Strength of Evidence in Systematic Reviews in Software Engineering

Tore Dybå
SINTEF ICT
NO-7465 Trondheim, Norway
Tel. +47 73 59 29 47
tore.dyba@sintef.no

Torgeir Dingsøyr
SINTEF ICT
NO-7465 Trondheim, Norway
Tel. +47 73 59 70 72
torgeir.dingsoyr@sintef.no

ABSTRACT
Systematic reviews are only as good as the evidence they are based on. It is important, therefore, that users of systematic reviews know how much confidence they can place in the conclusions and recommendations arising from such reviews. In this paper we present an overview of some of the most influential systems for assessing the quality of individual primary studies and for grading the overall strength of a body of evidence. We also present an example of the use of such systems based on a systematic review of empirical studies of agile software development. Our findings suggest that the systems used in other disciplines for grading the strength of evidence for and reporting of systematic reviews, especially those that take account of qualitative and observational studies are of particular relevance for software engineering.

Categories and Subject Descriptors
D.3 [**Software Engineering**]

General Terms
Management, Measurement, Experimentation, Theory.

Keywords
Systematic Review, Quality Assessment, Strength of Evidence.

1. INTRODUCTION
Systematic reviews are currently gaining popularity in software engineering, with reviews recently published on diverse topics including requirement elicitation [11], web engineering [40], cost estimation [27], and agile software development [15], as well as various aspects related to software engineering experiments, such as statistical power [17], theory use [22], effect size [28], and quasi-experimentation [29], Such reviews are important, as the volume of research that needs to be considered by software engineering (SE) practitioners and researchers is constantly expanding.

In many areas it has become almost impossible for the individual to read, critically evaluate, and synthesize the state of current knowledge, let alone keep updating this on a regular basis. As a result, reviews have become essential tools for anyone who wants to keep up with the new evidence that is accumulating in his or her fields of interest. Reviews are also required to identify areas where the available evidence is insufficient and further studies are required. However, because of the often poor quality of traditional narrative reviews, there has recently been an increasing focus on formal methods of systematically reviewing studies [9], [23], [30], [32], [43], [45].

Systematic reviews (SRs) evaluate and interpret the available research relevant to a particular research question, topic area, or phenomenon of interest [4], [32]. This is useful, in evidence-based software engineering (EBSE), which aims to improve decision making related to software development and maintenance by integrating current best evidence from research with practical experience and human values [18], [34]. This is an ambitious aim, particularly because the gap between research and practice can be wide. EBSE seeks to close this gap by encouraging a stronger emphasis on methodological rigor while focusing on relevance for practice.

SRs are a key tool for enabling evidence-based practice as they bring together, and combine, the findings from multiple studies. The quality of the primary studies in software engineering, however, is often poor [35], [47]. A central issue is then how much confidence we can place in the conclusions and recommendations arising from systematic reviews.

In this paper we present an overview of some of the most influential systems for assessing the quality of individual primary studies and for grading the overall strength of a body of evidence. We present an example of the use of such systems based on a systematic review of empirical studies of agile software development [15]. The remainder of this paper is organized as follows: Section 2 described systematic reviews in more detail, Section 3 presents methods to assess the quality of primary studies, Section 4 concerns grading the strength of a body of evidence, and Section 5.presents measures to assess the quality of systematic reviews. Finally, Section 6 concludes with recommendations for carrying out systematic reviews in software engineering.

2. WHAT ARE SYSTEMATIC REVIEWS
A systematic review is a concise summary of the best available evidence that uses explicit and rigorous methods to identify, critically appraise, and synthesize relevant studies on a particular topic. These methods are defined in advance and documented in a protocol so that others can critically appraise and replicate the review.

ESEM'08, October 9–10, 2008, Kaiserslautern, Germany.

Table 1. Stages of the SR process [32]

1. Planning the review a. Identification of the need for a review b. Development of a review protocol
2. Conducting the review a. Identification of research b. Selection of primary studies c. Study quality assessment d. Data extraction e. Data synthesis
3. Reporting the review

The strength of the SR methods lies in their explicit attempt to minimize the chances of drawing wrong or misleading conclusions as a result of biases in primary studies or from biases arising from the review process itself [19], [43], [45]. This is the key feature that distinguishes SRs from traditional narrative reviews.

SRs of multiple studies help establish whether scientific findings are consistent and can be generalized across populations, settings, and treatment variations, or whether findings vary by particular subsets. SRs can also identify crucial questions that have not been adequately addressed with past empirical research.

Kitchenham and Charters [32] describe a systematic review process in three stages as in Table 1. We give a brief overview of the first two steps, and relate the steps to experience reports from research groups who have carried out systematic reviews.

Planning the review includes defining explicit inclusion and exclusion criteria, which specify the types of study designs, interventions, populations, and outcomes that will be included in the review. A systematic search strategy [12] specifies the keyword strings and sources used to find relevant studies in bibliographic databases and other electronic sources. Several databases must usually be searched in order to assure good coverage on the topic [2]. It may also be necessary to search key journals and conference proceedings by hand to identify relevant studies that are not fully indexed. The search may be bounded by dates, journals, databases, and so forth, as long as the search procedures are transparent and replicable.

Staples and Niazi [49] discuss their use of an earlier version of Kitchenham and Charter's guidelines [32] for systematic review, and recommend using clear and narrow research questions when carrying out reviews. Brereton *et al*. [6] discuss problems encountered when doing systematic reviews, and in particular found problems with the quality of the software engineering indexing databases.

In the second stage, when conducting the review, decisions about full text retrieval, study eligibility, and coding are most often made by two independent reviewers to increase reliability. Similarly, data from primary outcome studies are extracted by independent reviewers onto paper or electronic forms. These data typically include characteristics of the study design, interventions, sample, outcome measures, and findings.

Dybå *et al*. [16] discussed challenges with assessing and synthesizing qualitative studies, using Noblit and Hare's [44] seven-step process for meta-ethnography. Noblit and Hare further distinguish between integrative and interpretive reviews.

