A checklist for integrating student empirical studies with research and teaching goals

Jeffrey C. Carver · Letizia Jaccheri · Sandro Morasca · Forrest Shull

Abstract The use of empirical studies with students in software engineering helps researchers gain insight into new or existing techniques and methods. However, due mainly to concerns of external validity, questions have been raised about the value of these types of studies. The authors of this paper draw on their experiences of conducting a large number of empirical studies in university courses in three countries (Italy, Norway, and the United States) to address this important issue. This paper first identifies the requirements that research and pedagogy place on a valid empirical study with students. This information is then used as the basis for a checklist that provides guidance for researchers and educators when planning and conducting studies in university courses. The goal of this checklist is to help ensure that these studies have as much research and pedagogical value as possible. Finally, an example application of the checklist is provided to illustrate its use.

J. C. Carver (✉)
Department of Computer Science, University of Alabama, Tuscaloosa, AL, USA
e-mail: carver@cs.ua.edu

L. Jaccheri
Department of Computer and Information Science, Norwegian
University of Science and Technology, Trondheim, Norway
e-mail: letizia@idi.ntnu.no

S. Morasca
Dipartimento di Scienze della Cultura, Politiche e dell'Informazione,
Università degli Studi dell'Insubria, Varese, Italy
e-mail: sandro.morasca@uninsubria.it

F. Shull
Fraunhofer Center for Experimental Software Engineering, College Park, MD, USA
e-mail: fshull@fc-md.umd.edu

1 Introduction

There are many reasons why researchers conduct empirical studies, including: investigating the feasibility of an approach, characterizing the strengths or weaknesses of products and processes, highlighting areas for improvements, or evaluating software engineering techniques. Different types of empirical studies may be conducted to accomplish these goals, e.g., controlled experiments, quasi-experiments, correlation studies, case studies, or surveys (Zelkowitz and Wallace 1998; Wohlin et al. 2000; Sjoeberg et al. 2005). The choice of empirical study depends on the study goals, the resources available and the constraints of the experimental environment.

A large number of the studies appearing in the empirical software engineering literature make use of student subjects (Singer and Vinson 2002). Such a study, which is an important type of *in vitro* study (i.e. one taking place in the laboratory, as it were), can be referred to as an Empirical Study with Students (ESWS). ESWSs are often the subject of much debate because, for example, they:

- use students whose level of experience is not fully representative of professional software developers; and
- apply practices to "toy" projects, rather than to full-strength industrial applications.

For these reasons, ESWSs are often viewed skeptically by researchers and practitioners. Furthermore, reviewers of scientific journals and conferences sometimes question the value of such studies, to the extent that ESWSs are sometimes rejected out of hand by reviewers and other researchers. It is true that these studies are often not of immediate interest to the industrial and research communities. However, a study's contribution has to be assessed relative to its goals and to the importance of those goals for practitioners and researchers. As Tichy points out in his guidelines for reviewing empirical papers, studies should not be dismissed simply for using students as subjects, but rather judged on whether the study goal justifies the use of student subjects (Tichy 2000).

Empirical studies with professionals, which are generally accepted by researchers and practitioners without much dispute, also suffer from similar generalizability problems. While students are not fully representative of software professionals, professionals in one environment also may not be representative of professionals in other environments, even within the same application domain. And, while the artifacts and domains used in an ESWS may not be fully representative of industrial-strength artifacts and domains, it is not clear whether the results from one industrial setting are really generalizable to other industrial settings. Often, models from a specific environment in one study cannot be reused "as-is" even in the same environment.

Therefore, just like any other empirical studies, ESWSs can be valuable to the industrial and research communities if they are conducted in an adequate way, address appropriate goals, do not overstate the generalizability of the results, and take into account threats to internal and external validity. For instance, an ESWS can be used to obtain preliminary evidence in support of or against some research hypothesis. As Tichy indicates (Tichy 2000), this evidence is useful both from a scientific point of view, i.e. establishing trends or eliminating alternative hypotheses, and from an industrial point of view, i.e. convincing professionals to participate in future studies or to use new techniques. The results of *in vitro* empirical studies in other disciplines are routinely published in conference proceedings and journals. In fact, according to McBurney: "Most

research in psychology is done using convenience samples: students enrolled in introductory psychology courses" (McBurney 2001). Even though there is no guarantee that *in vitro* results can be replicated *in vivo* (i.e. in real life), these *in vitro* studies still make a valuable contribution to the field. Note also that in other disciplines like pharmaceutical research, *in vitro* studies of new medicines are routinely conducted, even though there is no guarantee that a medication that is effective *in vitro* will also be effective *in vivo*.

In addition to their **research** value, ESWSs should also have **pedagogical** value. Whether conducted during classroom hours or as homework, an ESWS has to compete for scarce time, effort, and resources in the course. Therefore, in addition to the traditional *professional* and *research* stakeholders, the interests and viewpoints of two additional stakeholders, *students* and *instructors*, must be identified and considered. While some research and pedagogical interests may be complementary, others may be in conflict. Students are interested in what they can learn from participating in an ESWS, while the researcher is interested in the quality of the data and results. Unless the study is carefully planned, executed, and integrated into the course, it is likely that the study will neither provide the students with much educational value nor the researcher with much scientific value.

In an ESWS, the instructor plays the central role in identifying and balancing the stakeholders' interests. In many cases, the same person acts as both the researcher and the instructor—with potentially conflicting interests and goals. The researcher is usually interested in gathering data about a specific hypothesis, while the instructor is interested in ensuring the educational value of the study. In such a case, a feasible trade-off between these conflicting goals must be found so the study can provide both adequate data for the researcher and educational value for the students. While many ESWSs have been reported in the software engineering scientific literature (Sjoeberg et al. 2005), we find that many instructors in other sub-disciplines of computer science (e.g. high-performance computing or software reliability engineering) are often unsure of how they can run a useful ESWS in their class.

We, the authors, have been involved as researchers and/or instructors in several ESWSs, e.g. (Shull et al. 2000; Jaccheri 2001; Shull et al. 2001; Baresi and Morasca 2002; Baresi et al. 2003; Morasca 2003; Shull et al. 2005; Walia and Carver 2006). In an earlier paper, we addressed some issues in ESWSs by presenting a framework for assessing student experiments from four points of view: researcher, student, instructor, and professional. We defined costs and benefits for these stakeholders and noted that designing a valid and appropriate ESWS requires balancing the costs and benefits for all stakeholders (Carver et al. 2003). In this paper we extend that work by putting those issues in context through a literature review of research and pedagogy to identify requirements for successful ESWSs. Each requirement is labeled *Rx* where *x* is the requirement number. These requirements, along with the authors' experiences, are then used as the basis of a checklist for planning and conducting an ESWS. This checklist is designed to help both novice and experienced researchers keep the preparation tasks organized.

The paper is organized as follows. Section 2 highlights the research value of ESWSs as found in the literature and provides requirements for the checklist. Section 3 highlights the pedagogical value of ESWSs in the literature and provides additional requirements for the checklist. Section 4 then describes a checklist developed to meet these requirements. Section 5 provides examples of the use of the checklist in practice. Finally, the conclusions are given in Section 6.

2 The Research Value of ESWSs

Section 2.1 provides an overview of previous work to highlight the research value of ESWSs. Then, Section 2.2 provides research goals for which ESWSs can be used effectively. Finally, Section 2.3 covers the dimensions on which the research value of an ESWS can be judged.

