There appears to be no easy answer to this question. Each design will be a result of a number of tradeoffs, and it is not always possible to know how the decisions will influence the data. A good design can have various interpretations based on what are considered the goals for the experiment. One option is to use different designs involving different threats to validity and study the results as a whole.

- *What is the optimal sample size? Small samples lead to problems in the statistical analysis while large samples represent major expenses for the organization providing the subjects.*

  Organizations generally have limits for the amount of subjects they are willing to part with for an experiment, so the cost concerns are handled by the organizations themselves. A small sample size requires us to be careful in the design in order to get as many useful data points as possible. For this experiment, an example of such a tradeoff is that we chose to neglect learning effects in order to avoid spending subjects on control groups. This gave us more data points to be used in analyzing the difference between the two techniques, but at the same time we remained uncertain as far as the threat to internal validity caused by learning effects is concerned.

- *We need to adjust to various constraints - how far can we go before the value of the experiment decreases to a level where it is not worthwhile?*

  Our problem as experimenters is to maintain a certain level of validity while still generating sufficient interest for an organization to allow us to conduct the experiment. From an organization's point of view, an experiment should be closely tied to their own environment to see if the suggested improvement works with minimal effort in terms of environmental changes. From an experimental point of view, however, we are interested in a controlled environment where disturbing interaction effects are negligible.

- *To what extent can experimental aspects such as design, instrumentation and environment be changed when the experiment still is to be considered a replication?*

  One requirement for being considered a replication is that the main hypotheses are the same. Changes in design and instrumentation, in particular to overcome threats to validity, should also be considered "legal". However, one situation we should avoid is making substantial changes to the design based on the *results* from a previous experiment. This will introduce dependencies between the experiments that are highly undesirable from a statistical point of view.

For this experiment in particular, there are various problems that we need to study more carefully. The threats to validity should be carefully examined; in particular we feel the testing effects to be crucial. An experiment with a control group could be one way of estimating what the importance of these effects really are. We may also consider a more careful analysis of the NASA documents and environment in order to refine PBR to these particular needs. The results indicate that the choice of perspectives and associated scenarios do not match the needs of the NASA domain.

A more fundamental problem that should be considered is to what extent the proposed technique actually is followed. This problem with process conformance is relevant in experiments, but also in software development where deviations from the process to be followed may lead to wrong interpretation of measures obtained. For experiments, one problem is that the mere action of controlling or measuring conformance may have an impact on how well the techniques work, thus decreasing the external validity.

Conformance is relevant in this experiment because there seems to be a difference that corresponds to experience level. Subjects with less experience seem to follow PBR more closely ("It really helps to have a perspective because it focuses my questions. I get confused trying to wear all the hats!"), while people with more experience were more likely to fall back to their usual technique ("I reverted to what I normally do.").

There are numerous alternative directions for the continuation of this research. For further experimentation within NASA's SEL it seems to be necessary to tailor PBR to more closely match the particular needs of that domain. A possible way of further experimentation would be to do a case-study of a NASA SEL project to obtain more qualitative data.

We may also consider replication of the generic part of the experiment in other environments, perhaps even in other countries where differences in language and culture may cause effects that can be interesting targets for further investigation. These replications can take the form of controlled experiments with students, controlled experiments with subjects from the industry using their usual technique for comparison, or case studies in industrial projects.

One challenging goal of a continued series of experiments will be to assess the impact that the threats to validity have. Since it is often hard to design the experiment in a way that controls for most of the threats, a possibility would be to concentrate on certain threats in each replication to assess their impact on the results. For example, one replication may use control groups to measure the effect of repeated tests, while another replication may test explicitly for maturation effects. However, we need to keep the replications under control as far as threats to *external* validity are concerned, since we need to assume that the effects we observe in one replication will also occur in the others.

Acknowledgements

This research was sponsored in part by grant NSG-5123 from NASA Goddard Space Flight Center to the University of Maryland. We would also like to thank the members of the Experimental Software Engineering Group at the University of Maryland for their valuable comments to this paper.