Integrative reviews are concerned with combining or summarizing data for the purpose of creating generalizations [10]. It involves techniques, such as meta-analysis, that are concerned with assembling and pooling well specified data, or less formal techniques, such as providing a descriptive account of the data.

Interpretive reviews, on the other hand, achieve synthesis through subsuming the concepts identified in the primary studies into a higher-order theoretical structure (meta-ethnography). The primary concern is with the development of concepts and theories that integrate those concepts. An interpretive synthesis will therefore avoid specifying concepts in advance of the synthesis, but rather ground concepts in the data reported in the primary studies [14].

Noblit and Hare suggest that integrative reviews are primarily suitable for synthesizing quantitative studies, while interpretive reviews are more suitable for synthesizing qualitative studies [44]. Still, whilst most forms of synthesis can be characterized as being either primarily interpretive or primarily integrative, every integrative synthesis will include elements of interpretation, and every interpretive synthesis will include elements of aggregation of data. The nature of the primary studies will affect how the quality of the primary studies is evaluated, and also how the overall strength of the body of evidence is graded.

3. ASSESSING THE QUALITY OF PRIMARY STUDIES

Quality assessment of primary studies is necessary to limit bias in conducting the systematic review, to gain insight into potential comparisons, and to guide interpretation of findings [23], [32].

Assessing the quality of a study is not straightforward, however, as there is no general, agreed upon definition of "quality." There are also common problems in appraising the quality of published research as journal articles and, in particular, conference papers rarely provide enough detail of the methods used due to space limitations in journal volumes and conference proceedings. There is therefore a danger that what is being assessed is the quality of reporting rather than the quality of research.

In general, the "quality" of a study is closely linked to the research methods used and the validity of the findings generated by the study. In this sense, quality refers to the extent to which the design, conduct, and analysis of the primary studies are likely to prevent systematic errors or bias [23]. Since biased primary studies are more likely to provide misleading results, they are also likely to generate misleading systematic reviews. Therefore, it is important that reviewers critically appraise the methods of all primary studies.

3.1 Issues of Validity and Bias

In the context of a systematic review, the validity of a study is the extent to which its design and conduct are likely to prevent systematic errors, or bias [23]. The validity of a study may be considered to have two dimensions. The first is whether the study is asking an appropriate research question. This is often described as "external validity", and its assessment depends on the purpose for which the study is to be used. External validity is closely connected with the generalizability or applicability of a study's findings. The second dimension of a study's validity relates to whether it answers its research question in a manner free from

bias. This is often described as "internal validity", and it is this aspect of validity that is particularly relevant for assessing the quality of primary studies.

A bias is a systematic error, or deviation from the "truth", in results or inferences. Biases can vary in magnitude and direction: Some are small and trivial, others are substantial; some lead to underestimation while others lead to overestimation of the true intervention effect. There is also evidence that particular flaws in the design, conduct and analysis of empirical studies lead to bias [23]. There are particularly three methodological features that have been shown to influence the results of primary studies: randomization, blinding, and loss to follow-up [25], [29].

The aim of randomization is to avoid selection bias by making sure that each subject in the study has an equal chance of getting into each treatment group. In addition, steps must be taken to conceal the treatment sequence in order to preventing fore-knowledge of the forthcoming allocations.

We concur with Laitenberger and Rombach [37] who claim that quasi-experiments (in which study units are assigned to experimental groups non-randomly) represent a promising approach to increasing the amount of empirical studies in the SE industry. We also concur with Kitchenham [31] who suggests that researchers in SE need to become more familiar with the variety of quasi-experimental designs, because they offer opportunities to improve the rigor of large-scale industrial studies.

However, a recent systematic review of quasi-experiments in SE emphasizes the importance of paying attention to selection bias, and showed that quasi-experiments might lead to results other than those of randomized experiments [29]. It is important, therefore, that quasi-experimental studies are well designed and analyzed to control for selection bias.

Blinding of study participants and personnel may reduce the risk that knowledge of which treatment was received, rather than the treatment itself, affects outcomes and outcome measurements [25]. Blinding can be especially important for assessment of subjective outcomes, and can also ensure that treatment groups receive a similar amount of attention. Blinding may also be important in studies in which enthusiasm for participation or follow-up may be influenced by group allocation [23].

However, adequate blinding, in order to reduce experimenter and subject bias, is generally impossible in SE studies that rely on subjects performing human-intensive tasks [34]. Still, we can use blinding in a number of ways to reduce the opportunity for bias by reducing the direct interaction between subjects and experimenters during the course of an experiment. Techniques to consider include blind allocation of material, blind marking, and blind data collection and analysis [34], [35].

Missing outcome data, due to attrition (withdrawal and dropout) during the study or exclusions from the analysis, raise the possibility that the observed effect estimate is biased. However, if outcome data are missing in all treatment groups, but reasons for these are both reported and balanced across groups, then important bias would not be expected unless the reasons have different implications in the compared groups [23].

There are other sources of bias that are relevant only in certain circumstances. Some can be found only in particular research designs; some can be found across a broad spectrum of designs, but only for specific circumstances; and there may be sources of bias that are only found in a particular industrial setting.

Lee [38] and Klein and Myers [36], for example, discuss issues related to the quality of case studies and interpretative case studies. Tactics to reduce bias in such studies include:

- Multiple data sources and triangulation.
- Natural controls by using what is happening in the case to check rival theories.
- Member checking by presenting results to informants to reveal misunderstandings by the researchers.
- Use of alternative theories to explain the data from the case.

