Flexibility in Research Designs in Empirical Software Engineering

Vigdis By Kampenes [a,b], Bente Anda [a,b], Tore Dybå[a,c]

[a]Department of Software Engineering, Simula Research Laboratory, P.O. Box 134, NO-1325 Lysaker, Norway
[b]Department of Informatics, University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
[c]SINTEF ICT, NO-7465 Trondheim, Norway

vigdis@simula.no, bentea@simula.no, tore.dyba@sintef.no

Abstract

Problem outline: A common way of classifying empirical research designs is in *qualitative* and *quantitative designs.* Typically, particular research methods (e.g., case studies, action research, experiments and surveys) are associated to one or the other of these designs. Studies in empirical software engineering (ESE) are often exploratory and often involve software developers and development organizations. As a consequence, it may be difficult to pre-plan all aspects of the studies, and to be successful, ESE studies must often be designed to handle upcoming changes during the conduct of the study. A problem with the above classification is that it does not cater for the *flexibility* in the design.

Position: This paper suggests viewing research in ESE along the axis of flexible and fixed designs, which is both orthogonal to the axis of quantitative and qualitative designs, and independent of the particular research method. According to the traditional view of ESE, changes to the research design in the course of a study are typically regarded as threats to the validity of the study results. However, by viewing the study designs as flexible, practical challenges can be turned to useful information. The validity of the results of studies with flexible research designs can be established by applying techniques that are traditionally used for qualitative designs. This paper urges for increased recognition of flexible designs in ESE and discusses techniques for establishing the trustworthiness in flexible designs.

Keywords: *Empirical software engineering, flexible research design, fixed research design, trustworthiness.*

1. INTRODUCTION

Empirical software engineering (ESE) studies often involve humans, as individuals or as part of a software development organization, with their own constraints and expectations. It may be difficult for the researcher to predict these constraints and expectations, but in practice the researcher must adapt to them throughout the research (Anda, Hansen et al. 2006; Conradi, Dybå et al. 2006). Furthermore, studies in empirical software engineering are seldom based on established theories (Hannay, Sjøberg et al. 2007), and as a consequence elements of the research design, such as the research question or the concepts investigated, may need refinement during the study. These features of research in empirical software engineering demand the researcher to be flexible, managing research that takes unanticipated directions. To enable the researcher to be flexible, the research design must also be flexible.

We define flexibility simply as *the capacity to adapt* (Golden and Powell 2000), although a number of alternative definitions of flexibility of projects and organizations exist, see for example (DeLeeuw and Volberda 1996; Golden and Powell 2000; Olsson 2006).

Research designs are commonly classified into quantitative and qualitative designs, where only a qualitative design has flexibility. The terms quantitative and qualitative are also used for the data collected in the empirical studies. Our experience is that using the same terms both for the characteristics of the data collected and the features of the research design leads to some confusion among software engineering researchers. Qualitative designs can, for example, incorporate quantitative methods of data collection. To better describe the degree of flexibility in research designs, leaving type of data optional, Anastas and MacDonald (1994) and Robson (2002) use the terminology of *fixed* and *flexible designs* in social science. Because there is a need for increased awareness of the flexibility in ESE research designs, we suggest that this terminology or perspective of design is used also in ESE.

In fixed designs, the design is specified early in the research process, whereas in flexible designs, the design is allowed to evolve during the research. Type of design is orthogonal to type of data collected, that is, both qualitative and quantitative data may be used in both fixed and flexible designs. A research design may be either completely fixed, completely flexible or have degrees of flexibility. We believe that there are completely flexible designs conducted in ESE, but that these typically follow the traditional qualitative framework, collecting qualitative data in ethnographies, action research or exploratory case studies. In our experience, also other studies in empirical software engineering face some form of uncertainty in the planning phase, requiring a degree of flexibility in the design. Hence, our main concern is to find a design perspective that embraces also these studies. In particular we believe that many experiments need a degree of flexibility in the design. A literature review of type of evidence produced by empirical software engineers, performed by Segal (2005), shows that laboratory experiments dominate evaluations. Hence, our perspective might influence many empirical studies in software engineering.

Our main aims with this paper are to increase the awareness of why and how some ESE studies are flexible, and to initiate a discussion of how to handle this flexibility and simultaneously conduct methodologically sound research. We suggest using flexible designs, including appropriate techniques for establishing trustworthiness.