2.1 Related Work

As an indication of the increase in ESWSs, one survey documents that of 113 controlled experiments published between 1993 and 2002, 82 had student subjects, 21 had professional subjects, nine had both professionals and students, and for 1 the makeup of the sample was unknown (Sjoeberg et al. 2005). It is important to consider whether the results from these studies are valid contributions to the software engineering body of knowledge. To address this consideration, an analysis of the context dimensions which affect the value of a study's results must be conducted.

Höst, et al., write that "A key problem is the external validity of controlled experiments performed in a laboratory setting. It is often materialized in the form of comments regarding the use of students as subjects". They criticize the simplistic view that divides experiments into student-based and not student-based. Rather, they identify two main factors that affect the validity of the results: the incentives provided to the subjects, and the experience of the subjects. These two factors were used to classify experiments reported in the literature. The results showed that experiments with similar classifications produced similar results. Therefore, these factors were judged to be useful (Höst et al. 2005).

Another examination of the differences between student and professional subjects from a research point of view found significant differences between undergraduate and graduate students but only small differences between graduate students and professionals. As a result, Höst, et al., described conditions under which student experiments should be conducted and indicated that the educational goals of the course should be harmonized with the research goals (Höst et al. 2000). This conclusion highlights the importance of the relationship between the pedagogical and research goals. Based on these two papers, the first two requirements for a successful ESWS are:

R 1.   *External validity issues must be consciously considered.*
R 2.   *The ESWS must be properly integrated with the course.*

Singer and Vinson seem to be unique in addressing ethical issues in empirical studies. They discuss the peculiarities of student subjects and provide references to standards researchers should consider to design a valid ESWS whose results can contribute to the field. The main ethical issues that must be addressed are: full informed consent, the power relationship between instructor and student during subject recruitment, remuneration, and use of experimental data (Singer and Vinson 2002). Because researchers must consider each of these issues carefully, another requirement is:

R 3.   *Ethical issues must be adequately addressed by the study design.*

2.2 Reasons to Use ESWSs

ESWSs have often suffered from a prejudice concerning the utility of their results, i.e., their external validity. As discussed in Section 1, these validity threats are not unique to ESWSs. To accurately judge the usefulness of a particular ESWS, it should be evaluated based on

some additional, in-depth criteria. For example: For the measures of interest, can the researcher make a case that students are a good proxy for professionals? Can the researcher make a convincing argument that an ESWS is an effective way of addressing the study goals? Some goals that are well-suited for ESWSs and deserving of publication include:

- *Piloting experimental methodologies.* Because *in vivo* empirical studies in professional settings often require a large amount of time, effort, and resources, they need to be planned and executed carefully. To aid in this planning, it is useful to conduct an *in vitro* pilot study prior to the *in vivo* study. In this type of ESWS, the evaluation of whether the experimental methodology can address the problem of interest may be of equal or greater value than the results related to specific hypotheses. A desired outcome of this type of study is an increase in general knowledge about good experimental design for specific types of problems.

- *Studying issues related to a technology's learning curve or the behavior of novices.* This was exactly the goal of several ESWSs conducted as part of the DARPA High Productivity Computing Systems (HPCS) project[1] in which two of the authors were involved. One of the important goals of the HPCS project is to increase the size of the workforce capable of efficiently programming massively parallel supercomputers. In this case, the goal of the ESWSs was to understand whether various development approaches could quickly increase the effectiveness of novice programmers. Therefore, because the desired outcome of the ESWSs was to learn about the activities of novice developers, students were exactly the right test population (Shull et al. 2005; Hochstein et al. 2006; Basili et al. 2008).

- *Testing the feasibility of technologies.* Many critiques of individual ESWSs point out that evaluating a new development practice with unrepresentative subjects and/or toy problems does not help in understanding whether that practice will be beneficial for use on real projects. While this conclusion is often true, to overlook ESWSs for only this reason intrudes on the first step of the research and technology transfer process. Granted that to drive organizational change, it is necessary to evaluate the practice under realistic conditions. However, before the management of an organization is willing to invest in running such a relatively expensive study, they will want to see evidence of at least some level of feasibility, as noted by Tichy (Tichy 2000). An ESWS can be thought of as a filtering step that sacrifices some external validity for the benefit of inexpensively understanding the feasibility of new ideas (Shull et al. 2001). A desired outcome is that the resulting relatively weak evidence from the ESWS convinces someone else to invest in a replication in a more realistic setting.

- *Obtaining preliminary evidence in favor of or against a research hypothesis.* Suppose that the results obtained in an ESWS are opposite of what the researcher expects. For instance, suppose that a researcher expects a new technology under evaluation to positively impact a variable of interest, i.e. decreasing effort, but the result of the ESWS shows a strong negative effect, i.e. significantly increasing effort. Then, it would be sensible for the researcher to at least examine the causes of this unexpected result before conducting the same study, let alone introducing the technology, in an industrial environment. Tichy agrees that eliminating alternative hypotheses is a valid use for an ESWS (Tichy 2000). This approach is commonly used in other disciplines. For example, in pharmaceutical research, medications that show negative effects *in vitro* are hardly ever sent to the *in vivo* phase. There is no reason why software engineering

---

[1] http://www.highproductivity.org

should not adopt a similar practice. If the ESWS results provide support for the hypothesis, then it is sensible to carry out an in-depth analysis of the external validity of the ESWS. But the ESWS gives the researcher more confidence in the expected results from professionals.

This discussion provides another requirement for successful ESWSs:

R 4.   *The correct goal must be chosen for the study based on its environment.*

2.3 Dimensions for Evaluating an ESWS

In the previous discussion we introduced a number of issues that are important to the research value of an ESWS. These issues can be organized into three important dimensions to use for judging the value of an ESWS.

Realism of the environment: time vs. experience Researchers must balance the time required for the study with the experience of the subjects. One way to increase realism in experiments is by paying professionals to do certain tasks. In this case, the experience dimension is more realistic, but the time dimension is less realistic because professionals can only be paid for short tasks. Conversely, although students will often not have the same experience level as professional developers, an ESWS can be more realistic in the time dimension because it can occur in the context of projects that last weeks (Sjoeberg et al. 2002).

R 5.   *The study setting must be appropriate relative to its goals, the skills required and the activities under study.*

How well student behavior approximates professional behavior In the context of software estimation, Jorgensen, et al., discuss similarities and differences between students and professionals. While the students are less experienced than professionals, the study results show that on some tasks students performed better than professionals (Jorgensen et al. 2004). Although a common assumption is that students and professionals come from radically different populations, some researchers have investigated the extent to which these differences are important. In cases where the variables that differentiate students and professionals can be adequately addressed, such as the examples by Höst, et al. (Höst et al. 2000) mentioned earlier, then ESWSs can be conducted without introducing critical threats to validity.

R 6.   *The effect of differences between the subject population and the target population must be discussed.*

The appropriateness of the study goals Experiments do not always have to be conducted in a completely realistic setting. In some cases, giving up increased realism to use "cheaper" subjects makes sense. Whether that tradeoff is rational depends on the overall goals (e.g. trying to obtain initial insight into a new idea or testing feasibility). There are many examples in the literature of studies that use student subjects to match a specific research goal. For example, Basili, et al., validate object oriented metrics in the context of an undergraduate/graduate level course on object oriented analysis and design. In this example, the OO design properties under investigation should be constant regardless of the experience of the person who designed the system (Basili et al. 1996). Also, the guidelines for performing empirical investigations suggested by Kitchenham, et

al., mention that ESWSs can evaluate the use of a technique by novices or non-experts (Kitchenham et al. 2002). This discussion supports *R4* defined at the end of Section 2.2

3 The Pedagogical Value of ESWSs

The pedagogical value of an ESWS is judged by how well it supports the educational goals of the course. By incorporating the pedagogical perspective with the research perspective, researchers can design ESWSs that better meet the needs of more stakeholders. Of course, these two perspectives are not completely in opposition: An experiment that is pedagogically valid should enhance the students' motivation to participate and hence improve the research results.