A final concern is biased replications. Although we encourage replications of important SE studies, such replications should preferably be conducted by others than those conducting the original study in order to avoid bias. An example of the importance of such independent replications is the 15 differential replications of controlled experiments in SE identified by Sjøberg *et al*. [48]. Among these replications, seven of eight replications carried out by the same authors confirmed the results of the original experiments, while only one of seven replications carried out by other researchers confirmed the original results. This finding might be explained by studies that evaluate the ability to learn from experience, which have demonstrated biases that prevent people from using the information provided by such experience. Such biases include preferences for confirmatory evidence, assumptions about causality, and a disregard of negative information [5].

For all potential sources of bias it is important to consider the likely magnitude and direction of the bias. A useful classification of bias is into selection bias, performance bias, attrition bias, detection bias, and reporting bias (**Error! Reference source not found.**).

3.2 Assessment Tools and Checklists

There are a multitude of guidelines, tools, and checklists that can be used in assessing the quality of primary studies. Some of these build on the main threats to validity in experimental and quasi-experimental designs identified by Campbell and colleagues [7], [8], [46] while others focuses on the methodological characteristics of the study, e.g., [21] and [32].

One of the most well-known and widely used scales for assessing randomized controlled trials is the scale developed by Jadad *et al*. [25]. This scale reflects the main biases mentioned in the previous section and consists of only three, but important, items:

Table 2. Types of bias [23]

Selection bias: Systematic differences between the groups that are compared.
Performance bias: Systematic differences between groups or in exposure to factors other than the treatment.
Attrition bias: Systematic differences between groups in withdrawals from a study.
Detection bias: Systematic differences between groups in how outcomes are determined.
Reporting bias: Systematic differences between reported and unreported findings.

- Was the study described as randomized?
- Was the study described as double blind?
- Was there a description of withdrawals and dropouts?

Another widely used tool in the medical community is the CONSORT statement[1] for improving the quality of reports of parallel-group randomized trials [41]. The CONSORT statement consists of 22 checklist items pertaining to the content of the title, abstract, introduction, methods, results, and discussion sections of a paper. The statement also includes a flow diagram that depicts information from four stages of a trial: enrollment, intervention allocation, follow-up, and analysis.

In addition to a tool for assessing randomized controlled trials, the Critical Appraisal Skills Programme (CASP)[2], at the Public Health Resource Unit in Oxford, has developed a set of tools to help with the process of critically appraising articles of the following types of research: Systematic Reviews, Qualitative Research, Economic Evaluation Studies, Cohort Studies, Case Control Studies, and Diagnostic Test Studies.

Sjøberg *et al*. [47] discuss measures to increase the quality of empirical studies in SE in general, while Kitchenham *et al*. [35] have proposed a set of more concrete guidelines. These guidelines are intended to assist researchers, reviewers, and meta-analysts in designing, conducting, and evaluating empirical studies, and are based on a review of research guidelines developed for medical researchers and the authors' own experience in doing and reviewing SE research.

Based on a survey of guidelines from other disciplines, Jedlitschka and Pfahl [26] proposed a set of reporting guidelines for controlled experiments in SE. Recently, Kitchenham *et al*. [33] evaluated these guidelines and concluded that the current guidelines need to be revised and then subjected to further theoretical and empirical validation. Höst and Runeson [24] have suggested a checklist to use in case studies in SE, while the recent special issue of *Information and Software Technology* on qualitative SE research [13] provides several useful examples of approaches for study designs, data collection, and analysis that should be relevant for increasing the quality of qualitative studies.

3.3 Example of Quality Assessment of SE Studies

An example of the use of quality assessment of primary studies in SE is our systematic review of empirical studies of agile software development [15]. As part of that systematic review, both authors independently assessed the quality of the primary studies according to eleven criteria (Table 3). These criteria were informed by CASP and by principles of good practice for conducting empirical research in SE proposed by Kitchenham *et al*. [35]. The eleven criteria used to assess the quality of the studies covered four main issues:

- **Reporting**: Three criteria (1-3) were related to the quality of the reporting of a study's rationale, aims, and context.
- **Rigor**: Five criteria (4-8) were related to the rigor of the research methods employed to establish the validity of data collection tools and the analysis methods, and hence the trustworthiness of the findings.
- **Credibility**: Two criteria (9-10) were related to the assessment of the credibility of the study methods for ensuring that the findings were valid and meaningful.
- **Relevance**: The final criterion (11) was related to the assessment of the relevance of the study for the software industry at large and the research community.

Taken together, these criteria provided a measure of the extent to which we could be confident that a particular study's findings could make a valuable contribution to the review. A general observation was that we frequently found that methods were not well described; issues of bias, validity, and reliability were not always addressed; and methods of data collection and analysis were often not explained well.

The grading of each of the eleven criteria was done on a dichotomous scale with "1" awarded to a study when a question could be answered as "yes" and "0" when the answer was "no". The overall results of the quality assessment for the 33 studies included in the SR are shown in the parentheses in Table 3. As we only included research papers in the review, all studies were rated as OK on the first screening criterion. However, two of the included studies still did not have a clear statement of the aims of the research. All studies had some form of description of the context in which the research was carried out. For three of the studies, the chosen research design did not seem appropriate to the aims of the research. As many as 25 out of the 33 primary studies did not have a recruitment strategy that seemed appropriate for the aims stated for the research. Ten of the studies included one or more groups with which to compare agile methods. Seven and eight studies, respectively, did not adequately describe their data collection and data analysis procedures. In only one study was the recognition of any possibility of researcher bias mentioned.