The remainder of this paper is organized as follows. Section 2 describes the features of fixed and flexible designs and gives examples of how the need for flexibility occurs. Section 3 suggests factors to consider when choosing a design. Section 4 suggests techniques for establishing trustworthiness in studies with a flexible design, and Section 5 concludes.

2. TYPES OF RESEARCH DESIGNS

We consider a research design to consist of the following elements: *purpose(s)*, *theories*, *research questions*, *methods* and *sampling strategies*, see Figure 1. This model is described more thoroughly in (Robson 2002). Both the purpose(s) and the theory help specify the research questions. When the research questions are specified, decisions can be made regarding the methods to use and the sampling strategies. The methods include the research strategy, for example case study, survey or experiment; constructs and measures; the data collection methods, for example interviews, observations or questionnaires; analysis methods; techniques for establishing trustworthiness in the study and the research schedule. Finally, the sampling strategy includes descriptions of the study units and how to select them.

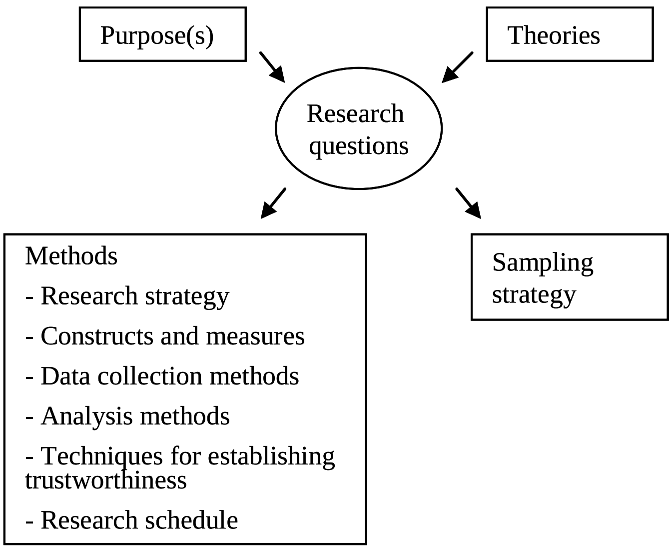


FIGURE 1: A framework for a research design

In a research study following a fixed design, the elements in Figure 1 must be specified before the data collection starts. More specifically, applying a fixed research design means to follow a procedure of research as visualized in Figure 2a. The idea generation and designing of the research are made in the beginning of the procedure, and here the plans for the data collection and analyses are made. This type of research includes any types of methods or data as long as they can be specified early in the research process.

Examples of typical types of fixed designs are experiments that test theories and use statistical methods as a decision tool for drawing conclusions (Arisholm, Gallis et al. 2007), replications (Laitenberger, Emam et al. 2001), systematic reviews of well understood phenomena (Kitchenham, Mendes et al. 2006), and surveys that are based on questionnaires (Dybå 2005).

Another example of fixed designs is studies that are not necessarily based on theories, but that has a short time schedule that allows no flexibility. For example, experiments performed at developer seminars (Grimstad and Jørgensen 2007).

