

A Controlled Experiment to Evaluate On-Line Process Guidance

Christopher M. Lott*
Bellcore, Morristown NJ 07960
lott@bellcore.com

March 31, 1997

Abstract

Process-centered software engineering environments are expected to improve an individual's comprehension of work activities, as well as improve communication and reduce conflicts among teams of software developers. We chose to investigate individual responses when using such an environment before looking for a group response. A controlled experiment compared off-line and on-line implementations of measurement-based process guidance to test three hypotheses: first, individuals accomplish their work more efficiently when using on-line process guidance as compared to off-line guidance; second, individuals are willing to use an on-line system for guidance; and third, individuals adjust their behavior based on quantitative quality models. The 20 subjects worked alone on two testing exercises. Subjects used either an off-line or an on-line guidance technique during the first exercise, and the other technique during the second exercise. The results did not confirm the hypotheses. On average, subjects worked more efficiently when using off-line guidance, preferred off-line guidance, and ignored quantitative quality models. Post hoc analysis identified a strong correlation between subject experience level and preference for the type of guidance, a result that deserves further investigation.

1 Introduction

Much research has focused on developing process-centered software engineering environments (surveys include [Dart et al., 1987, Perry and Kaiser, 1991, Lott, 1994]).

*This work was conducted while the author was with the Fraunhofer Institute for Experimental Software Engineering, Kaiserslautern, Germany.

However, few studies have demonstrated measurable, objective benefits of their use. The fundamental hypothesis concerning process-centered software engineering environments is that they improve communication and reduce problems among teams who cooperate to achieve some task. This hypothesis might be evaluated both in uncontrolled and controlled studies that complement each other's strengths and weaknesses. To attain maximum external validity, a field study (uncontrolled experiment) might follow teams through large tasks. To attain maximum internal validity, a controlled experiment could follow individuals through small tasks. A field study would be quite expensive, and controlling the threats to internal validity would be difficult. A smaller, more controlled experiment will help the community explore the issues and understand the threats involved in a larger field study. The study described here evaluated how individuals interact with an on-line system that offers measurement-based process guidance.

Automated support for measurement-based process guidance offers many potential benefits. First, all personnel see the same version of their process information. There is no need to update handbooks or other off-line documentation used regularly by personnel. Second, sophisticated visualization techniques can be applied to help personnel understand their processes. Third, these systems allow personnel to investigate and learn the consequences of their actions upon other personnel, and receive early notification of actions that may affect them. Finally, the intrusiveness and effort associated with data collection can be reduced by supporting these activities on the computer. For example, a user can invoke an electronic form at the appropriate point in time according to the process, and the data thus collected can be stored directly without requiring rekeying.

Some work has been done on evaluating process guidance systems. Christie reports on a comparison of ProcessWeaver and SynerVision to accomplish a shared task [Christie, 1995]. The study evaluated issues concerning ease of use as viewed by both the process developer and the end user (e.g., development of the process description, screen layout, and response time). Issues of acceptance and adoption among the end users (e.g., intrusiveness and impersonality of on-line guidance) were also evaluated. However, neither the amount of time required for the task nor the quality of the final result were evaluated. That study raises important questions, including whether resistance to working in an automated process environment will be a major impediment to adoption, and whether process automation products can realistically model industrial-strength projects. The study described in this paper specifically addresses the question of acceptance of on-line guidance by users.

The process guidance system that was used in this study is a research prototype that supports measurement-based guidance and feedback for individuals and teams [Lott et al., 1995]. The system uses integrated process and measurement information to assist software developers, and is named the "Multi-View Process

System” (MVP-S). Fundamentally, the system only provides information and collects data. MVP-S does not proscribe any actions on the computer, it does not provide any construction tools, nor does it store any of the documents that personnel may develop. The system informs users about the processes that they are expected to perform, displays information about the inputs, outputs, and personnel resources associated with each process, collects data about various aspects, evaluates those data, and provides feedback about the evaluation results. The MVP-S system architecture uses the client-server model, and supports distributed work across a local-area network. A process engine acts as the server for all information. The only client currently implemented is a user interface that provides personnel who perform technical, constructive activities with information maintained by the server.