Table 3. Quality assessment of 33 SE studies [15]

1.	Is the paper based on research (or is it merely a "lessons learned" report based on expert opinion)? (33)
2.	Is there a clear statement of the aims of the research? (31)
3.	Is there an adequate description of the context in which the research was carried out? (33)
4.	Was the research design appropriate to address the aims of the research? (30)
5.	Was the recruitment strategy appropriate to the aims of the research? (8)
6.	Was there a control group with which to compare treatments? (10)
7.	Was the data collected in a way that addressed the research issue? (26)
8.	Was the data analysis sufficiently rigorous? (25)
9.	Has the relationship between researcher and participants been adequately considered? (1)
10.	Is there a clear statement of findings? (33)
11.	Is the study of value for research or practice? (33)

[1] http://www.consort-statement.org
[2] http://www.phru.nhs.uk/Pages/PHD/CASP.htm

4. GRADING THE STRENGTH OF A BODY OF EVIDENCE

Several systems exist for making judgments about the strength of evidence in systematic reviews (see [20] for an overview). Most of these systems suggest that the strength of evidence can be based on a hierarchy with evidence from systematic reviews and randomized experiments at the top of the hierarchy and evidence from observational studies and expert opinion at the bottom of the hierarchy [36]. The inherent weakness with evidence hierarchies is that randomized experiments are not always feasible and that, in some instances, observational studies may provide better evidence.

4.1 The GRADE Approach

To cope with the weaknesses of evidence hierarchies, the Grades of Recommendation Assessment, Development and Evaluation (GRADE) Working Group has developed a system for grading the quality of evidence and strength of recommendations [20]. Several organizations, including the World Health Organization (WHO), have adopted the GRADE system, either in its original format or with minor modifications. The British Medical Journal (BMJ) encourages their authors to use the GRADE system and, recently, the Cochrane Collaboration has adopted the principles of the GRADE system for evaluating the quality of evidence for outcomes reported in systematic reviews [23].

The GRADE approach defines the quality of a body of evidence as the extent to which one can be confident that an estimate of effect or association is correct. Judgments about the quality of evidence involve considerations of the validity of the results of individual primary studies.

The GRADE approach specifies four grades of evidence (Table 4). The highest quality evidence comes from one or more well-designed and well-executed randomized controlled trials (RCTs) yielding consistent and directly applicable results. However, RCT evidence can be "downgraded" to moderate, low, or even very low quality evidence, depending on the seriousness of the limitations of the study. If there are very severe problems, RCT evidence may fall by more than one level. Evidence from sound observational studies will generally be graded as low quality. If, however, such studies yield large effects and there is no obvious bias explaining those effects, review authors may rate the evidence as moderate or even – if the effect is large enough – high quality.

The very low quality category includes poorly controlled observational studies and unsystematic clinical observations (e.g. case series or case reports). The system also clarifies that expert opinion is not a category of evidence, but rather represents an interpretation of evidence.

4.2 Factors that Decrease the Strength of Evidence

According to GRADE, the following factors may decrease the strength of a body of evidence (Table 5):

1. **Limitations to study quality**. Our confidence in an estimate of effect decreases if the studies have major deficiencies that are likely to result in a biased assessment of the treatment effect. For controlled experiments, these methodological limitations include lack of randomization, lack of allocation concealment, lack of blinding with subjective outcomes highly susceptible to bias, or large loss to follow-up.
2. **Inconsistency of results**. When studies yield important unexplained inconsistency in the results, our confidence in the estimate of effect for that outcome decreases. Such inconsistencies include differences in the direction of effect, the size of the differences in effect, and the significance of the differences.
3. **Uncertainty about directness**. Directness refers to the extent to which the subjects, settings, treatments, and outcome measures are similar to those of interest. Uncertainty about directness arises, e.g., when the subjects of interest are experienced professionals in industry while those in the studies are novices or students at universities. Indirectness may also apply to the treatments and outcomes when, e.g., small artificial tasks are used in the studies, while the systems of interest are large-scale industrial systems, or when defect counts are used as a surrogate for product quality.
4. **Imprecise or sparse data**. Data are considered sparse if the results include just a few subjects or observations. Data are imprecise if the confidence intervals are sufficiently wide that an estimate is consistent with either important net benefits or harms and thus consistent with divergent recommendations.

Table 4. Strength of evidence in the GRADE system [20]

High	Further research is very unlikely to change our confidence in the estimate of effect.
Moderate	Further research is likely to have an important impact on our confidence in the estimate of effect and may change the estimate.
Low	Further research is very likely to have an important impact on our confidence in the estimate of effect and is likely to change the estimate.
Very low	Any estimate of effect is very uncertain.

Table 5. Factors that may decrease or increase the strength of evidence [20]

Factors that may decrease the strength of evidence:

- Serious (–1) or very serious (–2) limitations to study quality
- Important inconsistency (–1)
- Some (–1) or major (–2) uncertainty about directness
- Imprecise or sparse data (–1)
- High probability of reporting bias (–1)

Factors that may increase the strength of evidence:

- Strong evidence of association (RR > 2 or RR < 0.5) based on consistent evidence from two or more observational studies, with no plausible confounders (+1)
- Very strong evidence of association (RR > 5 or RR < 0.2) based on direct evidence with no major threats to validity (+2)
- Evidence of a dose response gradient (+1)
- All plausible confounders would have reduced the effect (+1)

RR = relative risk

5. **Reporting bias**. The strength of evidence may be reduced if researchers fail to report studies or outcomes on the basis of result. Reporting bias also includes selected reporting of subgroup analyses or adjusted analyses. A prototypical situation that may elicit suspicion of reporting bias is when published evidence includes a number of small studies, all of which are industry funded [3].

These factors act cumulatively, e.g., if randomized experiments have both serious limitations and there is uncertainty about the directness of the evidence, the grade of evidence would drop from high to low.

4.3 Factors that Increase the Strength of Evidence

Although observational studies will generally yield low-quality evidence, the presence of certain factors could 'upgrade' such evidence as moderate or even high quality (Table 5):

1. **Strong or very strong evidence of association**. When methodologically strong observational studies yield large or very large, consistent and precise estimates of the magnitude of a treatment effect, one may be particularly confident in the results. The magnitude of the effect in these studies may move the assigned grade of evidence from low to moderate if the effect is large in the absence of other methodological limitations, or grade the quality as high if the effect is very large in the absence of other methodological limitations.
2. **Presence of a dose-response gradient**. The presence of a dose-response gradient may also increase our confidence in the findings of observational studies and thereby increase the strength of evidence.
3. **Underestimation of effect**. On occasions, all plausible biases from observational studies or experiments may be working to underestimate an apparent treatment effect. For example, if only novice developers receive an experimental treatment, yet they still fare better, it is likely that the actual treatment effect is larger than the data suggest. For instance, in a large quasi-experiment of pair programming conducted by Arisholm *et al*. [1] it is possible that the benefits of pair programming will exceed the results obtained in this experiment for larger, more complex tasks and if the pair programmers have a chance to work together over a longer period of time.