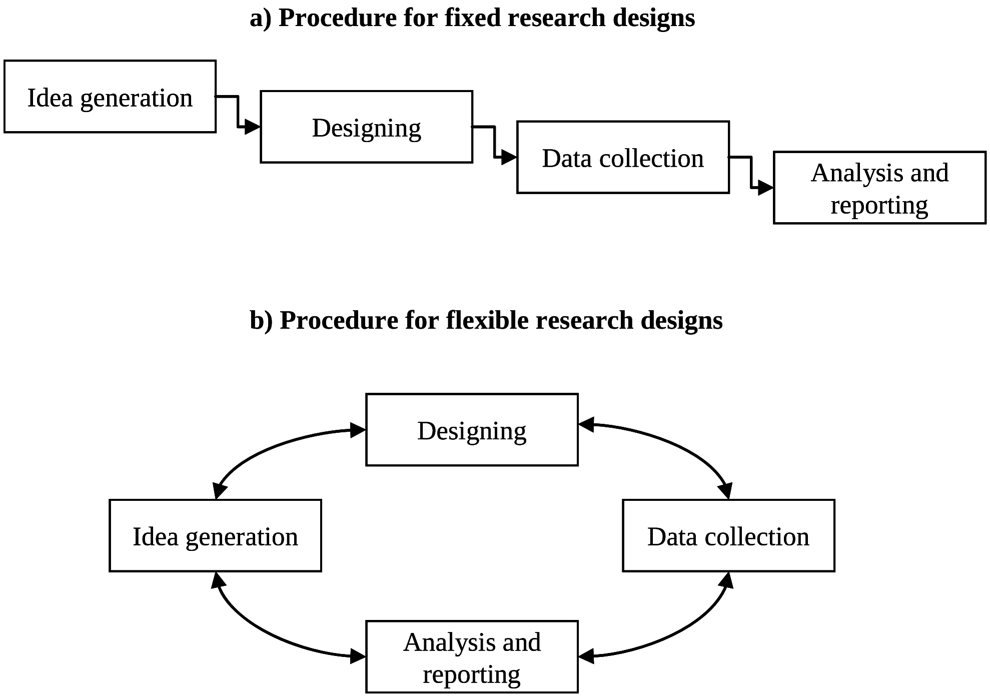


FIGURE 2: Visualizations of the procedures following a fixed research design and a flexible research design.

In flexible research designs, the design components in Figure 1 are specified during the course of the study. Hence, when applying a flexible research design, the methods of inquiry evolve incrementally in response to the data obtained (Robson 2002). Idea generation, designing, data collection and analysis and writing proceed together or in iterations rather than in separate stages, see Figure 2b. So, whereas the procedure of following a fixed research design is analogue to the waterfall method of designing software, the procedure of following a flexible research design is similar to iterative software development or agile methods.

The researcher's inability to fix one or several design elements in the beginning of the research, as well as practical constraints during the study, create the need for a flexible design. Examples are the specification of research questions, constructs and measures, and the research schedule:

- *Research questions.* A study may set out with a tentative research question that is refined in the course of the study, because the understanding of the phenomenon under study and of what can actually be studied empirically, increases.
- *Constructs and measures.* The mostly immature theories in software engineering mean that there will often be a corresponding lack of established constructs associated with the phenomenon under study. Also, the constructs may lack empirical validation. Consequently, constructs and measures may be refined during the study. Moreover, the knowledge about potential data sources and their quality may be limited at the outset of a study, so that the data collection must be adapted to the actual data available.
- *Research schedule.* The research schedule may have to be revised during the research, due to unforeseen events.

To give practical examples of how the need for flexibility occurs, we present in Table 1 experiences from two studies in ESE: a systematic review of the literature and a series of experiments. These studies were initially planned with a fixed design. However, because of lack of theories of the phenomenon under investigation, the designs appeared to have a need for some flexibility.

TABLE 1: Examples of how flexibility might occur in studies in ESE

Example 1.

A systematic review of 113 software engineering experiments (Sjøberg, Hannay et al. 2005; Kampenes, Dybå et al. 2007; Kampenes, Dybå et al. 2008)

The review was a quantitative investigation of the literature on experimentation over a decade. The first part of the review selected the relevant articles and summarized the characteristics of the experiments. The last part of the review investigated effect size reporting and quasi-experimentation.

Research question: In the last part of the review, the initial research question asked whether there was a difference in effect sizes between randomized experiments and quasi-experiments. This appeared to be difficult to answer, because the experiments did not include the necessary information for estimating the effect sizes. As a consequence, the review ended up with investigating the state of practice of effect size reporting and quasi-experimentation.

Constructs and measures: In the first part of the review, the operational definition of a software engineering experiment was refined throughout the review. Several researchers were involved in parts of the study and the final inclusion criteria were the results of several discussions. Also, the definition of effect size changed throughout the study, and ended up including the unstandardized effect size, because this type appeared to be reported in some articles and seemed very useful for describing the practical importance of the result. Furthermore, types of quasi-experiments in software engineering were not known in advance and therefore, the description of quasi-experiments was continuously changed.

Research schedule: The time schedule was continuously revised throughout the review.

Lessons learned: The decision of not following the initial plan, but account for new insight during the work, was important for the final quality of the review. All the refinements of research questions, and constructs and measures were valuable for the final results. However, the iterative process made the study more resource-demanding than planned; a flexible design requires a flexible budget. The iterative process was sometimes frustrating. If we had known the framework of flexible designs, we would probably have been more comfortable with all the refinements.