Section 3.1 discusses basic pedagogical theory and identifies requirements for conducting a pedagogically valid ESWS. Section 3.2 places ESWSs in the context of the software engineering education literature and identifies an additional requirement to ensure the ESWS is in line with that research.

3.1 Pedagogical Theory

Harmonizing a discussion of ESWSs with the prevailing perspectives and theories of education is challenging. Over 2,000 years, a number of educators, from Plato to our contemporaries, have debated the trade-offs between theory and practice and the importance of delegating responsibilities to the learner. Even educational researchers do not claim that any single theory accounts for all the ways individuals learn. In fact, it is likely that different individuals learn most effectively in different ways. These theories can provide useful insight into software engineering curricula and provide important requirements for successful ESWSs.

First, we point to the **social education theory** that says interaction among human beings is a source of learning and creativity. According to Vygotsky, in order to teach well, an instructor must understand the mental models that students use to perceive the world and the assumptions they make to support those models. The purpose of learning is for an individual to construct his or her own meaning, not just memorize the "right" answers and regurgitate someone else's meaning. Instructors also should rely heavily on open-ended questions and promote extensive dialogue among students (Vygotsky 1978).

Second, Bloom's taxonomy of six progressively complex levels of cognition (Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation) (Bloom 1956) provides useful guidance for writing lesson objectives in the software engineering education community. For example, an implication of the taxonomy is that education should move the students from simply comprehending a software engineering practice to being able to make their own evaluation of the value of that practice in their context. This comprehension also supports the social education theory by helping the students create their own understanding of the value of the practice. Participating in an ESWS and then understanding how it was conducted can help a student gain these important and useful skills.

R 7. *Students should learn the value of using empirical studies to evaluate products and processes and how to conduct them so they can later perform their own assessments.*

Social education theory can be extended to include an emotional perspective, e.g. the ability to deal with discord, group psychology, and the relationship between motivation and learning. Social education theory stresses that to learn and live, a person must have emotional strength and the ability to deal with discord (Jay 2002). A clear implication of this perspective is the necessity for students to learn how to work in team or group settings (where discord is likely to occur). Because the activities that occur during ESWSs are often collaborative, they provide students with a prime opportunity to learn and practice these skills. These two findings produce the following requirement for a successful ESWS:

R 8.   *Group work or collaborative work should be included in an ESWS.*

The more traditional mental approach to education, also defined as instructivism, emphasizes class lecture as the method of learning. We argue that because they provide a different approach to engaging groups of students, ESWSs can augment mental learning by supporting social and emotional learning.

3.2 The Software Engineering Education Literature

The relationship between software engineering education and empirical software engineering can be seen by the overlap in the conferences and journals from the two fields, e.g. software engineering experiments with student subjects, software engineering courses in which experimentation issues are investigated, software engineering student projects (often the subject of an ESWS) in which professionals play a significant role, and theoretical frameworks about software engineering student projects. Some of these frameworks take the education perspective (e.g., (Umphress et al. 2002)), while others the empirical software engineering research perspective (e.g., (Sjoeberg et al. 2005)).

Two aspects of the software engineering education literature are particularly relevant to ESWSs. First, literature that discusses the use of empirical studies in the context of software engineering education is relevant. Second, because many ESWSs are conducted in the context of a project (or introduce a project into a course that otherwise might not have one), literature about project-based software engineering education is relevant. The following discussion indicates the potential that ESWSs have to produce a significant pedagogical benefit.

3.2.1 Empiricism in Software Engineering Education

A theme that is becoming more common in the literature is the usefulness of including an empirical study in the curriculum. For example, the 2005 ACM Special Interest Group on Computer Science Education (SIGCSE) Technical Symposium on Education in Computer Science had a session specifically devoted to experimentation, experimentation methods, and the benefits of teaching these methods. That session produced two papers relevant to our work. First, Braught discussed the importance of empirical methods for computer science students in general and described a framework to introduce empiricism into a first year undergraduate course (Braught 2005). Second, Pastel described a human computer interaction course that integrates the student project with a corresponding research project. While undergraduate students work on their project, they act as the subjects for studies run by research students. This arrangement is a valuable scheme for symbiosis between design and research course assignments (Pastel 2005).

The International Conference on Software Engineering Education and Training (CSEE&T) has also devoted attention to the importance of empirical software engineering. Two examples are provided here for illustration. First, Port and Klappholz discussed the

benefits of ESWSs not only from a research point of view, but also as a tool to improve education by exposing students to contexts with a significant professional influence (Port and Klappholz 2004). Second, Höst highlighted the importance of using empirical methods in software engineering education as a tool to evaluate software product quality. If students are familiar with empirical methods, then they can better understand and use quality models to evaluate product and process quality (Höst 2002). This discussion supports R7, which was identified in Section 3.1.

3.2.2 Project-based Software Engineering Education

The literature search also identified a large body of work that discussed the use of projects in software engineering education. The Computing Curriculum 2001 Task Force stressed the importance of including a significant team project encompassing both design and implementation in the curriculum. They also highlight the possibility of working with local companies to allow students to engage in projects in a professional setting (CORPORATE 2001). In another paper, Way discussed a course that was based on interaction with software companies and provided a good list of references related to the use of software engineering projects to support educational goals (Way 2005).

In an *IEEE Software* special issue devoted to education, there is an article about a framework for teaching software project courses (Umphress et al. 2002). In addition, the editors include a list of fundamental software engineering education publications (Hilburn and Humphrey 2002). Two important references stand out. First, Denning discusses the roles of researchers, instructors, students, professionals, and innovators in software engineering education (Denning 1992) highlighting the need to account for multiple perspectives. Second, Bagert, et al., describe a software engineering body of knowledge and curriculum model. They also advocate for the inclusion of software projects in core software engineering courses (Bagert et al. 1999). These papers motivate the following requirement:

R 9.   *ESWSs should include development projects where possible.*

We would like to point out that in some cases ESWSs may address individual steps within the development process, i.e. an inspection, and therefore not be good candidates to use a full development project.

3.3 Summary

Cultural issues also play a significant role when discussing pedagogical value. There are some major differences in the educational approaches used by the three countries represented by the authors of this paper. For example, in Norway less theory is taught in all levels of school compared with Italy. In Norway, students do not receive grades until the eighth grade, while in Italy and in the USA they receive grades starting in the first grade. North American and Norwegian university students are more used to performing homework assignments compared with Italian students.

Because of these issues, discussing pedagogy in an international forum presents challenges. Realizing that the variation in educational contexts makes universal guidelines for ESWSs unlikely, we use our experiences with conducting multiple ESWSs in different contexts to develop a checklist that addresses the nine requirements identified, which are summarized in Table 1. This checklist provide a starting point, but the best way of ensuring pedagogical value is for instructors to engage in a dialogue with education experts at their own institutions and to learn and apply basic educational theories. To be successful in

Table 1 ESWS requirements

Requirements for a successful ESWS
R1 *External validity issues must be consciously considered.*
R2 *The ESWS must be properly integrated with the course.*
R3 *Ethical issues must be adequately addressed by the study design.*
R4 *The correct goal must be chosen for the study based on its environment.*
R5 *The study setting must be appropriate relative to its goals, the skills required and the activities under study.*
R6 *The effect of differences between the subject population and the target population must be discussed.*
R7 *Students should learn the value of using empirical studies to evaluate products and processes and how to conduct them so they can later perform their own assessments.*
R8 *Group work or collaborative work should be included in an ESWS.*
R9 *ESWS should include development projects where possible.*

software engineering education, it is important to learn from the disciplines that are dedicated to the study of education.