4.4 Application of the GRADE System in SE

To the best of our knowledge, there is only one example, so far, on the use of the GRADE system to grade the strength of a body of evidence in SE. In the systematic review of empirical studies of agile software development [15], we applied GRADE to rate the overall strength of evidence. We combined the four basic components of study design, study quality, consistency, and directness to grade the strength of the evidence regarding the benefits and limitations of agile methods, and for decisions related to their adoption.

Regarding study design, there were only three experiments in the review (two randomized trials), while the remaining studies were observational. Consequently, our initial categorization of the total evidence in the review based on study design was low.

With respect to the quality of the studies, methods were not, in general, described well; issues of bias, validity, and reliability were not always addressed; and methods of data collection and analysis were often not explained well. As many as 25 out of the 33 primary studies did not have a recruitment strategy that seemed appropriate for the aims stated for the research and 23 of the studies did not use other groups or baselines with which to compare their findings. Furthermore, in only one study was the possibility of researcher bias mentioned. Using these findings as a basis, we concluded that there were serious limitations to the quality of the studies that inevitably increases the risk of bias or confounding and that we must be circumspect about the studies' reliability.

With respect to consistency, i.e., the similarity of estimates of effect across studies, we found differences in both the direction of effects and the size of the differences in effects, i.e., we found no consistent evidence of association from two or more studies with no plausible confounders nor did we find direct evidence from studies with no major threats to validity. These inconsistencies might have been due to imprecise or sparse data, and reporting bias.

With respect to directness, i.e., the extent to which the people, interventions, and outcome measures are similar to those of interest, we found that most studies were concerned with XP. This left us with an uncertainty about the directness of evidence for other agile methods. However, since most of the studies regarding XP were performed with student subjects or professionals who had little or no experience in agile development, this also raised an issue regarding the directness of evidence for XP. In addition, very few studies provided direct comparisons of interventions; hence, we had to make comparisons across studies. However, such indirect comparisons leave greater uncertainty than direct comparisons because of all the other differences between studies that can affect the results. Our judgment was thus that there were major uncertainties about the directness of the included studies.

Combining the four components of study design, study quality, consistency, and directness, we found that the strength of the evidence in the review regarding the benefits and limitations of agile methods, and for decisions related to their adoption, was very low. This means that any estimate of effect that is based on evidence of agile software development from current research is very uncertain. This is also consistent with criticisms that have been raised regarding the sparse scientific support for many of the claims made by the agile community [39].

5. ASSESSING THE QUALITY OF SRs

As with any research, the quality of systematic reviews is likely to be variable depending on how rigorously the authors have conducted the review. Several publications have described the science of reviewing research, differences among narrative reviews, systematic reviews, and meta-analyses, and how to carry out, critically appraise, and apply such secondary research in practice (see [32] and [47] for an overview).

In addition to CASP, two tools are particularly relevant for assessing the quality of SRs: the QUOROM statement [42] and the MOOSE statement [50]. The QUOROM (Quality Of

Reporting Of Meta-analyses) statement sets out to achieve the same improvement in the quality of reporting of meta-analyses as the CONSORT statement is attempting to do for clinical trials. Many health-related journals have adopted the QUOROM guidelines and require reporting of SRs and meta-analyses according to the QUOROM checklist and also that a detailed flowchart is included, which illustrates the inclusion and exclusion of studies from the review.

The checklist describes a way to present the abstract, introduction, methods, results, and discussion sections of a report of a meta-analysis. It is organized into 21 headings and subheadings regarding searches, selection, validity assessment, data abstraction, study characteristics, and quantitative data synthesis, and in the results with "trial flow", study characteristics, and quantitative data synthesis. The flow-diagram provides information about both the numbers of trials identified, included, and excluded and the reasons for exclusion of trials.

QUOROM is primarily aimed at systematic reviews and meta-analyses of randomized controlled trials. However, in many situations randomized controlled designs are not feasible and only data from observational studies are available; in other situations, observational studies may provide better evidence. Systematic reviews of observational studies and mixed methods studies present particular challenges because of inherent biases and differences in study designs; yet, they may provide a tool for helping to understand and quantify sources of variability in results across studies.

Methodological and interpretational concerns make it clear that thorough reporting of systematic reviews of observational studies is absolutely essential. This is exactly the aim of the MOOSE statement, which is a checklist of items for reporting that builds on similar activities for randomized controlled trials, but is intended for use by authors, reviewers, editors, and readers of systematic reviews of observational studies (Table 6).

The MOOSE guidelines do not require the use of a flow-diagram like the QUOROM statement does, but we would nevertheless recommend the use of a flow-diagram because studies will be excluded at different stages in the review for different reasons, and because we believe that visualizing the inclusion and exclusion criteria at each stage will help readers better understand the process (see Figure 1 for en example).