This paper describes a controlled experiment that compared the use of measurement-based process guidance as implemented off-line (on paper) versus on-line (using MVP-S) [Lott, 1996]. Because relatively little empirical work has been done on using process-centered systems to assist software developers, the research presented here is best classified as exploratory, and can be expected to uncover more questions than it answers. Section 2 sets the context of this work using a simple characterization scheme for software engineering experiments. Section 3 states hypotheses, Section 4 gives the experiment plan, Section 5 discusses the procedures used to conduct the experiment, and finally Section 6 presents results and interpretations.

2 Context of the Experiment

Many factors can be explored in an experiment that evaluates how a person or team can be assisted with a software engineering task. Table 1 lists some of these factors [Basili, 1994]. The factors listed in the table include methods (e.g., structured programming), empirical models (e.g., defect profiles), goals (e.g., “write maintainable code”), process definitions (e.g., a process view based on a formal process modeling language), tracking and evaluation (e.g., collecting and evaluating data from process), and replanning (e.g., information about how to recover from some situation). Some of the individual factors can benefit from automated support.

A collection of factors up to and including level n might be called an approach. Level 0 (“ad hoc”) means that no supportive methods or other information are used. Thereafter, each level n adds a new factor to the approach characterized by level $n - 1$. The experiment presented here can be characterized using Table 1 as comparing the approaches at levels 4 and 5. All subjects used methods (testing techniques), empirical models (expectations of time and failure counts), goals (“work efficiently”), and process definitions. The fundamental variable of interest

Level 0	Ad hoc
Level 1	Methods
Level 2	Empirical models
Level 3	Goals
Level 4	Process definitions
Level 5	Process definitions with automated support
Level 6	Tracking and evaluation
Level 7	Replanning with automated support

Table 1: Factors tested in software engineering experiments

was the use of off-line and on-line process guidance (without automation versus with automation).

The challenge of this experiment is the choice of procedures rather than choice of design. The procedures must lead each subject to use the on-line process guidance system in a credible, natural way. This consideration heavily influences decisions about the information supplied to the subjects as well as the pretesting and training activities.

3 Hypotheses

The experimental hypotheses ask questions about how an individual interacts with off-line and on-line process definitions during a carefully planned exercise of short duration. The hypotheses have a strong exploratory character, and testing them is expected to contribute towards understanding the issues of coordinating teams using process-centered software engineering environments.

Hypothesis 1: Subjects who use an on-line guidance system accomplish their work more efficiently as compared to subjects who use off-line guidance.

Hypothesis 2: Subjects accept guidance from an on-line system as willingly as they accept conventional instructions on paper.

Hypothesis 3: Subjects comprehend on-line versions of process guidance better than they comprehend conventional, off-line guidance.

Hypothesis 4: Subjects take quantitative quality models seriously (e.g., a suggested maximum amount of time for a step), and adjust their behavior accordingly.

Hypothesis 5: Subjects supply data to a process-centered software engineering environment about starting and ending processes, as well as other data about those processes, on a timely basis.

4 Experiment Plan

This section discusses the design, instrumentation, and other materials for the experiment.

4.1 Experimental design

A randomized, within-subjects design was used in the experiment. The within-subjects design (i.e., blocked on subjects) permits the detection of differences between levels of independent variables despite wide variations in subject ability.

4.1.1 Independent variables

The following four independent variables are manipulated in the 2^4 design.