4 A Checklist-based Approach for Balancing Pedagogical and Research Issues

Because conducting an ESWS requires a significant amount of effort to be invested both by a researcher (e.g., preparation, coordination with instructor, and analysis of results) and by an instructor (e.g. integrating with the course and coordination with the researcher), the benefits gained from the study need to be maximized. Small discrepancies or mistakes can reduce the ESWS's benefit or even invalidate its results. Furthermore, failure to consider the educational consequences could result in a study that is not beneficial to the students. To help researchers address the requirements in Table 1, we draw on our own positive and negative experiences to develop a checklist of items that will help researchers make ESWSs as effective as possible. The items on this checklist help the study designer elicit the research goals (Section 2) and the pedagogical goals (Section 3) then design a study to address both types of goals.

Table 2 provides a high-level overview of the checklist items grouped by when they should occur. Sections 4.1, 4.2, 4.3, 4.4, 4.5, 4.6, 4.7, 4.8, 4.9, 4.10 then describe each of the items in more detail, with a focus on the researcher's viewpoint as the driving force behind any empirical study. However, these activities also impact the goals and risks of the other three stakeholders (instructor, student, and professional). For each checklist item, we first provide a brief description. Then we discuss the rationale for including it as a necessary step in conducting an ESWS. Finally, we indicate, and justify, which of the requirements in Table 1 it fulfills. We mention the requirements that are more explicitly fulfilled, but other requirements may also be indirectly and more weakly addressed. Section 4.11 then provides a summary table to indicate which checklist items fulfilled which requirements.

4.1 Ensure Adequate Integration of the Study into the Course Topics

As discussed in Section 3.1, special care should be taken to integrate the ESWS with the pedagogical goals and topics of the course. An ESWS that is too focused on the researcher's goals can easily produce invalid results if the students are not well prepared or

Table 2 ESWS checklist

Items to consider when designing and conducting an ESWS

Before the class begins	
1.	Ensure adequate integration of the study into the course topics.
2.	Integrate the study timeline with the course schedule.
3.	Reuse artifacts and tools as appropriate.
4.	Write up a protocol and have it reviewed.
As Soon as the Class Begins	
5.	Obtain subjects' permission for their participation in the study.
6.	Set subject expectations.
When the Study Begins	
7.	Document information about the experimental context in detail.
8.	Implement policies for controlling/monitoring the experimental variables.
When the Study is Completed	
9.	Plan follow-up activities.
10.	Build or update a lab package.

if it is not clearly related to the course in the students' minds (which decreases motivation and performance).

Thus, while the researchers set research goals for the ESWS, the instructors, who are most familiar with the course material and able to determine how the study fits into the class materials, need to set educational goals. For example, the research goal for a study might be to *Compare the effectiveness of two inspection techniques*, while the educational goal would be to *Give students hands-on experience with the software inspection process*. The instructors also need to communicate the pedagogical value of the study in sufficient detail to the students so they are adequately motivated to participate in the study. We recommend that a short statement of the anticipated educational benefits be explicitly stated on the assignment sheet.

Considerations:

- If an ESWS is inserted into a course too early in subjects' education, they are likely to lack necessary skills and produce results that are neither internally nor externally valid.
- If a study is inserted into a specialized course later in subjects' education, care must be taken that it is in-line with the normal course topics, otherwise subjects will not be sufficiently motivated to pay proper attention or overcome the learning curve.
- The topic of experimentation itself should be considered for inclusion in the curriculum. Teaching developers-in-training about empirical study helps them learn how to evaluate processes and methodologies.
- Consider whether the use of team projects is reasonable for addressing the research and educational goals.
- Determine if the use of a student project is reasonable for the course.

Requirements Addressed:

- R1 - By making the subject population more similar to the target population, i.e. software professionals, an ESWS's external validity improves. Carefully integrating the study with the course helps ensure that all the subjects have sufficient knowledge about the method or technique under study, thus making them closer to professionals.

- R2 - This checklist item clearly fulfills the requirement for proper course integration.
- R4 - Researchers and instructors should be able to obtain a good assessment of the skills acquired by the subject during the course, allowing them to set a reasonable and realistic goal for the study.
- R5 - Integration of the study with the course allows the instructor to ensure that the subjects are taught the necessary skills, thereby obtaining a setting that is appropriate for the study's goals.
- R7 - Introducing experimentation into the curriculum helps ensure that students learn how to evaluate products and processes on their own.
- R8 - Integrating teamwork into the assignment / curriculum would address this requirement.
- R9 - Integrating a student project into the assignment / curriculum would address this requirement.

4.2 Integrate the Study into the Course Schedule

Students enroll in multiple courses and must properly allocate their effort among the various commitments. The schedule pressures caused by these various courses may affect the students' motivation for the ESWS. Students tend to allocate their effort to the activities that give them the highest perceived benefit (i.e. studying for a mandatory exam that is worth 30% of the grade in one course versus completing an optional assignment related to an ESWS that is only worth 5% in another course). Conflicts among courses cannot be totally avoided, but to maximize the quality of the student response, the experimenters should plan the ESWS to minimize conflicts. We have found it useful to check with colleagues and students at the university before scheduling the ESWS.

   Considerations:

- Students who are overloaded with commitments are usually seen as a threat to internal validity, because they may not actually complete all of the assigned tasks or may cut corners. However, this threat should be balanced against the threat to external validity that results from subjects with no schedule conflicts. For example, if the study goal is to provide insight into a professional environment, where schedule pressures are typically quite intense, then, the lack of schedule pressures in a classroom setting would be unrealistic.

   Requirements Addressed:

- R1 - Providing realistic conditions improves the external validity.
- R2 - Ensuring that the aspects of the study (training and activities) appear at the right time during the semester helps ensure proper integration with the course.
- R5 - Choosing the right time during the semester to conduct the study will ensure that the subjects have the proper skills to perform the required activities.

4.3 Reuse Existing Artifacts and Tools as Appropriate

During ESWS preparation, researchers should make a careful search for existing artifacts and tools that can be reused. In addition to saving time, reuse can facilitate comparisons with other studies that used the same artifacts, thereby increasing the value of the study. The scientific literature and the internet are good sources of artifacts, but local industry or

governmental offices may also provide real-life artifacts to make the ESWS more closely fit their needs.

Regardless of the source, researchers must carefully tailor artifacts to fit the study goal(s), when necessary, without introducing unexpected issues through last-minute modifications and without decreasing the comparability with other studies that used the same artifacts.

Considerations:

- Reusing artifacts saves considerable time and effort over creating new artifacts that are useful and representative.
- Reusing artifacts that have already been "tested" and "debugged" saves even more effort, and reduces the risk of unexpected threats to validity.
- Reusing artifacts and tools from real-world applications also reduces the threats to the external validity without incurring prohibitive cost. At the same time, this reuse gives the students more insights into professional development practices.

Requirements Addressed:

- R1 - The reuse of tools and artifacts that have been validated in other studies increases external validity.