Pre-review mapping and piloting the review protocol, as suggested by Brereton, Kitchenham et al. (2007), might have helped to reduce the number of iterations. However, we experienced that many changes appeared late in the process and a flexible approach would still have been valuable for this type of review.

Example 2.

A series of three laboratory experiments investigating effects of different ways of applying use cases in the construction of class diagrams (Anda and Sjøberg 2003; Syversen, Anda et al. 2003; Anda and Sjøberg 2005)

The first experiment was a pilot study with 26 students as participants, then an experiment was conducted with 53 students as participants, and finally an experiment was conducted with 22 professional software developers. The experiments were motivated by common claims in software engineering textbooks, but there were no established theories on the topic.

Research question: The initial research question was whether there was a difference, regarding time spent on design and quality of the final class diagrams, between a use case driven development process and a responsibility-driven development process. During the analysis and writing up of the experiments, we realized, however, that the experiment had compared a more specific aspect of the two processes, the transition from use cases to class diagrams. Consequently, the research question was changed to whether there was a difference, regarding time spent on design and quality of the final class diagrams, when classes where derived by analyzing the use cases compared to when the use cases are used to validate the class diagram.

Constructs and measures: The exploratory nature of the experiments meant that the constructs used for the independent variable, the process, and for one of the dependent variables, quality, were not well established at the outset. Therefore, qualitative data was collected during the experiments to allow us to understand how the participants worked when solving the experimental tasks. The assessment of the quality of the final solutions was qualitative and was slightly refined based on the actual data.

The procedure for data collection mostly remained as planned at the outset of the study, but there were some changes to how data was collected due to specifics of each experiment. Also a few of the participants did not manage to follow the process description that was part of the experimental material, and their solutions were discarded.

Research schedule: The study procedure was revised for each experiment.

Lessons learned: Revising the initial research questions and central constructs during the course of the study was important for the quality of the final study, as it allowed us to take into account what we had previously learned. Furthermore, the collection of qualitative data, in particular about how the participants worked during the study, was valuable in ensuring the validity of the results.

Conducting a pilot experiment is recommended in empirical research before fixing the design for the main experiment. Therefore, some revisions of the research design are catered for also in the existing literature on software engineering experiments. In this case the first experiment can be characterized as a pilot. Our experience is, however, that it was difficult to fix all aspects of the design based on the relatively small pilot and some flexibility was useful also in the later experiments.

3. CHOOSING AN APPROPRIATE RESEARCH DESIGN

The researcher must early in the research process decide whether to use a fixed design or a design that accounts for a certain degree of flexibility. In this decision process, we suggest considering *the maturity of the research*, *the purpose of the research*, *the research setting* and *the time schedule of the research*.

Maturity of research can be catalogued into the extent of previous work in the field, for example nascent, intermediate and mature theory, see Table 2.

Research purpose is commonly divided into exploratory, descriptive and explanatory, see Table 3. The purpose of the research often depends on the maturity of the research, but not in a deterministic way. They represent two different perspectives, and both must be considered when choosing a design. In general, the less known about a specific topic, the larger flexibility in the design. However, the research setting and the time schedule must also be considered.

TABLE 2: The maturity of the research

Mature theory presents well-developed constructs and models that have been studied over time with increasing precision by a variety of scholars.

Intermediate theory presents provisional explanations of phenomena, often introducing a new concept and proposing relationships between it and established constructs. Although the research question may allow the development of testable hypothesis, similar to mature theory research, one or more of the constructs involved is often still tentative, similar to nascent theory research.

Nascent theory proposes tentative answers to novel questions and suggests new connections among phenomena.

(Edmondson and McManus 2007)

TABLE 3: The purpose of the research

Exploratory research: Research in which the primary purpose is to examine a little understood issue or phenomenon to develop preliminary ideas and move toward refined research questions by focusing on the “what” question.

Descriptive research: Research in which the primary purpose is to “paint a picture” using words or numbers and to present a profile, a classification of types, or an outline of steps to answer questions such as who, when, where, and how.