1. Exercise (two levels: FT/cmdline and ST/tokens). All subjects performed two exercises. In exercise 1, subjects applied principles of functional testing to program 'cmdline.' In exercise 2, subjects applied principles of structural testing to program 'tokens.' The test technique is therefore confounded with the program, so this experiment does not offer a fair comparison of test techniques.
2. Guidance technique (two levels: off-line and on-line). This was the primary variable of interest. Subjects were assisted with their exercises once by the off-line guidance technique and once by the on-line guidance technique.
3. Order of trials with respect to guidance (two levels: off/on and on/off).
4. Order of trials with respect to exercise (two levels: FT/ST and ST/FT).

The independent variables for order are confounded for each subject, so if a significant result is found for order, it will be impossible to separate one from the other. Uncontrolled independent variables include the subject's experience level, motivation for the experiment, and process conformance.

For each subject, the design randomized the match between exercise and guidance technique, as well as the orderings. To do so, subjects were randomly assigned to one of the four groups S_{1A} , S_{1B} , S_{2A} and S_{2B} . For example, a subject

in group S_{1A} would perform exercise 1 using the off-line guidance technique and subsequently perform exercise 2 using the on-line guidance technique. The experimental design is summarized in Table 2.

Exercise	Guidance technique		Ordering of trials	
	Off-line	On-line	Off-line/On-line	On-line/Off-line
1: FT of 'cmdline'	S_1	S_2	S_A	S_B
2: ST of 'tokens'	S_2	S_1		

Table 2: Experimental design

4.1.2 Dependent variables

A number of dependent variables were defined to measure the effects of manipulating the independent variables. Traceability between the hypotheses and the dependent variables (measures) is shown in the following list.

H. 1: Efficiency (off-line versus on-line)

- M. 1: Amount of time the subject needed to complete the exercise, excluding breaks.
- M. 2: Percentage of possible unique failures that were detected by the subject. This is derived from counts of failures possible, failures revealed by the subject's test cases, and failures recorded by the subject.

H. 2: Acceptance of on-line guidance

- M. 3: Subject's willingness to use an on-line process-centered guidance system, recorded as "prefer on-line" or "prefer off-line guidance."
- M. 4: Subject's opinion about the restrictiveness of the guidance, recorded as "too restrictive" or "not restrictive."
- M. 5: Subject's preference about entering data on paper versus on screen.

H. 3: Comprehension (off-line versus on-line)

- M. 6: Subject's opinion of the comprehensibility of the guidance offered to them, recorded as "comprehensible" or "incomprehensible."

- M. 7: Subject's opinion about the sufficiency of supplied information, recorded as "sufficient" or "insufficient."
- M. 8: Observer's opinion of a subject's understanding of a process, formed by asking the subject to list the inputs, outputs, and exit criteria of a given process while performing that process. Recorded as "good" or "fair" conformance.

H. 4: Use of quantitative quality models

- M. 9: Subject's opinion about the quantitative quality models used as start and termination criteria, recorded as "helpful" or "harmful."
- M. 10: Observer's opinion of whether the subject adjusted his or her behavior to conform to the quantitative quality models, recorded as "yes, behavior was adjusted" or "no, it was not."

H. 5: Supplying data in a timely manner

- M. 11: Observer's opinion of how long a subject waited to report results to the process-centered system or record on paper, recorded as "no waiting," "short waits," or "long waits." This is additionally supported during the on-line guidance scenario by time stamps on messages.

U: Uncontrolled independent variables

The following measures were defined to assess the uncontrolled independent variables.

- M. 12: Subject's experience level, recorded as M.S. or Ph.D. candidacy.
- M. 13: Subject's subjective assessment of motivation for the exercise, recorded on a 5-point scale.
- M. 14: Observer's opinion of the subject's process conformance, recorded as "followed the defined process faithfully" or "significant deviations from the process."