4.4 Write Up a Protocol and Have It Reviewed

Prior to the study, a protocol, or set of steps to follow, should be developed and documented. Protocol development includes designing the study from a technical point of view, i.e. blocking, randomization approaches, and the identifying subject tasks. The protocol is tightly linked with the experimental design and it provides the researchers with a "recipe" to follow when conducting the study. It also facilitates communication with other researchers who might want to replicate the study. Once developed, the protocol should be reviewed by at least two sets of people: colleagues (research and education) and the governing ethics body, e.g. an Institutional Review Board (IRB), if the research is conducted in a country where such a review is required by law (e.g. the United States).

First, the importance of the review by colleagues is to ensure that both the research value and the pedagogical value are maximized. For research value, a protocol review should help ensure that the research questions are appropriate, proper measures are used, and the design is valid. For pedagogical value, a protocol review should help ensure that the subjects (students) will receive adequate educational value from the ESWS. In addition, when there are trade-offs between research value and pedagogical value (which is often the case) an independent and more objective review can help ensure that proper decisions have been made. This step is especially important when the same person assumes both the researcher and instructor roles.

Second, prior to conducting research that involves human subjects, researchers must obtain approval. Universities in the United States have an IRB that is responsible for reviewing and approving such studies. The IRB has the mandate of ensuring that human subject populations (especially vulnerable populations like students) are protected. In most cases, the researcher will need to complete a form to describe, in detail, the goals of the research and the protocol to be followed. The protocol description indicates what activities the subjects will perform and any potential risks, such as loss of privacy or reputation, the subjects face. Without approval, studies using human subjects cannot be conducted or published in the literature.

Considerations:

- The protocol review should be done far enough in advance to allow time to make any necessary adjustments to the study design.
- When designing the study, researchers and instructors should consider whether the use of group work is feasible in the design of the study.
- When designing the study, researchers and instructors should determine whether the study is a good candidate to include a development project.

Requirements Addressed:

- R1 - Third-party review of the protocol and design will help to identify as many external validity threats as possible.
- R2 - A review of the activities by the instructor of the course (or an educational colleague) will help ensure the study is properly integrated with the course.
- R3 - The review by the IRB or other research colleagues who can focus on ethical issues helps identify any important ethical concerns that must be addressed.
- R4 - Research colleagues can help identify mismatches between the study goal and the course within which the study will be conducted.
- R5 - The instructor of the course (or an educational colleague) can help ensure that the chosen course is the appropriate setting for the study based on its goals and design.
- R8 - Integrating teamwork into the assignment / curriculum would address this requirement.
- R9 - Integrating a student project into the assignment / curriculum would address this requirement (Note that not all studies can make use of a project, e.g. an inspection study.)

4.5 Obtain Subjects' Permission for Their Participation in the Study

Prior to the study, the instructor/researcher should inform the students of the high-level goals of the study, any possible adverse consequences for participation, and the measures taken to keep data anonymous. The subjects should then give explicit permission, using a consent form, before they participate in the ESWS and be given a chance to opt out of the study later. Many universities require that students give consent before any data collected can be used. Depending on the kind of data collected, the subject's permission may be required by law, so, even when permission is not required, we advise that it be asked for due to the increasing awareness of privacy and data property issues

Considerations:

- The consent form is a written agreement between the instructor/researcher and the students explaining if and how the results of the empirical study will influence the students' grades.
- A possible validity threat of requiring consent is that it may cause a self-selection, because only some students may be willing to participate in the study. However, it is not clear whether this threat would be greater than forcing unwilling subjects to participate. Making it clear to subjects that their data will remain anonymous helps researchers obtain better, less biased data.
- If the students are properly informed, they will feel more at ease with participating in an ESWS, feel less like "lab rats," and feel more like active participants. It is

also easier for the students to absorb the educational contents of the ESWS, rather than trying to guess the consequences that may result from activities performed during the study.

Requirements Addressed:

- R1 - Subjects who freely consent to participate are likely to provide more accurate data than those who feel coerced. Obtaining more accurate data will reduce the threat to external validity.
- R3 - This step fulfills one of the main ethical requirements: subject consent.

4.6 Set Subject Expectations

Subject motivation is fundamentally important to a valid ESWS. Special care should be devoted to explaining what is required of the students, if and how they will be compensated, and — to the extent possible — the pedagogical goals of the study. It is important to give students realistic time estimates, though human factors and other circumstances (e.g., the other commitments) may have an influence. When in doubt, we recommend erring on the side of over- rather than under-estimating time requirements. If the time requirements are underestimated, students may quit or become frustrated when the activity takes longer than expected, thereby reducing the quality of the data provided.

Starting with the consent form, instructors must specify prior to the ESWS if and how the empirical study will affect students' grades. An ESWS needs to have an educational value and may replace other forms of teaching. Thus, it may not necessarily be used for grading, especially when the ESWS involves a cutting-edge technological issue. In some cases, however, ESWSs may have some kind of grade incentives, so the criteria must be explained before the ESWS in the same way that grading criteria for other exercises are explained. Our experience shows that the grading criteria should be based on the process conformance and data quality, rather than the "quantity" of data generated. For example, in a study about a new inspection technique, it would not be appropriate to grade the students based on the number of defects reported. Rather, they should be graded based on how well they followed the technique and how much detail they provided about the defects they found. Otherwise, the students may artificially inflate the number of defects reported (even reporting things that are not true defects), and disregard the real goal of the exercise. At any rate, compensating students for their participation in an ESWS somewhat mirrors what happens in professional environments where employees are compensated according to the quality with which they perform their assignments.

Considerations:

- Sensible time and effort estimates and clear grading criteria make the students comfortable and minimize the risk of drop-outs or a rush to complete their task, resulting in missing or low quality data.
- When explaining the goals of the ESWS, the researcher should not disclose information that may bias the study. For example, the researcher could state that the ESWS will compare different methods for software validation rather than stating it would evaluate a new inspection technique. This approach eliminates the risk that students may (consciously or not) try to please the researchers by acting in a way to confirm the researchers' hypotheses. For example, they could put more effort or enthusiasm into application of what they perceive to be the new technique being researched.

Requirements Addressed:

- R1 - Setting appropriate subject expectations will increase the likelihood of obtaining representative data, thereby increasing external validity.
- R3 - By making the subjects aware of the activities they will perform this step addresses another important ethical consideration.

4.7 Document Detailed Information about the Experimental Context

Empirical studies in software engineering are in many respects more similar to those in the social sciences than those in the 'hard' sciences, due to the large influence of human factors. So, to run an effective study, it is important to gather and record context information, i.e. the specific characteristics and constraints that make the environment unique. Examples of the type of information that should be reported include: the name and general content of the course, the experience of the subjects (e.g., year in school, expertise in performing study tasks), and any unique constraints placed on the subjects (e.g., the experimental tasks had to be performed as homework rather than in-class work). When this context information is reported along with the results of the study, it helps other researchers and educators evaluate the ESWS along the dimensions introduced in Section 2.3.

Considerations:

- To properly interpret the results and compare them with the results from other studies, researchers normally collect background information about the subjects. In ESWSs, the instructors must also provide information about the goals of the course, the topics covered, and the teaching methods used.
- Interaction with professional organizations allows researchers to better understand the difference between the academic and professional environments and what kind of background information is useful to collect.

Requirements Addressed:

- R1 - A detailed description of the experimental context helps other researchers evaluate the external validity of the study and whether the results are relevant to their context.
- R2, R4 and R5 - A clear description of the course helps readers judge whether the course was an appropriate setting for the study.
- R6 - A clear description of the subjects who participated in the study is important for judging the differences between the subject population and the target population and evaluating the impact of those differences.