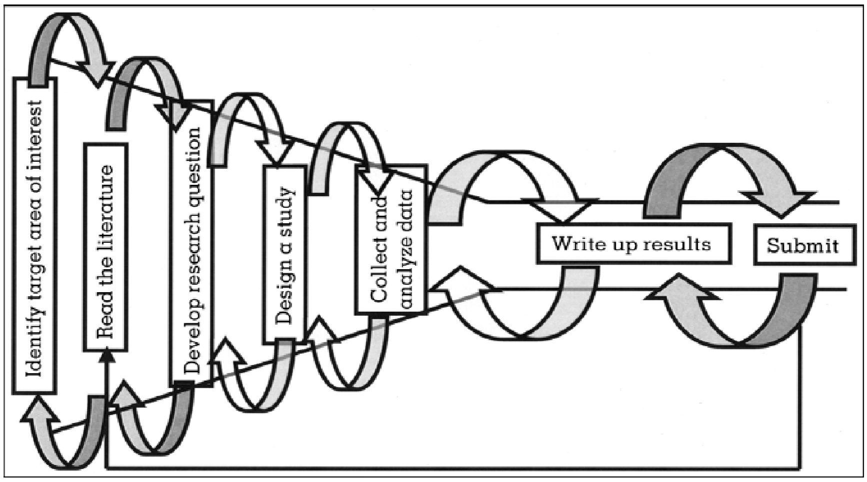
Explanatory research: Research in which the primary purpose is to explain why events occur and to build, elaborate, extend, or test theory.

(Neuman 2006)

The research setting can be divided into two categories, based on the extent of control. In laboratories and classrooms, more control is possible compared to studies conducted in a field setting. The controlled setting may enable a fixed design, even if the study is exploratory, whereas the field setting often requires a flexible design. Edmondson and McManus (2007) describe the process of management field research as a journey that may involve almost as many steps backward as forward, in an iterative way. We interpret their description to fit well into the perspective of a flexible design. Moreover, they argue that this iteration is present in all types of management field research, but the timing and intensity of the iterations depends on the maturity level of the research. They argue that field research is exposed to so many unforeseen events that it must be viewed as a continuous learning process. The aim with the learning process is to achieve the best possible methodological fit between prior work and research methodology. They provide suggestions for optimal fit and we present part of their work in Table 4. They suggest applying qualitative data for nascent research, a combination of data types (hybrid or mixed methods) for intermediate research and quantitative data for mature research. This is in line with our view of type of data being orthogonal to the choice of fixed and flexible design. However, we believe that quantitative data is sometimes useful for nascent research and qualitative data may contribute fruitfully to mature research.

A fourth factor to consider is the time schedule of the research. Studies with a short time schedule often require a fixed design, whereas studies that have a long perspective often require a flexible design. For example, an experiment performed during a one hour lecture has a short time schedule, whereas experiments performed during a whole day or longer have a long time schedule. Sometimes, participants in experiments are performing tasks at different points in time. Such experiments might last over several weeks, allowing the researcher to influence the later part of the experiment based on experiences obtained in the early part. Also, the chances for other unexpected events increase with time.

TABLE 4: Categories of methodological fit for research in a field setting (Edmondson and McManus 2007)

State of prior theory and research	**Nascent**	**Intermediate**	**Mature**
Research questions	Open-ended inquiry about a phenomenon of interest.	Proposed relationships between new and established constructs.	Focused questions and/or hypotheses relating existing constructs.
Constructs and measures	Typically new constructs, few formal measures	Typically one or more new constructs and/or new measures	Typically relying heavily on existing constructs and measures
Type of data	Qualitative, initially open-ended data that need to be interpreted for meaning.	Hybrid (both qualitative and quantitative).	Quantitative data: focused measures where extent or amount is meaningful.
Research schedule			

4. ESTABLISHING TRUSTWORTHINESS

An important part of the research design is to establish trustworthiness. In a fixed research design, the trustworthiness is established by a convincing argument for the importance of the research, a corresponding sensible plan, which includes control with potential biases that can influence the result, and a performance according to the plan. Central concepts, when talking about trustworthiness in fixed designs, are *validity* and *reliability*, see for example descriptions in (Shadish, Cook et al. 2002) and in (Wohlin, Runeson et al. 1999). Examples of ways of establishing trustworthiness in fixed designs are randomization, blinding, random sampling, and computations of researcher's agreement scores.

Also in flexible designs, trustworthiness is established by convincing arguments regarding the importance of the research, but there is no fixed plan up front to compare performance to by the end of the study, and there might be different types of biases than in fixed designs.