4.1.3 Threats to validity

Threats to external validity include the short duration of the exercise, and both the uniqueness and prototype nature of the process guidance system. Threats to internal validity include selection and maturation effects [Campbell and Stanley, 1966]. A selection effect could be caused by using subjects who are predisposed towards (or against) the use of a process-centered system. However, no preliminary test was administered to the subjects that might have assessed this bias in order to avoid

problems of reactivity and sensitization. Selection effects can be measured to some extent because all subjects are observed twice. However, the drawback of multiple observations is a maturation effect that might result from a subject learning from the first exercise. Different exercises were used to mitigate such maturation effects.

4.1.4 Use of quantitative quality models

Quantitative quality models are a basic part of measurement-based process guidance [Lott and Rombach, 1993]. All subjects received quantitative quality models, regardless of the guidance technique used. This was intended to approximate the situation in which an organization maintains a database about prior test efforts. For example, a quality model about a test process could predict the number of failures based on the component size or complexity.

4.2 Exercises

The testing exercises consist of generating test cases, executing them, and evaluating the output to detect failures. The target duration of the exercises was 60–70 minutes.

4.2.1 Exercise 1: Functional testing of “cmdline”

In this exercise, principles of functional (black-box) testing are applied to program “cmdline.” The subjects never see the source code. In step 1 the subjects receive an instruction sheet and prepare for the exercise by fetching the needed computer files. In step 2 the subjects receive the specification. They identify equivalence classes in the input data and construct test cases using the equivalence classes, paying special attention to boundary values. In step 3 the subjects type in their test cases using a format required by a simple test harness. In step 4 the subjects use the test harness to execute their test cases, and fix any typographical or similar mistakes in their test cases. They are instructed *not* to construct additional test cases. In step 5, the subjects use the specification to detect failures that were revealed in their output.

4.2.2 Exercise 2: Structural testing of “tokens”

In this exercise, principles of structural (clear-box) testing are applied to program “tokens.” In step 1 the subjects receive an instruction sheet and prepare for the exercise by fetching the needed computer files. In step 2 the subjects use the specification and the source code to construct test cases that will achieve 100% coverage of all branches, multiple conditions, loops, and relational operators. For example, 100% coverage of a multiple condition using a single “logical and” operator means

that all four combinations of true and false must be tested, and 100% coverage of a loop means that it must be executed zero, one, and many time(s). In step 3 the subjects type in their test cases using a format required by a simple test harness. In step 4 the subjects use an instrumented version of the program to execute their test cases and view reports of attained coverage values.¹ The subjects construct additional test cases as part of step 4 until they reach 100% coverage, or believe that they cannot achieve better coverage due to various pathological cases. In step 5 the subjects use the specification to detect failures in their output.

4.3 Instruments

The instruments include the guidance information, the programs, and the quantitative quality models.

4.3.1 Guidance Information

The guidance information consisted of a set of instructions for performing the tasks of testing each program. The instructions were fairly detailed, and described the steps of fetching the needed files, generating test cases, entering the test cases in a way that permitted automatic execution, running the test cases, evaluating the results, and recording the findings. Information about each step was offered both on-line (measurement-based guidance with automation) and off-line (measurement-based guidance without automation). The steps were described using the same text in both versions; the description of each step required text that occupied about one-third of a sheet of paper.

In the off-line version, subjects could see instructions for all steps simply by spreading out the sheets. In the on-line version, the text for each step was displayed in a separate window, and subjects could open as many or as few windows as they required. The amount of screen space on the 19-inch monitors was the only limitation on the amount of information that could be seen at one glance.

Subjects used the on-line system in two ways. The primary use involved a menu that called up a window with information about an individual process step. This was the subject's source of guidance information. Another use required manipulating a menu and a pop-up window to notify MVP-S of the events "start process" or "complete process." This permitted the system to record and display each subject's progress. No other interactions were necessary to accomplish the exercise.

¹Instrumentation and reporting was supported by GCT, the Generic Coverage Tool, software developed by Brian Marick and available by anonymous ftp from host cs.uiuc.edu in directory /pub/testing/gct.files.