4.8 Implement Policies for Controlling/Monitoring the Experimental Variables

Factors influencing the study need to be controlled and monitored. Several different quantitative and qualitative measures may be collected during the empirical study. The same methods can be used to collect data during an ESWS that are used in empirical studies with practitioners, e.g., forms, interviews, timesheets, automated tools for extracting information from artifacts. The issues in ESWSs are similar to those that arise in other empirical studies, for instance: evidence should be collected in a timely fashion and not reconstructed a posteriori, if possible; data collection procedures should be minimally invasive; self-reported data may be less reliable than

automatically collected data; some types of evidence are difficult to collect, so cost-effectiveness must be assessed; some data may be sensitive, so protection and anonymity are important. Like the previous checklist item, this information also provides insight into the dimensions discussed in Section 2.3.

Considerations:

- An important constraint on controlled variables is that the educational value of the ESWS should be the same for all students. Suppose that the researcher wants to compare a new technique to an existing one. Ideally, the study design would have two groups with each using one technique. In this case, the students would learn different things during the ESWS. However, this problem can be remedied by having each group learn the other technique subsequent to the study.
- Students may be more comfortable with some of their classmates than with others and naturally group together if they are allowed to self-group. Such groupings may create an unrealistic situation (i.e. in the work environment, employees are structured into workgroups based on corporate needs rather than their own preferences) or invalid experimental conditions (i.e., one group may be biased over another on a particular variable of interest). Blocking or randomization may be used to avoid these natural groups and prevent the associated problems.
- Accuracy, invasiveness, and cost of data collection are especially important in a professional environment. An ESWS is able to test the data collection procedures and show the associated costs.

Requirements Addressed:

- R1 - By properly monitoring the experimental variables, the researcher increases confidence in the quality of data collected, thereby increasing external validity.
- R3 - By ensuring that each student receives the same value from the study, this step helps fulfill an important aspect of pedagogical ethics.
- R6 - A careful monitoring and documentation of the experimental variables, especially subject variables, will provide information needed to judge the subject population against the target population.

4.9 Plan Follow-up Activities

This important step is often overlooked in empirical studies. Follow-up activities can be performed in several ways, e.g., questionnaires, interviews, and class discussions in which the goals of the study and preliminary results are presented. Professionals may benefit from the discussion as well. Professionals may also provide important insights into the practical usefulness of the results and suggest improvements to the study design.

Considerations:

- Follow-up activities provide very valuable feedback to researchers. Questionnaires and interviews make it easier to obtain information from students who feel somewhat uncomfortable during class discussions. Interviews, though more time-consuming, are often useful, even after questionnaires, because they can be more focused. Class discussions provide information that may not show up on a questionnaire. Once students hear the comments of others, they may provide feedback not recorded on a

questionnaire. Any of these activities can help identify, at least after the fact, possible threats to validity. These activities can also help researchers obtain more detail about process conformance or feedback about alternate explanations not yet considered.

• Follow-up activities are an essential teaching opportunity for both the object of study in the ESWS, i.e. the students learn about the results of the study, and for empirical techniques, i.e. the students begin to learn how to conduct their own studies.

• The feedback also fulfills the researchers' ethical obligations both to the students and to other researchers. For the students, the feedback session provides the researcher with the opportunity to elaborate any important study details that were kept secret from the subjects during the study (to prevent biasing the results). For other researchers, the feedback session helps the researcher ensure that the conclusions are consistent with what actually occurred during the study.

Requirements Addressed:

• R1 - Evaluating the validity of the results through follow-up helps the researcher properly report any external validity problems.

• R2 - Feedback from the students will help the researcher and instructor understand whether the study was properly integrated with the course.

• R3 - By providing full disclosure of all study details, the researcher fulfills an important ethical obligation.

• R6 - If the follow-up session includes the presence of a member of the target population (a professional), the researchers will be able to better understand which characteristics of the subject population do not match.

• R7 - During the follow-up discussion, the students learn how the study was conducted. This information helps them understand how to conduct their own studies.

4.10 Build or Update a Lab Package

Replication of an ESWS in an educational or professional environment is extremely important (Shull et al. 2008). Building a lab package after a study is completed helps save effort on a replication and makes explicit any possible mistakes that may have been made. By documenting experiences and mistakes the researcher allows the community to learn. To build an accurate lab package, all the details of an empirical study must be recorded and tracked. This step will facilitate the reuse of artifacts and tools in future studies (see Section 4.3).

It is important that a lab package not be seen as a statement of an "exemplary" design to be reused "as is." While replicators may indeed choose to apply a design from a lab package with no changes, replications which vary some aspects of the design and the artifacts increase generalizability. Replications which obtain the same results using different designs or artifacts greatly increase the confidence in those results. Replications that reuse the same design without change run a risk of replicating the mistakes along with the rest of the experiment, and hence may end up providing additional support for a spurious result (Daly 1996; Wood et al. 1999; Miller 2005).

Considerations:

• Lab packages should also be viewed as a mechanism to support communication among researchers. Often a lab package is the only means other researchers have for examining the protocols and artifacts in sufficient detail to provide well-founded critiques.

• It is certainly in the researchers' best interest to build a lab package that can be used by other researchers to confirm the results, and corroborate external validity. Also,

recording all the details of an empirical study may help explain possible discrepancies among the results found in different environments and highlight the source of outliers.

- Students who participate in future replications will benefit from a lab package because the artifacts and tools can be evaluated and used by other students and professionals. No design or artifact is perfect the first time. When made available for review and extension, the entire community can evolve and improve artifacts, rather than only a single group.
- This step is one of the most beneficial to professionals, because a lab package can be readily reused in a professional environment. It may require some tailoring and modifications, but will cost less than building the study from scratch.

   Requirements Addressed:

- *R1-R9* - The production of the lab package does not directly address any of the requirements. But, by providing a detailed description of the information about the study, the lab package helps other researchers and educators understand how all of the requirements were addressed.

4.11 Mapping of Requirements to Checklist Items

To summarize the discussion in Section 4, Fig. 1 shows the mapping of the checklist items to the requirements they address. Some of the requirements are definitely addressed when the checklist item is completed. These items are marked with a '√'. Other items are only addressed if certain conditions are met, as explained in the earlier sections. These items are marked with a '?'.

5 Examples of Checklist Use

As empiricists, we realize it is important to provide some evidence of the feasibility or usefulness of the reported results. Although we cannot yet quantify the benefits of using the checklist, we note that versions of this checklist have been used to support the design and execution of 20 ESWSs within HPCS over the last 3 years. (Hochstein et al. 2006). In addition, to provide some

Checklist Items	R1	R2	R3	R4	R5	R6	R7	R8	R9
1	√	√		√	√		√	?	?
2	√	√			√				
3	√								
4	√	√	√	√	√			?	?
5	√		√						
6	√		√						
7	√	√		√	√	√			
8	√		√			√			
9	√	√	√			?	√		
10	√	√	√	√		√	√	√	√

Mapping of Requirements to Checklist Items
(√ = definitely addressed; ? = may be addressed)

Fig. 1 Mapping of requirements to checklist items. (√=definitely addressed; ?=may be addressed)

initial validation of checklist, one of us has participated in conducting a study that followed the checklist. As a method of illustration, we will discuss how each checklist item was followed and provide a description of the positive research and educational aspects that resulted.

The study focused on understanding whether architectural approach and task order affected the maintainability of object-oriented software. Subjects were given a system that was implemented following one of two architectural approaches (Delegated or Centralized). Then students were asked to make a series of changes to that system that ranged in difficulty from easy to difficult (Wang et al. 2007; Wang and Arisholm 2008). A detailed discussion of the technical details of the study is not relevant to this paper, so we focus on the use of the checklist and its impact on the results of the study.