In the remaining of this section, we describe techniques for establishing trustworthiness in flexible designs. We will use the definitions of validity and reliability that is suited to all types of research, suggested in (Hinds, Scandrett-Hibden et al. 1990). We start with describing validity.

> Validity is established when the findings reflect reality, and the meaning of the data is accurately interpreted.
>
> (Hinds, Scandrett-Hibden et al. 1990, p.431)

One main threat to validity in studies with a flexible design comes from the researcher's involvement in the study. It is the researcher's role to be deeply involved in every iteration and decision in the study. In contrast to using a fixed design, where the researcher can concentrate on the planning in a specific time period followed by phases of practical work and analyses according to the plan, in a flexible design, he or she must continuously handle all aspects of the research: planning, performance and analyses. This is a very demanding situation. The researcher must avoid that the research is more influenced by his or hers personal assumptions than by the data.

This threat to researcher bias and valid interpretation can be reduced or eliminated by the techniques described in the literature of qualitative research, see for example (Kvale 1989; Huberman and Miles 2002; Creswell 2007).

In addition to the potential researcher bias, we believe there are two other main threats to validity in flexible designs. One is collecting the data that is not best suited for answering the research questions. This might occur when the research question changes in response to the research and the data collection procedure is not sufficient flexible to account for these changes. This threat can be reduced by collecting more data than is necessary to answer the initial research questions. A second threat to validity occurs when the researcher does not account for the design flexibility when analysing and reporting the results. The flexibility in the design will influence the way inferences can be made from the results. For example, the assumptions for the statistical analyses might not be fulfilled. In such cases, the results can be regarded as hypothesis generating, instead of conclusive. Furthermore, the reporting of the study must account for the insight obtained through the flexible approach. Hence, both the limitations and the gains obtained through the flexibility must be reported.

We suggest considering these threats to validity and corresponding techniques to reducing them, when performing studies in ESE that needs flexibility in the design. In particular, we are concerned with those studies that traditionally do not use such techniques, for example experiments, systematic reviews, and other studies that use quantitative data. We recommend the following, which are mostly based on (Robson 2002):

- **Strive for the right researcher skill**. The researcher must be flexible, manage unanticipated directions in the research - balance adaptiveness and rigour. Moreover, the researcher must know the issue under investigation, because the information gathered is interpreted, not only recorded. And finally, he must be open to contrary findings and ask for critical views on the work.
- **Use multiple researchers.** There is probably more need for multiple researchers in the conduct and analyses in flexible designs than in fixed designs. Arrange *peer debriefing* and *support groups sessions*.
- **Be aware your value system.** Write a description of your pre-assumptions and value-system and continuous do reflective journal-keeping.
- **Document everything.** Produce an archive of your activities, raw data, analysis notes, etc. and let others inspect it (*Audit trial).* Also, document the analysis process to be able to trace the route by which you came to your interpretation.
- **Use the strategy of *triangulation***. Use of multiple sources to enhance the rigour of the research. For example, collect both qualitative and quantitative data.
- **Collect data on a broad basis**. Be open for the need for data that are related to, but that not directly contribute to answering the initial research questions.
- **Perform *member checking.*** Check with the respondents whether your interpretations are correct from their view. For example, interview the participants in experiments after their performance.
- **Account for the flexibility in the analysis and reporting of the study.** Both the limitations and the gains obtained through the flexibility must be considered in the analysis and reporting.

Generalizability is one part of validity. Generalizability is possible in flexible design, by providing sufficient information in the reporting of the study, enabling the reader to determine whether the findings are applicable to a new situation (Robson 2002).

Reliability is the second concept of trustworthiness.

> Reliability is established when the repeatability of scientific observations, and sources that could influence the stability and consistency of those observations, have been identified and evaluated.
>
> (Hinds, Scandrett-Hibden et al. 1990, p.431)

Subjectivity and objectivity in research are often connected to the question of reliability. The researcher's role in the flexible design makes it easy to consider flexible design to be subjective, and thereby unreliable. Patton (1990) prefers to avoid using the words subjectivity and objectivity. He strives for *"emphatic neutrality"* and with that, he means to be *non-judgemental* and report what is found in a *balanced* way. Phillips (1990) claims that "All good research is objective in the sense that it has been open to criticism and withstood serious scrutiny. Hence, a way of establishing reliability in flexible designs is to let other researchers evaluate all aspects of the research.