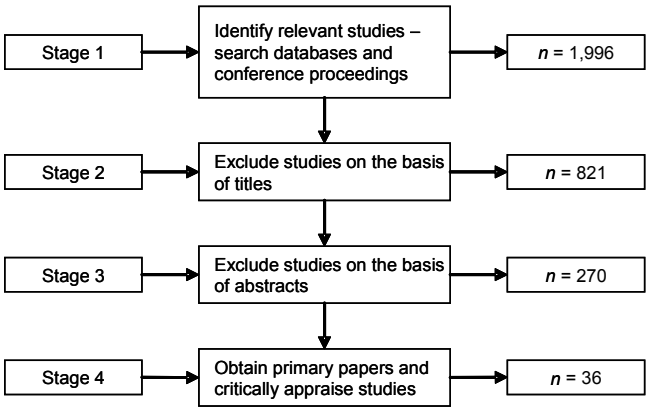

Figure 1. Flow-diagram of study selection process [15]

Table 6. MOOSE checklist [50]

Reporting of background should include:
- Problem definition
- Hypothesis statement
- Description of study outcome(s)
- Type of exposure or intervention used
- Type of study designs used
- Study population

Reporting of search strategy should include:
- Qualifications of searchers (e.g., librarians and investigators)
- Search strategy, including time period included in the synthesis and keywords
- Effort to include all available studies, including contact with authors
- Databases and registries searched
- Search software used, name and version, including special features used (e.g., explosion)
- Use of hand searching (e.g., reference lists of obtained articles)
- List of citations located and those excluded, including justification
- Method of addressing articles published in languages other than English
- Method of handling abstracts and unpublished studies
- Description of any contact with authors

Reporting of methods should include:
- Description of relevance or appropriateness of studies assembled for assessing the hypothesis to be tested
- Rationale for the selection and coding of data (e.g., sound clinical principles or convenience)
- Documentation of how data were classified and coded (e.g., multiple raters, blinding, and inter-rater reliability)
- Assessment of confounding (e.g., comparability of cases and controls in studies where appropriate)
- Assessment of study quality, including blinding of quality assessors; stratification or regression on possible predictors of study results
- Assessment of heterogeneity
- Description of statistical methods (e.g., complete description of fixed or random effects models, justification of whether the chosen models account for predictors of study results, dose-response models, or cumulative meta-analysis) in sufficient detail to be replicated
- Provision of appropriate tables and graphics

Reporting of results should include:
- Graphic summarizing individual study estimates and overall estimate
- Table giving descriptive information for each study included
- Results of sensitivity testing (e.g., subgroup analysis)
- Indication of statistical uncertainty of findings

Reporting of discussion should include:
- Quantitative assessment of bias (e.g., publication bias)
- Justification for exclusion (e.g., exclusion of non–English-language citations)
- Assessment of quality of included studies

Reporting of conclusions should include:
- Consideration of alternative explanations for observed results
- Generalization of the conclusions (i.e., appropriate for the data presented and within the domain of the review)
- Guidelines for future research
- Disclosure of funding source

The study selection process of the systematic review, used as an example in this paper, consisted of four distinct stages as depicted in the flow-diagram in Figure 1. The search strategy applied at stage 1 of the review resulted in a total of 2,946 "hits" that included 1,996 unduplicated citations. At stage 2, both authors went through the titles of the studies from stage 1 and, based on these, excluded 1,175 studies, leaving 821 for the next stage. At stage 3, we made independent reviews of all 821 abstracts and as a result of this process; we excluded another 551 articles leaving 270 for a detailed quality assessment. At stage 4, we applied the quality assessment described in Section 3.3, and excluded another 234 lessons-learned or single-practice articles, leaving 33 primary and 3 secondary studies for data extraction and synthesis.

Regarding the quality of the SR that we have used as an example in this paper, it is associated with (at least) two limitations, which we believe are common to several other SRs too, i.e., bias in the selection of publications and inaccuracy in data extraction. Even though we identified keywords and search terms that would enable us to identify the relevant literature, it is important to recognize that SE keywords are not standardized and that they can be both discipline- and language-specific. Therefore, due to our choice of keywords and search strings, there is a risk that relevant studies were omitted. To try to avoid selection bias, we piloted every part of the review process, and in particular, the search strategy and citation management procedure, in order to clarify weaknesses and refine the selection process. Furthermore, we tried to ensure an unbiased selection of articles by utilizing a multistage process that involved three researchers who documented the reasons for inclusion and exclusion at every step, as suggested by Kitchenham and Charters [32]. So, even though we tried to limit the amount of selection bias, we cannot eliminate it.

When we piloted the data extraction process, we found that several articles lacked sufficient details about the design and findings of a study and that, due to this, we differed too much in what we actually extracted. As a consequence, all data from all the 33 primary studies were extracted by the two authors in consensus meetings according to a predefined extraction form. However, we often found that the extraction process was hindered by the way some of the primary studies were reported. Many articles lacked sufficient information for us to be able to document them satisfactorily in the extraction form. More specifically, we frequently found that methods were not described adequately, that issues of bias and validity were not always addressed, that methods of data collection and analysis were often not explained well, and that samples and study settings were often not described well. There is therefore a possibility that the extraction process may have resulted in some inaccuracy in the data.

6. CONCLUSION

Both traditional narrative reviews and systematic reviews are retrospective, observational studies and are therefore subject to systematic and random error. The quality and worth of a review, thus, depends on the extent to which scientific review methods have been used to minimize error and bias. It is important, therefore, that users of systematic reviews know how much confidence they can place in the conclusions and recommendations arising from such reviews.

Since a large portion of empirical SE studies is qualitative and observational in nature, we expect a large portion of systematic reviews in SE to be interpretive rather than integrative. As both the GRADE system and MOOSE acknowledge the challenges of grading the overall strength of evidence for and reporting of systematic reviews of observational studies, we find these systems to be of particular relevance for systematic reviews in SE.

Three issues have been identified in this paper, which are essential to help author provide more rigorous systematic reviews as well as to help users of systematic reviews to make judgments regarding the utility of a specific review: (1) explicit assessment of the quality of the primary studies, (2) grading the strength of the total body of evidence, and (3) explicit discussion of the limitations of the systematic review itself. None of these issues are trivial. And further research is needed to identify alternative ways of dealing with these issues in systematic reviews in software engineering. As we see it, there are specific challenges related to interpretive reviews that should be given attention. We hope that this paper can serve as a starting point for discussing these issues, as we believe that such discussions are crucial for the development of a more evidence-based approach to software engineering.