1. ***Ensure adequate integration of the study into the course topics*** — The research goal was to *Evaluate the impact of the architectural approach (Delegated vs. Centralized) on the duration and correctness of a series of change tasks*. The educational goal of the study was that *Students should learn to define and explain central concepts in the software architecture (SWA) domain and learn to use and describe design patterns, methods to design SWA, methods and techniques to achieve software qualities, methods to document SWA, and methods to evaluate SWA* (Wang et al. 2007). The study was conducted in a Software Architecture course whose goals included: learning how to use various design approaches, understanding software quality (e.g. maintainability) and participating in a practical, hands-on experience. The experiment took the place of an assignment on design patterns that is normally used in this course. Therefore, by conducting the study in the place of another comparable assignment and by using the assignment to provide the students with a hands-on experience, the study was properly integrated into the course.

2. ***Integrate the study timeline with the course schedule*** — This step was completed by specifically choosing the point in the semester when the students would be most adequately prepared to carry out the tasks required by the study. The timeline for the study was then planned according to this information.

3. ***Reuse artifacts and tools where appropriate*** — This study was a replication. Therefore, all of the artifacts from the previous study were reused. In addition, the researchers used an infrastructure tool to support the experimentation process.

4. ***Write up a protocol and have it reviewed*** — Because this study was a replication, there was no need for to write up a formal protocol and have it reviewed. Rather than having a new protocol reviewed, the replicating researchers obtained the same benefit by making use of the lessons learned by the original experimenters.

5. ***Obtain subjects' permission for their participation in the study*** — This study was conducted outside of class time. The subjects were allowed to perform an alternate task if they chose not to participate. Therefore, by showing up for the study, the students consented to participate.

6. ***Set subject expectations*** — Prior to the study, the students were told the purpose of the study (in enough detail not to bias the results), what materials they should bring, how much time it would require, and the amount of compensation they would receive.

7. ***Document information about the experimental context in detail*** — Detailed information was documented about the experience of the subjects, the nature of the course, and the tasks that were performed. This information is included in the research report.

8. ***Implement policies for controlling/monitoring the experimental variables*** — An automated environment was used to allow the students to download the code and task descriptions, upload task solutions and answer questionnaires. This environment kept

the subjects attention on their main task rather than on recording data. In addition, the students were given an anonymous survey upon completion of the study to gather subjective feedback.

9. ***Plan follow-up activities*** — Later in the semester, the researchers presented the results of the study to the students. During this presentation, the researchers informed the students about the background and context of the study, the study methods that were used, the experimental design and the analyses conducted. Finally, the results of an anonymous survey taken at the end of the study were presented. The students were given an opportunity to provide oral feedback on the entire experience. Therefore, in addition to learning the technical course material, the students also learned something about how to conduct empirical studies.

10. ***Build or update a lab package*** — Because this study was a replication, there was already an existing lab package.

At the conclusion of the study, the students were asked to evaluate how well the experiment fit into the course. Two-thirds of the students agreed that the study was "relevant to the course." When asked whether empirical studies should be conducted in courses, 60% agreed. Next, an open-ended question gathered the benefits of using the experiment in the course. Finally, as an external judge on validity, this study produced a conference paper related to the educational aspects (Wang et al. 2007) and a journal paper related to the research aspects (Wang and Arisholm 2008).

6 Conclusions and Future Work

In this paper, we have presented knowledge gained as software engineering researchers and educators from our experiences in planning and conducting a large number of ESWSs in three countries (Italy, Norway, and the United States). A discussion of relevant work from the empirical software engineering literature and the software engineering education literature was used to generate requirements for conducting valid ESWSs. Furthermore, to ensure that the discussion properly built on educational theory, we also reviewed relevant education principles.

The literature review conducted for this work was of an interdisciplinary nature, at the intersection between software engineering theory and pedagogical theory. The purpose of the review was both to provide a summary of the pedagogical theory that has guided our work and to stimulate the reader to appreciate the importance of pedagogical issues when conducting ESWS. This appreciating might drive some to seek out dialogue with educational experts in their institutions. In the future, the literature review could be expanded by including both software engineering researchers and education researchers in the review team.

Using the information from the literature, we presented a discussion of how to assess both the experimental and pedagogical value of an ESWS. We note that in the literature there is a relatively large amount of work focused on experimental value, but relatively little focused on pedagogical value. However, in order for ESWSs to be successful, they must have both experimental and pedagogical value. We then presented a checklist to guide researchers and instructors in conducting ESWSs. The checklist contains information about the steps that should be taken before, during, and after ESWSs to help ensure a successful study that balances the needs of stakeholders and meets the important requirements identified. As an extension of this work we plan

to implement simple web support for the checklist. This support will enable the checklist users to navigate from goals to checklist items, examples, and other experimental material related to the checklist.

While we do provide an initial evaluation of the checklist through use on a real study, it has been developed largely on the experiences of the four authors. Therefore, our future work includes a more thorough, and external, evaluation. We plan to evaluate the checklist by surveying and interviewing representatives of the four types of stakeholders. The goal of this investigation will be to improve the checklist in three ways. First, we want to ensure that we have the correct set of stakeholders with their most important goals and risks. Second, we want to ensure that the checklist contains the right set of steps. This improvement may require the addition, deletion or merging of steps. Finally, we want to ensure the clarity and completeness of description of each checklist item and to clarify any items that are confusing or too abstract.

References

Bagert D, Hilburn TB, Hislop G, Lutz M, McCracken M, Mengel S (1999) Guidelines for Software Engineering Education. SEI Technical Reports. CMU/SEI-99-TR-032

Baresi L, Morasca S (2002) An empirical study on the design effort of web applications. Proceedings of 3rd International Conference on Web Information Systems Engineering, Singapore. 345–354

Baresi L, Morasca S, Paolini P (2003) Estimating the design effort of web applications. Proceedings of Ninth International Software Metrics Symposium. 62–72

Basili VR, Briand LC, Melo WL (1996) A validation of object-oriented design metrics as quality indicators. IEEE Trans Softw Eng 22(10):751–761. doi:10.1109/32.544352

Basili VR, Carver JC, Cruzes D, Hochstein LM, Hollingsworth JK, Shull F, Zelkowitz MV (2008) Understanding the high-performance-computing community: a software engineer's perspective. IEEE Softw 25(4):29–36. doi:10.1109/MS.2008.103

Bloom BS (ed) (1956) Handbook I, cognitive domain. Taxonomy of educational objectives: the classification of educational goals. Longman, New York

Braught G (2005) Teaching empirical skills and concepts in computer science using random walks. Proceedings of 36th SIGCSE technical symposium on Computer science education, St. Louis, Missouri, USA, ACM Press. 41–45

Carver J, Jaccheri L, Morasca S, Shull F (2003) Issues in using students in empirical studies in software engineering education. Proceedings of Ninth International Software Metrics Symposium (METRICS 2003). 239–249

CORPORATE (2001) Computing Curricula 2001. J Educ Resour Comput 1(3es):1

Daly J (1996) Replication and a multi-method approach to empirical software engineering research. Department of Computer Science. University of Strathclyde. PhD

Denning PJ (1992) Educating a new engineer. Commun ACM 35(12):82–97. doi:10.1145/138859.138870

Hilburn TB, Humphrey WS (2002) The impending changes in software education. IEEE Softw 19(5):22–24. doi:10.1109/MS.2002.1032848

Hochstein L, Nakamura T, Basili VR, Asgari S, Zelkowitz MV, Hollingsworth JK, Shull F, Carver J, Voelp M, Zazworka N, Johnson P (2006) Experiments to understand Hpc time to development. *CTWatch Quarterly*. November: 24–32

