

C. Wohlin, M. Höst and A. Wesslén, "Can the Personal Software Process be used for Empirical Studies?", Proceedings ICSE workshop on Empirical Studies of Software Development and Evolution", Los Angeles, USA, 1999.

Can the Personal Software Process¹ be used for Empirical Studies?

Claes Wohlin, Martin Höst and Anders Wesslén
Dept. of Communication Systems, Lund University
PO Box 118, SE-221 00 Lund, Sweden
E-mail: (claesw, martin, wesslen)@tts.lth.se

Abstract

This paper discusses the use of the Personal Software Process (PSP) as a context for conducting empirical studies. It is argued that the PSP provides an interesting environment for doing empirical studies. Student experiments are frequently used to perform empirical studies to improve the understanding of software development as an engineering discipline. The PSP may be one context in which student experiments can be performed, in particular if the PSP is taught independently of the objective of doing empirical studies. In other words, the empirical study becomes a natural by-product in teaching the PSP. This has some benefits in comparison with other types of studies. The PSP provides a well-defined context and eases replication of the studies. The objective of this paper is to present the idea and discuss the opportunities in combining the PSP with empirical studies.

1. Introduction

The PSP can be viewed from two different perspectives when it comes to empirical studies. First, it is important to evaluate the effect of the PSP. This means that the PSP is the object of the study. We have seen results published [Humphrey96, Ferguson97], but further studies are needed. This includes both reports on the outcome from taking the PSP as a course and from industrial use of the PSP. It is, however, not the intention here to discuss this matter. Second, assuming that we have started to teach and use the PSP, it can be used as a context for empirical studies and hence as a vehicle for evaluating different methods and techniques, and to study different relationships in software engineering.

2. Advantages

A key question to consider is, of course, why should we use the PSP as a context for empirical studies? Or we could negate the question, and ask why should we not use the PSP as a context for empirical studies? We would like to argue, since we often have to resort to or see the value in using students in empirical studies, that it is suitable to use the PSP. It provides a context which includes collection of several measures which are valuable when performing empirical studies. Moreover, we also believe that the PSP can be the starting point for case studies, i.e. before going out in industry and perform a major study, we can gain a first and valuable insight by performing studies using the PSP

1. PSP and the Personal Software Process are service marks of Carnegie-Mellon University.

and its outcome. Thus, the main advantages of using the PSP as a context for empirical studies are:

- Defined context.

The context is given by the definition of the PSP as described by Watts Humphrey [Humphrey95]. We may want to change the proposed PSP slightly, but basically the context is provided, and hence we do not have to define the context and describe it very carefully to allow for others to understand our study from the context perspective.

- Replication.

The context also forms the basis for replication. A major problem in empirical studies is that in order to come up with generally valid observations, we must be able to perform a study several times to build up general experience. Thus, the PSP may be one way to ease replication. In other words, experiments and case studies can be conducted at several places using the PSP simultaneously. The PSP provides a well-defined process and the process description is generally available.

- Measures.

This is also closely related to the PSP. Measures are collected as an integrated part of the PSP, and it is fairly easy to add measures of specific interest for an empirical study. Thus, the PSP provides a good starting point for collecting measures to use for hypothesis testing or model building.

Although, several advantages have been identified, it is clear that the PSP is not the silver bullet for empirical studies. The PSP has some disadvantages, both in terms of limitations on the type of studies that can be conducted and regarding the validity of the results obtained through using the PSP as a context for empirical studies. For a specific study, it can probably be argued that a better context exists. The major advantage with the PSP is that it is well-documented and generally available. This is the key point, and it certainly increases the replication opportunities. A replication could either be a replication in terms of performing a similar study, although the results come out differently, or a true replication in terms of also getting the same results. In the first case, it is important to address the reasons for the differences. The validity problems using the PSP are discussed in the next section.

3. Validity

The major challenge in using the PSP as context is the ability to scale the observation to other environments, and in particular to large-scale software development. On the one hand, it is difficult to scale individual results to large project, on the other hand the PSP is supposed to act as a down-scaled project.