References

(Campbell, 1963) Campbell, Donald T. and Stanley, Julian C. 1963. *Experimental and Quasi-Experimental Designs for Research* . Boston, MA: Houghton Mifflin Company.

(Edington 1987) Edington, Eugene S. 1987. *Randomization Tests*. New York, NY: Marcel Dekker Inc.

(Fagan, 1976) Fagan, M. E. 1976. *Design and code inspections to reduce errors in program development.* IBM Systems Journal, 15(3):182-211.

(Hatcher, 1994) Hatcher, Larry and Stepanski, Edward J. 1994. *A Step-by-Step Approach to Using the SAS® System for Univariate and Multivariate Statistics*. Cary, NC: SAS Institute Inc.[2]

(Heninger, 1980) Heninger, Kathryn L. 1985 *Specifying Software Requirements for Complex Systems: New Techniques and Their Application.* IEEE Transaction on Software Engineering, SE-6(1):2-13

(Linger, 1979) Linger, R. C., Mills H. D. and Witt, B. I. 1979. *Structured Programming: Theory and Practice.* In The Systems Programming Series. Addison Wesley.

(Parnas, 1985) Parnas, Dave L. and Weiss, David M. 1985. *Active design reviews: principles and practices.* In Proceedings of the 8th International Conference on Software Engineering, p.215-222.

(Porter, 1995) Porter, Adam A., Votta, Lawrence G. Jr. and Basili, Victor R. *Comparing Detection Methods For Software Requirements Inspections: A Replicated Experiment.* IEEE Transactions on Software Engineering, June 1995.

(SAS, 1989) SAS Institute Inc. 1989. *JMP® User's Guide.* Cary, NC: SAS Institute Inc.[3]

(SEL, 1992) Software Engineering Laboratory Series. 1992. *Recommended Approach to Software Development, Revision 3*, SEL-81-305, p. 41-62.

(Votta, 1993) Votta, Lawrence G. Jr. 1993 *Does every inspection need a meeting?* In Proceedings of ACM SIGSOFT '93 Symposium on Foundations of Software Engineering. Association of Computing Machinery, December 1993.

A. Sample Requirements

Below is a sample requirement from the ATM document which tells what is expected when the bank computer gets a request from the ATM to verify an account:

Functional requirement 1

Description: The bank computer checks if the bank code is valid. A bank code is valid if the cash card was issued by the bank.

Input: Request from the ATM to verify card (Serial number and password)

Processing: Check if the cash card was issued by the bank.

Output: Valid or invalid bank code.

We also include a sample requirement from one of the NASA documents in order to give a picture of the difference in nature between the two domains. Below is the process step for calculating adjusted measurement times:

Calculate Adjusted Measurement Times: Process

1. Compute the adjusted Sun angle time from the new packet by

$$t_{s,adj} = t_s + t_{s,bias}$$

2. Compute the adjusted MTA measurement time from the new packet by

$$t_{T,adj} = t_T + t_{T,bias}$$

3. Compute the adjusted nadir angle time from the new packet.

a. Select the most recent Earth_in crossing time that occurs before the Earth_in crossing time of the new packet. Note that the Earth_in crossing time may be from a previous packet. Check that the times are part of the same spin period by

$$t_{e-in} - t_{e-out} < E_{\max} T_{spin,user}$$

b. If the Earth_in and Earth_out crossing times are part of the same spin period, compute the adjusted nadir angle time by

$$t_{e-adj} = \frac{t_{e-in} + t_{e-out}}{2} + t_{e,bias}$$

4. Add the new packet adjusted times, measurements, and quality flags into the first buffer position, shifting the remainder of the buffer appropriately.

5. The Nth buffer position indicates the current measurements, observation times, and quality flags, to be used in the remaining Adjust Processed Data section. If the Nth buffer does not contain all of the adjusted times ($t_{s,adj}, t_{b,adj}, t_{T,adj}$, and $t_{e,adj}$), set the corresponding time quality flags to indicate invalid data.

Footnotes

[1] ISERN is the International Software Engineering Research Network whose goal is to support experimental research and the replication of experiments.

[2] SAS® is the registered trademark of SAS Institute Inc.

[3] JMP® is a trademark of SAS Institute Inc.