Höst M (2002) Introducing empirical software engineering methods in education. Proceedings of 15th Conference on Software Engineering Education and Training, 2002. (CSEE&T 2002). 170–179

Höst M, Regnell B, Wohlin C (2000) Using students as subjects-a comparative study of students and professionals in lead-time impact assessment. Empir Softw Eng 5(3):201–214. doi:10.1023/A:1026586415054

Höst M, Wohlin C, Thelin T (2005) Experimental context classification: incentives and experience of subjects. Proceedings of 27th international conference on Software engineering, St. Louis, MO, USA, ACM Press. 470–478

Jaccheri L (2001) Software quality and software process improvement course based on interaction with the local software industry. Comput Appl Eng Educ 9(4):265–272. doi:10.1002/cae.10000

Jay R (2002) How to build a great team, Financial Times Management

Jorgensen M, Teigen KH, Molokken K (2004) Better sure than safe? Over-confidence in judgement based software development effort prediction intervals. J Syst Softw 70(1–2):79–93. doi:10.1016/S0164-1212(02)00160-7

Kitchenham BA, Pfleeger SL, Pickard LM, Jones PW, Hoaglin DC, El Emam K, Rosenberg J (2002) Preliminary guidelines for empirical research in software engineering. IEEE Trans Softw Eng 28(8):721–734. doi:10.1109/TSE.2002.1027796

McBurney DH (2001) Research methods. Wadsworth Thomson Learning, Stamford

Miller J (2005) Replicating software engineering experiments: a poisoned chalice or the Holy Grail. Inf Softw Technol 47(4):233–244. doi:10.1016/j.infsof.2004.08.005

Morasca S (2003) A Bayesian approach to software testing evaluation. Proceedings of Software Engineering and Knowledge Engineering, San Francisco Bay, USA. 706–713.

Pastel R (2005) Integrating science and research in a Hci design course. St Louis, ACM Press

Port D, Klappholz D (2004) Empirical research in the software engineering classroom. Proceedings of 17th Conference on Software Engineering Education and Training, 2004. 132–137

Shull F, Lanubile F, Basili VR (2000) Investigating reading techniques for object-oriented framework learning. IEEE Trans Softw Eng 26(11):1101–1118. doi:10.1109/32.881720

Shull F, Carver J, Travassos G (2001) An empirical methodology for introducing software processes. Proceedings of The Joint 8th European Software Engineering Conference and 9th ACM SIGSOFT Foundations of Software Engineering, Vienna, Austria. 288–296

Shull F, Carver J, Hochstein L, Basili VR (2005) Empirical study design in the area of high performance computing (Hpc). Proceedings of International Symposium on Empirical Software Engineering, Noosa Heads, Australia. 305–314

Shull F, Carver J, Vegas S, Juristo N (2008) The role of replications in empirical software engineering. Empir Softw Eng 13(2):211–218. doi:10.1007/s10664-008-9060-1

Singer J, Vinson NG (2002) Ethical issues in empirical studies of software engineering. IEEE Trans Softw Eng 28(12):1171–1180. doi:10.1109/TSE.2002.1158289

Sjoeberg DIK, Anda B, Arisholm E, Dyba T, Jorgensen M, Karahasanovic A, Koren EF, Vokac M (2002) Conducting realistic experiments in software engineering. Proceedings of 2002 International Symposium on Empirical Software Engineering. 17–26.

Sjoeberg DIK, Hannay JE, Hansen O, Kampenes VB, Karahasanovic A, Liborg NK, Rekdal AC (2005) A survey of controlled experiments in software engineering. IEEE Trans Softw Eng 31(9):733–753. doi:10.1109/TSE.2005.97

Tichy WF (2000) Hints for reviewing empirical work in software engineering. Empir Softw Eng 5(4):309–312. doi:10.1023/A:1009844119158

Umphress DA, Hendrix TD, Cross JH (2002) Software process in the classroom: the capstone project experience. IEEE Softw 19(5):78–81. doi:10.1109/MS.2002.1032858

Vygotsky LS (1978) Mind in society: development of higher psychological processes. Harvard University Press, Cambridge

Walia GS, Carver J (2006) Requirements error abstraction and classification: an empirical study. Proceedings of The 5th International Symposium on Empirical Software Engineering, Rio de Janeiro. 336–345

Wang AI, Arisholm E (2008) The effect of task order on the maintainability of object-oriented software. Information and Software Technology. Accepted for Publication: Technical Report Available: http://www.idi.ntnu.no/grupper/su/publ/alfw/idi-tr-02-07.pdf

Wang AI, Arisholm E, Jaccheri L (2007) Educational approach to an experiment in a software architecture course proceedings of The 20th Conference on Software Engineering Education & Training, 2007. 291–300

Way TP (2005) A company-based framework for a software engineering course. Proceedings of 36th SIGCSE technical symposium on Computer science education, St. Louis, Missouri, USA, ACM Press. 132–136

Wohlin C, Runeson P, Host M, Ohlsson MC, Regnell B, Wesslen A (2000) Experimentation in software engineering: an introduction, Kluwer Academic Publishers

Wood M, Daly J, Miller J, Roper M (1999) Mulit-method research: an empirical investigation of object-oriented technology. J Syst Softw 48(1):13–26. doi:10.1016/S0164-1212(99)00042-4

Zelkowitz MV, Wallace DR (1998) Experimental models for validating technology. IEEE Comput 31(5):23–31

Jeffrey Carver is an Assistant Professor in the Department of Computer Science at the University of Alabama. He received his PhD in Computer Science from the University of Maryland. His main research interests include empirical software engineering, software quality, software inspections, software architecture, and human factors in software engineering. He is a member of the IEEE and the ACM.

Letizia Jaccheri is a Professor at the Department of Computer and Information Science at the Norwegian University of Science and Technology (NTNU) since 2002. In the last five years, she has been involved in the supervision of at least 10 PhD students. She has been working with software research issues since 1988. She wrote her PhD on software process modeling in 1994. She is interested in software intensive processes with special focus on artistic software and open source software.

Sandro Morasca is a Professor of Computer Science at the Università degli Studi dell'Insubria in Como, Italy. In the past, he was an Associate Professor and Assistant Professor at the Politecnico di Milano in Milano, Italy. He was a Faculty Research Assistant and later a Visiting Scientist at the Department of Computer Science of the University of Maryland at College Park. Sandro Morasca has been actively carrying out research in the Software Engineering field in Empirical Software Engineering, Specification of Concurrent and Real-time Software Systems, Software Verification, and Open Source Software, and has published about 20 journal papers and 50 conference papers. Sandro Morasca has been involved in a number of national and international projects. He is currently the Leader of the activity related to the trustworthiness of Open Source Software in the QualiPSo project, financed by the European Union. Sandro Morasca has served on the PC of a number of international software engineering conferences and serves on the editorial board of "Empirical Software Engineering: An International Journal," published by Springer-Verlag.

Dr. Forrest Shull is a Senior Scientist at the Fraunhofer Center for Experimental Software Engineering in Maryland (FC-MD), a nonprofit research and tech transfer organization, where he leads the Measurement and Knowledge Management Division. At FC-MD, he has been a lead researcher on projects for NASA's Office of Safety and Mission Assurance, the NASA Safety Center, the U.S. Department of Defense, the National Science Foundation, the Defense Advanced Research Projects Agency (DARPA), and companies such as Motorola and Fujitsu Labs of America. He is an associate adjunct professor at the University of Maryland College Park, and Associate-Editor-in-Chief of *IEEE Software*.