The validity of the observations and findings is crucial. We would like to be able to generalize the observations. In order to do this, we should consider different types of validity or threats to the empirical studies. Four types of validity may be addressed [Cook79]:

1. Conclusion validity.

This validity is concerned with the relationship between the treatment and the outcome. We want to make sure that there is a statistical relationship, i.e. with a given significance.

2. Internal validity.

If a relationship is observed between the treatment and the outcome, we must make sure that it is a causal relationship, and that it is not a result of a factor of which we have no control or have not measured. In other words, we should be certain that the treatment causes the outcome (the effect).

3. Construct validity.

This validity is concerned with the relation between theory and observation. If the relationship between cause and effect is causal, we must ensure two things: 1) that the treatment reflected the construct of the cause well and 2) that the outcome reflected the construct of the effect well.

4. External validity.

The external validity is concerned with generalization. If there is a causal relationship between the construct of the cause, and the effect, can the result of the study be generalized outside the scope of our study?

The interpretation of the different types of threats is illustrated in Figure 1.

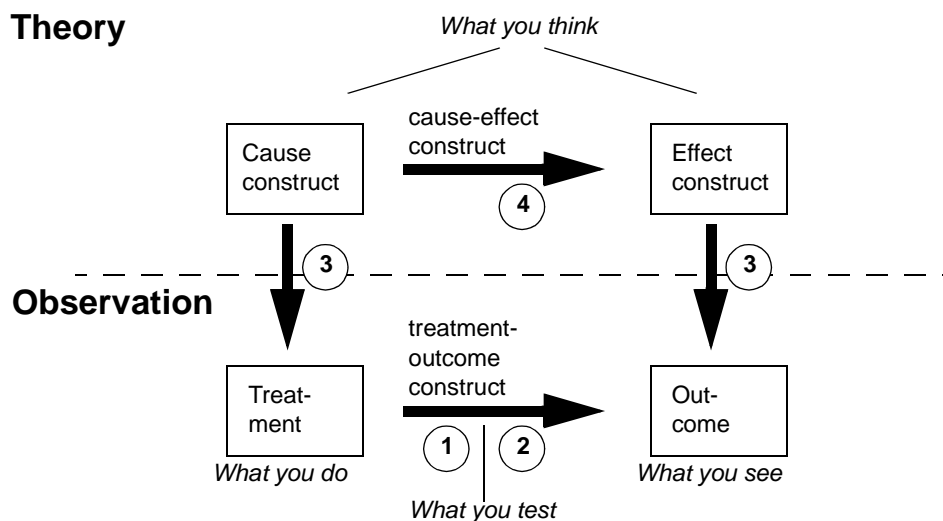


FIGURE 1. Validity threats to be considered [Wohlin99].

The threats can be addressed partly when considering the PSP as a context for empirical studies in general, but the actual validity of a specific study should be addressed separately, as the validity is highly dependent on the study and what we intend to generalize. The threats addressed here are the threats in which empirical studies within the PSP are better or worse than using students in empirical studies in general. The threats differentiating the PSP as a context from other empirical studies in a student setting are:

1. Conclusion validity

The threats regarding conclusion validity are less than when using student subjects in general due to the fact that the context is well-defined and generally available, for example, process scripts and measures are largely predefined. It is, however, worth noting that the data validity of the PSP has been questioned lately [Disney98]. It has been reported that approximately 5% of the data collected is erroneous. Unfortunately, this problem probably exists in most empirical studies where a large amount of data is recorded. Thus, it is not a strong threat to the validity, although it is a problem.

The type of error in the data should also be taken into consideration. We can identify three different types, which affect the outcome quite differently, namely random, omissions and intentional. The two first types are not that critical, since random errors should not affect the outcome in a systematic way. If some data is omitted, the persons responsible for these omissions can be removed from the study. The intentional errors are more difficult, but they can probably be found in any empirical investigation, where participants have the opportunity to make the data look better than it normally should.