Above, we have presented ways of establishing trustworthiness in the research to handle the challenges that arise from the flexibility of the design. In addition, there are considerations to make regarding worldviews and particular choices of research methods and type of data gathered. For example, Lee (1989) discuss conducting case studies consistent with the conventions of positivism, Klein and Myers (1999) discuss how to conduct interpretive field studies, Host and Runeson (2007) have suggested a checklist to use in case studies in software engineering, see also the book by Yin (2003) for general descriptions of case studies. Moreover, the recent special issue of *Information and Software Technology* on qualitative software engineering research provides many useful examples of approaches for study designs, data collection, and analysis that should be relevant for future studies of software development employing flexible designs (Dittrich, John et al. 2007). Finally, issues regarding mixed methods are presented by Tashakkori and Teddlie (2003).

5. CONCLUSION

This paper suggests a new perspective on research designs in empirical software engineering – the flexibility in the designs. The rationale for this perspective is that studies in empirical software engineering are often exploratory, immature or performed in a field setting. Moreover, the studies involve people, for whom we cannot predict exact behaviour or skill. Because such studies are difficult to pre-plan in detail, the researcher must be flexible and prepared to manage unanticipated directions of the research. This requires the use of flexible research designs.

Our impression is that most research in ESE use fixed designs, in the form of experiments and surveys, probably because this type of design is traditionally regarded as most reliable, or most easy to perform. This strategy might imply that the full potential of the study is not achieved, for example, deviations from the plan are regarded as threats to validity. Using a flexible design, such deviations are regarded as learning opportunities and used to adjust the remaining of the research as well as being part of the results. Moreover, flexible research requires a flexible budget. Hence, planning for flexibility will help realistic budgeting.

A flexible design can be used in all types of empirical research in software engineering, the extent and timing of the flexibility being study specific. In order to establish trustworthiness, techniques for reducing researcher bias must be used and the reporting of the study must account for both the limitations and the insight obtained through the flexible approach.

Our aim is to initiate a discussion on how to handle the need for flexibility in research designs in ESE and simultaneously perform methodologically sound studies.

ACKNOWLEDGEMENT

Thanks to Barbara Kitchenham, Per Runeson and Margunn Aanestad for introducing us to the topic and for interesting discussions.

REFERENCES

Anastas, J. W. and MacDonald, M. L. (1994) *Research Design for Social Work and the Human Services*. New York: Lexington.

Anda, B. C. D., Hansen, K., Gullesen, I. and Thorsen, H. K. (2006) Experiences from using a UML-based development method in a large safety-critical project. *Empirical Software Engineering,* **11,** 555-581.

Anda, B. C. D. and Sjøberg, D. I. K. (2003) Applying use cases to design versus validate class diagrams - a controlled experiment using a professional modelling tool. *Second International Symposium on Empirical Software Engineering (ISESE),* Rome, Italy, 30 September-1 October, pp.~ 50-60, IEEE Computer Society.

Anda, B. C. D. and Sjøberg, D. I. K. (2005) Investigating the role of use cases in the construction of class diagrams. *Empirical Software Engineering,* **10,** 285-309.

Arisholm, E., Gallis, H., Dybå, T. and Sjøberg, D. I. K. (2007) Evaluating pair programming with respect to system complexity and programmers expertise. *IEEE Transactions on Software Engineering,* **33,** 65-86.

Brereton, P., Kitchenham, B.A., Budgen, D., Turner, M. and Khalil, M. (2007) Lessons from applying the systematic literature review process within the software engineering domain, *The Journal of Systems & Software* **80**, 571-583.

Conradi, R., Dybå, T., Sjoberg, D. and Ulsund, T., Eds. (2006) Software Process Improvement: Results and Experience from the Field. Springer.

Creswell, J. W. (2007). *Qualitative Inquiry & Research Design. Choosing Among Five Approaches*. Sage.

DeLeeuw, A. C. J. and Volberda, H. W. (1996) On the concept of flexibility: a dual control perspective. *Omega, International Journal of Management Science,* **24,** 121-139.

Dittrich, Y., John, M., Singer, J. and Tessem, B. (2007) For the special issue on qualitative software engineering research. *Information and Software Technology,* **49,** 531-539.

Dybå, T. (2005) An empirical investigation of the key factors for success in software process improvement. *Transactions on Software Engineering,* **31,** 410-424.