7. ACKNOWLEDGMENTS

The work in this paper was supported by the Research Council of Norway through the project Evidence-Based Software Engineering (181685/I30).

8. REFERENCES

[1] Arisholm, E., Gallis, H., Dybå, T., and Sjøberg, D. (2007) Evaluating Pair Programming with Respect to System Complexity and Programmer Expertise, *IEEE Transactions on Software Engineering*, 33(2): 65–86

[2] Bailey, J., Zhang, C., Budget, D., and Turner, M., "Search Engine Overlaps: Do They Agree or Disagree?," *Second International Workshop on Realising Evidence-Based Software Engineering (REBSE'07)*, 2007.

[3] Bhandari, M., Busse, J.W., Jackowski, D., Montori, V.M., Schünemann, H., Sprague, S., Mears, D., Schemitsch, E.H., Heels-Ansdell, D., and Devereaux, P.J. (2004) Association between Industry Funding and Statistically Significant Pro-Industry Findings in Medical and Surgical Randomized Trials, *CMAJ*, 170: 477–480

[4] Biolchini, J., Mian, P.G., Natali, A.C.C., and Travassos, G.H. (2005) *Systematic Review in Software Engineering*, Univ. Rio de Janeiro, TR, ES 679/05.

[5] Brehmer, B. (1989) In One Word: Not from Experience, *Acta Psychologica*, 45(1-3): 223–241.

[6] Brereton, P., Kitchenham, B. A., Budgen, D., Turner, M., and Khalil, M., "Lessons from Applying the Systematic Literature Review Process within the Software Engineering Domain," *Journal of Systems and Software*, no. 4, vol. 80, pp. 571–583, 2007.

[7] Campbell, D.T. and Stanley, J.C. (1963) *Experimental and Quasi-Experimental Designs for Research*, Boston: Houghton Mifflin Company

[8] Cook, T.D. and Campbell, D.T. (1979) *Quasi-Experimentation: Design & Analysis Issues for Field Settings*, Boston: Houghton Mifflin Company

[9] Cooper, H. (1998) *Synthesizing Research* (3rd Ed.), Thousand Oaks, CA: Sage.

[10] Cooper, H. and Hedges, L.V. (Eds.) (1994) *Handbook of Research Synthesis*, New York: Russell Sage Foundation.

[11] Davies, A., Dieste, O., Hickey, A., Juristo, N., and Moreno, A.M. (2006) Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review, *Proceedings 14th IEEE International Requirements Engineering Conference (RE'06)*, IEEE Computer Society, pp. 179–188.

[12] Dieste, O. and Padua, A.G. (2007) Developing Search Strategies for Detecting Relevant Experiments for Systematic Reviews, *Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, 20-21 Sept., IEEE Computer Society, pp. 215–224.

[13] Dittrich, Y., John, M., Singer, J., and Tessem, B. (2007) For the Special issue on Qualitative Software Engineering Research, *Information and Software Technology*, 6(49): 531–539.

[14] Dixon-Woods, M., Agarwal, S., Jones, D., Young, B., and Sutton, A. (2005) Synthesising Qualitative and Quantitative Evidence: A Review of Possible Methods, *J. of Health Services Research & Policy*, 10(1): 45–53.

[15] Dybå, T. and Dingsøyr, T. (2008) Empirical Studies of Agile Software Development: A Systematic Review, *Information and Software Technology*, 50(9-10): 833–859.

[16] Dybå, T., Dingsøyr, T., and Hanssen, G.K. (2007) Applying Systematic Reviews to Diverse Study Types: An Experience Report, *Proceedings of the 1st International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, 20-21 Sept., IEEE Computer Society, pp. 225–234.

[17] Dybå, T., Kampenes, V.B. and Sjøberg, D.I.K. (2006) A Systematic Review of Statistical Power in Software Engineering Experiments, *Information and Software Technology*, 48(8):745–755.

[18] Dybå, T., Kitchenham, B.A., and Jørgensen, M. (2005) Evidence-Based Software Engineering for Practitioners, *IEEE Software*, 22(1): 58–65.

[19] Egger, M., Smith, G.D., and Altman, D.G. (2001) *Systematic Reviews in Health Care: Meta-analysis in Context* (2nd Ed.), London: BMJ Publishing Group.

[20] GRADE Working Group (2004) Grading Quality of Evidence and Strength of Recommendations," *BMJ*, 328: 1490.

[21] Greenhalgh, T. (2006) *How to Read a Paper: The Basics of Evidence-Based Medicine* (3rd Ed.), London: BMJ Publishing Group.

[22] Hannay, J., Sjøberg, D.I.K., and Dybå, T. (2007) A Systematic Review of Theory Use in SE Experiments, *IEEE Transactions on Software Engineering*, 33(2): 87–107.

[23] Higgins J.P.T. and Green, S. (Eds.) (2008), *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.0.0* (updated February 2008), The Cochrane Collaboration, available from www.cochrane-handbook.org.

[24] Höst, M. and Runeson, P. (2007) Checklists for Software Engineering Case Study Research, *Proceedings of the First International Symposium on Empirical Software Engineering and Measurement (ESEM'07)*, Madrid, Spain, 20-21 Sept., IEEE Computer Society, pp. 479–481

[25] Jadad, A.R., Moore, R.A., Carroll, D., Jenkinson, C., Reynolds, D.J., Gavaghan, D.J., and McQuay, H.J. (1996) Assessing the Quality of Reports of Randomized Clinical Trials: Is Blinding Necessary?, *Controlled Clinical Trials*, 17(1): 1–12.

[26] Jedlitschka, A. and Pfahl, D. (2005) Reporting Guidelines for Controlled Experiments in Software Engineering, *Proceedings of the 4th International Symposium on Empirical Software Engineering (ISESE'05)*, Noosa Heads, Australia, 17-18 Nov, IEEE Computer Society, pp. 95–104.