4.3.2 Programs

The programs are C-language artifacts that were developed for use in studies such as this one. Program “cmdline” checks a set of command line arguments for syntactical and semantical correctness. Program “tokens” sorts and counts words in its input. Static source-code measures of the programs are summarized in Table 3, and reveal a size difference in excess of 2:1. Previous experience with these exercises in the context of a controlled experiment [Kamsties and Lott, 1995] indicated that the structural testing exercise required more time than the functional testing exercise for the same program, so the size difference was necessary to keep the exercise duration as similar as possible.

Program	Total lines	Blank lines	Lines w/ comments	NB, NC lines	Semi-colons	Fault count
cmdline	299	34	4	261	122	10
tokens	127	9	1	117	72	5

Table 3: Size measures and fault counts for the programs

Faults were seeded into the programs. The two-faceted fault classification scheme defined by Basili and Perricone [Basili and Perricone, 1984] was used to select a mix of faults (type {omission, commission} and fault class {initialization, computation, control, interface, data, cosmetic}). The number of faults was chosen to attain a similar fault density in both programs. All faults caused unique, visible failures given suitable inputs.

4.3.3 Data used to derive quality models

The data supplied as quantitative quality models (e.g., estimated amount of time required) were derived from previous experimental trials that used the same techniques and instruments [Kamsties and Lott, 1995]. Table 4 summarizes the previous observations and the quality models that were supplied to all subjects. The guidance materials explained that the quantitative quality models were based on “similar” components and exercises. With respect to the quality model for failure count, in a preset experiment using testing exercises it is possible to supply true (perfect) values. However, this information is not available in the typical project situation, so using that information would constitute a threat to external validity. Instead, a value was supplied that was close.

Functional testing of “cmdline”	Observed or true values	Supplied value
Total time (min.)	Obs: 60, 80, 86, 55, 71, 50	60
Valid equiv. classes (count)	Obs: 7, 9, 17, 3, 20, 26	16
Invalid equiv. classes (count)	Obs: 2, 3, 6, 0, 5, 8	4
Test cases (count)	Obs: 22, 22, 15, 10, 26, 16	20
Unique failures (count)	True: 10	8
Structural testing of “tokens”	Observed or true values	Supplied value
Total time (min.)	Obs: 83, 80	70
Overall coverage (percent)	Obs: 87, 86	85
Test cases (count)	Obs: 5, 9	6
Unique failures (count)	True: 5	6

Table 4: Quantitative quality models supplied to subjects

4.4 Subjects

Subjects were recruited from the staff of the Research Group on Software Engineering at the University of Kaiserslautern in Kaiserslautern, Germany. The population was made up of master’s degree candidates (quarter-time appointments) and Ph.D. degree candidates (full-time appointments). Twenty subjects (6 Ph.D. and 14 M.S. candidates) completed the exercises during their regular working hours. One threat to validity was that the subjects were in part self-selected [Brooks, 1980]. All employees were experienced in window, icon, menu, pointer (WIMP) interfaces. Only volunteers who had a working knowledge of the C programming language were accepted as subjects. No prior knowledge about the testing techniques was required.

Another possible threat to validity was biased subjects. All subjects had been previously exposed to principles and ideas in empirical software engineering, either through course work or research projects. They may have been biased towards accepting the use of explicit process models and quantitative quality models.

4.5 Data collection procedures

Both off-line and on-line data collection methods were used to gather data for the measures defined in Section 4.1.2. The primary instrument was a data-collection form. Subjects who used the on-line guidance technique also used an on-line data collection form, and subjects who used off-line guidance completed a paper form. Second, the MVP-S system sent a mail message after each interaction. Time