Edmondson, A. C. and McManus, S. E. (2007) Methodological fit in management field research. *Academy of Management Review,* **32,** 1155-1179.

Golden, W. and Powell, P. (2000) Towards a definition of flexibility: in search of the Holy Grail? *Omega, The International Journal of Management Science,* **28,** 373-384.

Grimstad, S. and Jørgensen, M. (2007) The impact of irrelevant information on estimation of software development effort. *The Australian Software Engineering Conference,* Melbourne, pp.~359-368, IEEE Computer Society.

Hannay, J. E., Sjøberg, D. I. K. and Dybå, T. (2007) A systematic review of theory use in software engineering experiments. *IEEE Transactions on Software Engineering,* **33,** 87-107.
Hinds, P. S., Scandrett-Hibden, S. and McAulay, L. S. (1990) Furter assessments of a method to estimate reliabilty and validity of qualitative research findings. *Journal of Advanced Nursing,* **15,** 430-435.
Host, M. and Runeson, P. (2007) Checklists for software engineering case study research. *First International Symposium on Empirical Software Engineering and Measurement (ESEM),* 479-481.
Huberman, A. M. and Miles, M. B. (2002). *The Qualitative Researcher's Companion*. Sage.
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,** 1073-1086.
Kampenes, V. B., Dybå, T., Hannay, J. E. and Sjøberg, D. I. K. (2008) A systematic review of quasi-experiments in software engineering, To appear in *Information & Software Technology,*
Kitchenham, B., Mendes, E. and Travassos, G. H. (2006) A systematic review of cross-vs. within-company cost estimation studies. *Proceedings for the 10th International Conference on Evaluation and Assessment in Software Engineering (EASE), Keele University, UK, 10-11 April,* pp.~89-98, British Computer Society.
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,** 67-94.
Kvale, S., Ed. (1989) Issues of Validity in Qualitative Research. Chartwell Bratt Ltd.
Laitenberger, O., Emam, K. E. and Harbich, T. G. (2001) An internally replicated quasi-experimental comparison of checklist and perspective- based reading of code documents. *IEEE Transactions on Software Engineering,* **27,** 387-421.
Lee, A. S. (1989) A scientific methodology for MIS case studies. *MIS Quarterly,* **12,** 33-50.
Neuman, W. L. (2006) *Social Research Methods: Quantitative and Qualitative Approaches*. Allyn & Bacon.
Olsson, N. O. E. (2006) Management of flexibility in projects. *International Journal of Project Management,* **24,** 66-74.
Patton, M. Q. (1990) *Qualitative Evaluation and Research Methods.* Sage Publications, Newbury Park, CA.
Phillips, D. C. (1990) Subjectivity and objectivity: an objective inquiry. In Eisner and Peshkin (eds.), *Qualitative Inquiry in Education: The Continuing Debate*. Teachers College Press, New York.
Robson, C. (2002). *Real World Research*. Blackwell Publishers Inc.
Segal, J., Grinyer, A. and Sharp, H. (2005) The type of evidence produced by empirical software engineers, *Proceedings of the Workshop on Realising Evidence-Based Software Engineering*, St. Louis, Missouri, USA, May 17, pp.~1-4, ACM.
Shadish, W. R., Cook, T. D. and Campbell, D. T. (2002) *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin, Boston.
Sjøberg, D.I.K., Hannay, J.E., Hansen, O., Kampenes, V.B., Karahasanovic, A., Liborg, N-K., and Rekdal, A.C. (2005) A survey of controlled experiments in software engineering. *Transactions on Software Engineering,* **31***,* 733-753.
Syversen, E., Anda, B. and D.I.K.Sjøberg (2003) An evaluation of applying use cases to construct design versus validate design. *Hawaii International Conference on System Sciences (HICSS-36),* Big Island, Hawaii, 6-9 January, IEEE Computer Society.
Tashakkori, A. and Teddlie, C., Eds. (2003) Handbook of Mixed Methods in Social & Behavioral Research. USA, Sage.
Wohlin, C., Runeson, P., Høst, M., Ohlsson, M. C., Regnell, B. and Wesslen, A. (1999) *Experimentation in Software Engineering: An Introduction*. Kluwer Academic Publishers.
Yin, R. K. (2003). *Case Study Research : Design and Methods*. Sage, Thousand Oaks, CA.