate between academic studies to be included or excluded in SE domain secondary research is demonstrated. The use of these algorithms can produce work savings of as much as 90% with an average recall 0.79 and precision of 0.32. This allows for considerable time and resource savings not only in terms of researchers’ time, but also the effort needed to select studies based on the consistent application of criteria. This technique as holds promised for members of the practice community seeking a means to cull the ever-growing deluge of information available.

The significant time savings will allow researchers the opportunity to participate more fully in the business decision as the time allotted can be utilized for analysis and synthesis instead of information search and culling. The timeline for reviews can be reduced from months or years to days or weeks while maintaining an acceptable level of risk with regard to the information utilized in the decision process.

Likewise, the application of the VSM or LSA classifier will ease the maintenance and updating of SLRs in the future. By leveraging the inclusion/exclusion decisions made in previous versions of an SLR, the time required for selection of relevant articles from the continuous stream of academic publications is drastically reduced. Thus researchers can produce updated SLRs more frequently with less effort.

5.4 Continuing and Future Work

A version of the work presented in the second essay (Chapter 3) has already been published at the 18th International Conference on Evaluation and Assessment in Software Engineering (EASE) and is available through the Association for Computing Machinery (ACM) digital library. A second community workshop has also been held to validate the findings of this initial effort and extend it by collecting data on the desired features of a tool to support the SLR
process. Work in currently in progress to complete the manuscript presenting the results of workshop. The target journal for publication is Information and Software Technology.

At a more macro level, the work presented in this manuscript serves as a roadmap for advancing the adoption of EBSE in both research and practice. It identifies specific gaps in the literature relating directly to practice and shows where synthesis and additional research are needed in relation to the research side of the EBSE paradigm. It indicates that closer ties between research and practice are needed, and that there are opportunities to encourage the adoption of EBSE in the education of those entering the field.

Finally, the results of this investigation provide a basis for both researchers and practitioners to reduce the effort needed to acquire empirically founded, juried evidence on which to base decisions. For researchers, this means improved support in the production of SLRs. For practitioners, it is hoped that the advancements made in improving the SLR production process can be leveraged as part of the due-diligence process utilized in business decisions.
REFERENCES


APPENDIX A

IRB Certificate

April 3, 2014

Jeffrey Carver, Ph.D.
Department of Computer Science
College of Engineering
The University of Alabama

Re: IRB # 13-OR-256 (Revision # 2) “Advanced Systematic Literature Review for Software Engineering”

Dear Dr. Carver:

The University of Alabama Institutional Review Board has reviewed the revision to your previously approved expedited protocol. The board has approved the change in your protocol.

Please remember that your approval period expires one year from the date of your original approval, July 31, 2013, not the date of this revision approval.

Should you need to submit any further correspondence regarding this proposal, please include the assigned IRB application number. Changes in this study cannot be initiated without IRB approval, except when necessary to eliminate apparent immediate hazards to participants.

Good luck with your research.

Sincerely,

Stuart Usdan, Ph.D.
Chair, Non-Medical Institutional Review Board
The University of Alabama
UNIVERSITY OF ALABAMA

INSTITUTIONAL REVIEW BOARD FOR THE PROTECTION OF HUMAN SUBJECTS
REQUEST FOR APPROVAL OF RESEARCH INVOLVING HUMAN SUBJECTS

I. Identifying information

Principal Investigator
Names: Dr. Jeffrey C. Carver

Second Investigator
Dr. Nicholas A. Kraft

Third Investigator

Department: Computer Science

Engineering

University: Alabama

Address: Box 870290

Phone: 8-9829

Fax: 8-0219

E-mail: carver@cs.ua.edu

Title of Research Project: Advanced Systematic Literature Review Infrastructure for Software Engineering

Date Submitted: March 13, 2014

Funding Source: National Science Foundation

Type of Proposal

☐ New

☐ Revision

☐ Renewal

☐ Completed

☐ Exempt

Please attach a renewal application

Please attach a continuing review of studies form

Please enter the original IRB # at the top of the page

II. NOTIFICATION OF IRB ACTION (to be completed by IRB):

Type of Review: Full board Expedited

IRB Action:

☐ Rejected

☐ Tabled Pending Revisions

☐ Approved Pending Revisions

☐ Approved, this proposal complies with University and federal regulations for the protection of human subjects.

Approval is effective until the following date: 7-30-14

Items approved:

☐ Research protocol (dated )

☐ Informed consent (dated )

☐ Recruitment materials (dated )

☐ Other (dated )

Approval signature ___________________________ Date 4-3-14