stamps on the messages indicated whether the system was used to record completed steps on an ongoing basis, or whether the subject waited until all steps were complete before notifying the system. Third, subjects were interrupted rather informally at least once during each trial and asked to explain, as if to an uninformed person, the current process that they were performing. The principal investigator asked about input products, output products, and termination criteria in order to form an opinion about how well the subject understood that process. Fourth, the principal investigator was either in the same room with the subject(s) or close by for the entire duration of each trial, and completed a form at the conclusion of each trial. Being present throughout the trial enabled the subject to seek help immediately if needed, and also permitted indirect observation of the subject that facilitated the collection of limited data about interactions between the user and the guidance technique. Finally, all subjects completed a paper evaluation form after they had completed both exercises to report various subjective evaluations. The subjective evaluations included the subject's comfort level with using the on-line system, their understanding of the process, their preference of guidance and data-collection technique, etc. Each subject subsequently participated in an individual debriefing session, which was an opportunity for the principal investigator to validate the subject's responses and gain a deeper understanding of them.

4.6 Data analysis procedures

The experimental design yields sets of randomized, paired comparisons. The analysis for the main effects uses inferential statistics to test the pairs for significant differences [Box et al., 1978, p. 97–101]. To do so, the difference between each subject's pair of data points is computed (e.g., compute the result of using off-line guidance less the result of using on-line guidance), the average difference over all subjects is determined, and a t-test is applied to evaluate the probability that the sample mean is zero. According to [Box et al., 1978, p. 101], the t-test is an acceptable approximation of a randomization test. Since randomization tests do not assume a normal distribution in the results, and randomization was incorporated into the design, the t-test is an appropriate choice of inferential analysis approach. In accordance with generally accepted practice, we state that a significant result was detected if the probability value p is less than 0.05. In these analyses, the data for the more experienced subjects was not separated from the less experienced subjects due to the small number of observations.

5 Experiment Procedures

The procedures that were used for conducting the experiment are presented next.

5.1 Training activities

Training in the testing and guidance techniques was sharply limited, for two reasons. First, the exercise itself should be as unfamiliar as possible to the subjects. This forced the subjects to rely on the information supplied by the guidance technique. Second, we wanted to avoid issues of reactivity or sensitization of the subjects to the true goals of the experiment.

No training in the test techniques was offered. All subjects were given a 10-minute introduction to the MVP-S before they worked with it. This introduction followed a script so that each subject received approximately the same information. Part of the introduction involved having the subject try the interactions required to complete the exercise under the direction of the principal investigator.

5.2 Conducting the experiment

The design presented in Section 4.1 was implemented using a random drawing to assign each subject to one of the four groups (i.e., orderings) shown in Table 5. Subjects were assigned an identification number to preserve their anonymity. Three subjects started but failed to complete both exercises, so recruiting continued until 20 subjects had completed both exercises.

Group	Experimental trials	Subject IDs
S_{1A}	FT/cmdline off, ST/tokens on-line	2, 8, 11, 14, 17
S_{1B}	ST/tokens on, FT/cmdline off-line	3, 12, 15, 18, 19
S_{2A}	ST/tokens off, FT/cmdline on-line	4, 7, 10, 16, 20
S_{2B}	FT/cmdline on, ST/tokens off-line	1, 5, 6, 9, 19

Table 5: Groups used to partition subjects

All subjects were given an explicit goal of detecting as many failures (runtime behavior that deviates from the specification) in as little time as possible. This goal was stated verbally and written on the instruction sheets. The true goal behind the experiment, namely the comparison of off-line and on-line guidance techniques, was explained to the subjects in the final questionnaire and discussed in the individual debriefing sessions.

Subjects who worked with off-line guidance received a package of materials that included a detailed instruction sheet, a data-collection form, an explanation of the approach for constructing test cases, and the specification for the program. Subjects who worked with the structural-testing exercise additionally received a hard-copy of the program's source code. Subjects who worked with on-line guidance received a different package of materials. Instead of a detailed instruction sheet and data-collection form, they received a brief explanation of how to invoke the on-line system (MVP-S for the needed information and a forms-based tool to collect data). The other materials (explanation of the approach for constructing test cases, and specification for the program and possibly the source code) were identical to those used in the off-line technique.