2. Internal validity

The internal validity is also better in the PSP since it is based on performing several assignments (normally 10 development assignments) over a period of time, hence minimizing, for example, the risk that rare events disturb the study. The length of the experiment also influences other aspects, such as maturation and mortality. The former includes one negative aspect, i.e. the subjects get bored and loose interest, and one positive aspect, i.e. learning effects can be removed by not analysing all of the assignments. The latter is concerned with people dropping out of the study, which may be a risk due to the length of the study, but on the other hand it is structured as a course and people want to learn the whole package. Other internal threats include instrumentation issues, and in this case the PSP is strong due to the well-defined context in terms of process scripts and different templates. Finally, a social threat to the internal validity is that people perform differently since they know that they are part of an experiment. We believe this threat is smaller in the PSP, since people primarily focus on the course as such and not on participating in an experiment.

3. Construct validity

The construct validity is threatened by the fact that it is an experiment and there may be interactions between the testing and the treatment, and people may start guessing the hypothesis to be tested and hence intentionally or unintentionally influencing the construct validity. Since the PSP is normally run primarily as a course and not as an empirical study, we believe that this threat is smaller in the PSP context than when setting up a separate empirical study. The fact that it is a course is, however, also a threat since the students will be graded. Thus, they may behave differently due to the grading. In addition to this, it must be ensured that the measures collected during the empirical study reflect what they are supposed to. In other words, that the measures within the PSP, or any measure added to the PSP, represent a true picture of the constructs, i.e. both cause and effect constructs. The measures defined should represent the cause construct and the collected measures should reflect the effect construct. In summary, we believe that the advantages and disadvantages regarding construct validity outweigh each other for the PSP and student studies in general.

4. External validity

The external validity is normally very important, since we would like to generalize our results. The use of the PSP as a context in comparison with other student experiments is similar in terms of threats. An obvious threat is that the PSP is concerned with developing small programs, and mostly in the domain of statistical programs. Other student experiments are also normally concerned with small assignments, hence the difference is not that big. An advantage for the PSP is that the objective is to scale down large-scale software development to the individual level hence systemati-

cally practising important activities for large-scale software development. Thus, the external validity is believed to be the same for using the PSP as a context and using students as subjects in general.

In summary, it is our opinion that if we teach the PSP anyway, the threats to the validity is smaller when using the PSP as context for empirical studies than when using students as subjects in general.

4. Summary

The PSP provides opportunities for empirical studies. We may study the use of different techniques and methods, or investigate the relationships between different attributes. The main limitation of using the PSP as a basis for empirical studies is that we have limited opportunities to study group activities. It is possible to experiment with using different reading techniques on an individual basis, but we are unable to study the use of inspections and group meetings within the PSP. Thus, we cannot expect to use the PSP as a context for experimentation for all types of empirical studies, but it is our firm belief that it opens some new opportunities and that the major inhibiting factor is our imagination.

Acknowledgment

This work was partly funded by The Swedish National Board for Industrial and Technical Development (NUTEK), grant 1K1P-97-09673.

References

- [Cook79] T. D. Cook and D. T. Campbell, "Quasi-Experimentation – Design and Analysis Issues for Field Settings", Houghton Mifflin Company, 1979.
- [Disney98] A. M. Disney and P. M. Johnson, "Investigating Data Quality Problems in the PSP", Proceedings of the Sixth International Symposium on the Foundations of Software Engineering, 1998.
- [Ferguson97] P. Ferguson, W. Humphrey, S. Khajenoori, S. Macke and A. Matvya, "Results of Applying the Personal Software Process", IEEE Computer, Vol. 30, No. 5, pp. 24-31, 1997.
- [Humphrey95] W. S. Humphrey, "A Discipline of Software Engineering", Addison-Wesley, 1995.
- [Humphrey96] W. S. Humphrey, "Using a Defined and Measured Personal Software Process", IEEE Software, pp. 77-88, May 1996.
- [Wohlin99] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell and A. Wesslén, "Introduction to Experimentation in Software Engineering", Technical report Department of Communication Systems, Lund University, Sweden, 1999. (In preparation for publication).