[27] Jørgensen, M. and Shepperd, M. (2007) A Systematic Review of Software Development Cost Estimation Studies, *IEEE Transactions on Software Engineering*, 33(1): 33–53.

[28] Kampenes, V.B., Dybå, T., Hannay, J.E., and Sjøberg, D.I.K. (2007) A Systematic Review of Effect Size in Software Engineering Experiments, *Information and Software Technology*, 49(11-12): 1073–1086.

[29] Kampenes, V.B., Dybå, T., Hannay, J.E., and Sjøberg, D.I.K. (In Press) A Systematic Review of Quasi-Experiments in Software Engineering, Accepted to *Information and Software Technology*.

[30] Khan, K.S. ter Riet, G., Glanville, J., Sowden, A.J., and Kleijnen, J. (Eds.) (2001) *Undertaking Systematic Review of Research on Effectiveness, CRD's Guidance for those Carrying Out or Commissioning Reviews*, CRD Report Number 4 (2nd Ed.), NHS Centre for Reviews and Dissemination, University of York.

[31] Kitchenham, B.A. (2007) Empirical Paradigm – The Role of Experiments, in V.R. Basili *et al*. (Eds.), *Empirical Software Engineering Issues: Critical Assessment and Future Directions*, Proceedings from Int. Workshop, Dagstuhl Castle, June 26-30, 2006, Lecture Notes in Compute Science 4336, Springer, pp. 25–32.

[32] Kitchenham, B.A. and Charters, S. (2007) *Guidelines for performing Systematic Literature Reviews in Software Engineering*, Version 2.3, Keele University, EBSE Technical Report, EBSE-2007-01.

[33] Kitchenham, B.A., Al-Khilidar, H., Babar, M.A., Berry, M., Cox, K., Keung, J., Kurniawati, F., Staples, M., Zhang, H., and Zhu, L. (2008) Evaluating Guidelines for Reporting Empirical Software Engineering Studies, *Empirical Software Engineering*, 13(1): 97–121.

[34] Kitchenham, B.A., Dybå, T., and Jørgensen, M. (2004) Evidence-Based Software Engineering, *Proceedings of the 26th International Conference on Software Engineering (ICSE 2004)*, IEEE CS Press, pp. 273–281.

[35] Kitchenham, B.A., Pfleeger, S.L., Pickard, L.M., Jones, P.W., Hoaglin, D.C., El Emam, K., and Rosenberg, J. (2002) Preliminary Guidelines for Empirical Research in Software

Engineering, *IEEE Transactions on Software Engineering*, 28(8): 721–734.

[36] Klein, H.K. and Myers, M.D. (1999) A Set of Principles for Conducting and Evaluating Interpretive Field Studies in Information Systems, *MIS Quarterly*, 23(1): 67–93.

[37] Laitenberger, O. and Rombach, D. (2003) (Quasi-) Experimental Studies in Industrial Settings, in N. Juristo and A.M. Moreno (Eds.), *Lecture Notes on Empirical Software Engineering*, Singapore: World Scientific (Series on Software Engineering and Knowledge Engineering 12), pp. 167–227.

[38] Lee, A.S. (1989) A Scientific Methodology for MIS Case Studies, *MIS Quarterly*, 13(1): 33–50.

[39] Mcbreen, P. (2003) *Questioning Extreme Programming*, Boston, MA, USA: Pearson Education.

[40] Mendes, E. (2005) A Systematic Review of Web Engineering Research, *Proceedings of the 4th International Symposium on Empirical Software Engineering (ISESE'05)*, Noosa Heads, Australia, 17-18 Nov, IEEE Computer Society, pp. 481–490.

[41] Moher D, Schultz KF, Altman D (2001) The CONSORT Statement: Revised Recommendations for Improving the Quality of Reports of Parallel-Group Randomized Trials, *Lancet*, 357:1191–1194, April 14.

[42] Moher D., Cook, D.J., Eastwood, S., Olkin, I., Rennie, D., and Stroup, D.F. (1999) Improving the Quality of Reports of Meta-Analyses of Randomised Controlled Trials: The QUOROM Statement, *Lancet*, 354: 1896–1900.

[43] Mulrow, C. and Cook, D. (Eds.) (1998) *Systematic Reviews: Synthesis of Best Evidence for Health Care Decisions*, Philadelphia: Am. College of Physicians.

[44] Noblit, G.W. and Hare, R.D. (1988) *Meta-Ethnography: Synthesizing Qualitative Studies*, Thousand Oaks: Sage.

[45] Petticrew, M. and Roberts, H. (2006) *Systematic Reviews in the Social Sciences: A Practical Guide*, Oxford, UK: Blackwell.

[46] Shadish, W.R., Cook, T.D. and Campbell, D.T. (2002), *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*, Boston: Houghton Mifflin.

[47] Sjøberg, D.I.K., Dybå, T., and Jørgensen, M. (2007) The Future of Empirical Methods in Software Engineering Research, *29th International Conference on Software Engineering (ICSE'07)*, *Future of Software Engineering (FOSE'07)*, Minneapolis, Minnesota, USA, 20-26 May, IEEE Computer Society Press IEEE Computer Society, pp. 358–378.

[48] Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanović, A., Liborg, N.-K., and Rekdal, A.C. (2005) A Survey of Controlled Experiments in Software Engineering, *IEEE Transactions on Software Engineering*, 31(9):733–753,

[49] Staples, M. and Niazi, M. (2007) Experiences Using Systematic Review Guidelines, *Journal of Systems and Software*, 80(9): 1425–1437.

[50] Stroup, D.F., Berlin, J.A., Morton, S.C., Olkin, I., Williamson, G.D., Rennie, D., Moher, D., Becker, B.J., Sipe, T.A., and Thacker, S.B. (2000) Meta-analysis of Observational Studies in Epidemiology: A Proposal for Reporting, *JAMA*, 283(15): 2008–2012.