Termination criteria were partially specified by the quantitative quality models in terms of duration, but the subject was expected to decide when to stop working. The use of a hard time limit was seen as a threat to external validity and therefore avoided. Subjects were instructed to record data on the relevant data-collection instrument immediately after completing the task about which data was required. To prevent subjects from disseminating knowledge about faults and thereby spoiling the exercises for other subjects, all materials were collected after each trial, and subjects were asked not to discuss the experiment.

Most subjects performed both exercises on the same day. However, the exercises were not conducted in large groups. Trials were conducted on a one-on-one basis between subject and principal investigator, although occasionally two subjects worked in parallel. This approach permitted close contact with each subject, as well as iterative enhancement of the study over the course of the trials. However, no fundamental problems were found in the instruments or other procedures. Minor enhancements were made to the guidance materials and program specifications, primarily fixing typographical mistakes and other superficial problems.

6 Results

This section presents data and interpretations, discusses some implications of the results, gives lessons learned about MVP-S, offers a short critique of the experiment, and mentions some issues that deserve further study.

6.1 Data and interpretations

Some raw data from the experiment are shown in Table 6. The following analyses make extensive use of the data for total time and percentage of possible failures detected. These unrelated variables can be combined to obtain a rate with the

ID	L	G	Ex. 1: FT/cmdline (10 possible failures)				Ex. 2: ST/tokens (5 possible failures)			
			R	D	T	E	R	D	T	E
1	Ph.D.	2B	7	4	160	15.0	3	1	140	8.6
2	Ph.D.	1A	5	3	40	45.0	2	1	48	25.0
3	Ph.D.	1B	7	6	100	36.0	4	3	115	31.3
4	Ph.D.	2A	5	3	60	30.0	3	2	112	21.4
5	Ph.D.	2B	2	2	65	18.5	3	1	75	16.0
6	Ph.D.	2B	6	4	62	38.7	2	2	86	27.9
7	M.S.	2A	2	2	65	18.5	1	1	80	15.0
8	M.S.	1A	5	3	38	47.4	1	1	130	9.2
9	M.S.	2B	6	5	66	45.5	3	2	60	40.0
10	M.S.	2A	5	4	116	20.7	3	1	147	8.2
11	M.S.	1A	7	5	56	53.6	3	2	140	17.1
12	M.S.	1B	4	3	55	32.7	3	1	70	17.1
13	M.S.	2B	7	2	77	15.6	2	1	50	24.0
14	M.S.	1A	5	5	71	42.3	2	1	85	14.1
15	M.S.	1B	1	1	47	12.8	1	0	150	0.0
16	M.S.	2A	3	2	38	31.6	3	2	130	18.5
17	M.S.	1A	5	3	75	24.0	2	0	80	0.0
18	M.S.	1B	7	6	51	70.6	3	3	73	49.3
19	M.S.	1B	6	3	35	51.4	3	2	85	28.2
20	M.S.	2A	6	2	40	30.0	3	2	95	25.3
Mean			5.0	3.4	65.9	34.0	2.5	1.5	97.6	19.8
Standard deviation			1.8	1.4	30.3	15.4	0.8	0.8	32.8	12.3

Table 6: Raw data from the experiment. ID is the subject’s identifier, L is experience level, G is group, R is revealed failures, D is detected failures, T is time in minutes, and E is efficiency in percentage of possible failures detected per hour.

unit ‘percentage of possible failures per hour.’ The normalization that results in this somewhat odd unit is necessary because the programs differed in size but had very similar fault densities, making it unrealistic to compare raw data on failures detected per hour. Other than this normalization, the data were not transformed in any other way to perform the analyses. Only the data obtained from the 20 subjects who completed both exercises and the debriefing questionnaire are used in the following analyses. As explained in Section 4.6, all analyses are paired comparisons based on the two observations of each subject.

6.1.1 Hypothesis 1: Efficiency

Hypothesis 1 focuses on how efficiently the subjects worked. The influence of the uncontrolled independent variables (e.g., the subject’s motivation and process conformance) on the results was checked first. All subjects reported that they were well motivated for the exercises, which is consistent with a mostly self-selected sample. No significant correlation of reported motivation with the results was detected. Although the observer detected no significant deviations from the prescribed processes, and process conformance was judged to be acceptable, assessing process conformance is extremely difficult [Sørumgård, 1996]. Next the influence of the controlled independent variables exercise, guidance technique, and order on the rate at which failures were detected is analyzed.

Measure	n	Mean	Std. dev.
Difference in time	20	-31.7	37.7
Difference in % failures	20	5.0	13.9
Difference in rate	20	14.2	11.8

Table 7: Results for effect of exercise (ex. 1 – ex. 2)

Independent variable 1: Exercise. Table 7 summarizes the data used for analyzing the effect of this independent variable, and Figure 1 plots the differences in failure-detection rate attained by each subject (exercise 1 minus exercise 2). Although the exercises were designed to be identical, the figure shows clearly that failures were detected more efficiently in exercise 1. This is confirmed by the statistical analysis: a t-test on the rate data indicates that the differences are significant.

The data show large average differences in the amount of time and small differences in the percentage of failures detected. This suggests that subjects were less efficient when performing exercise 2 because they required extra time, not because

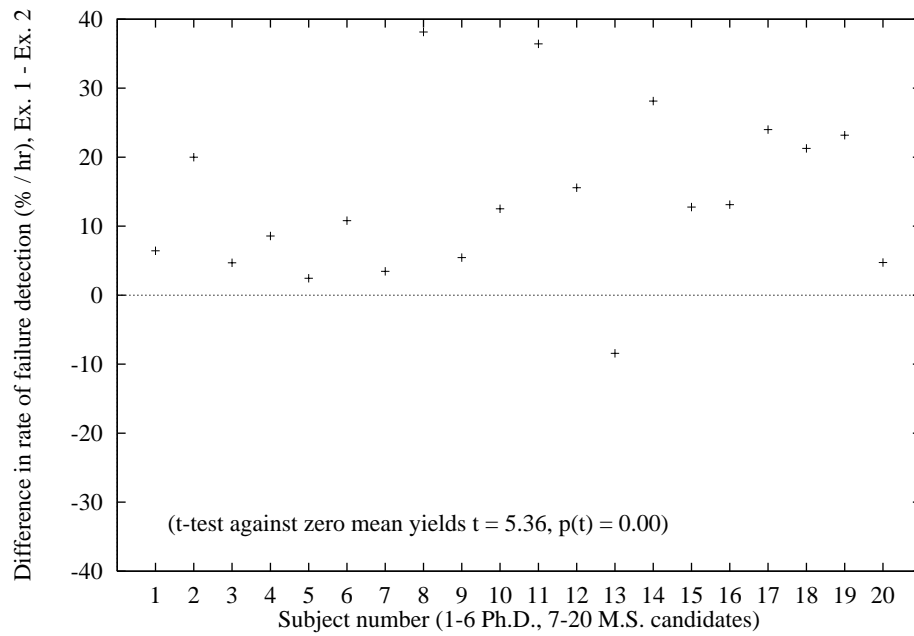


Figure 1: Difference in failure-detection rate, ex. 1 – ex. 2

they detected fewer failures. A possible explanation is the degree of exercise difficulty, as determined by the technique, program size, and complexity. Because the testing technique was confounded with the program, these results cannot be used to compare the relative efficiencies of the testing techniques.

Independent variable 2: Guidance technique. Table 8 summarizes the data used for analyzing the effect of this independent variable, and Figure 2 plots the differences in failure-detection rate attained by each subject (off-line minus on-line). No obvious result is apparent in the figure, but statistical analysis of the rate data indicates a significant difference that favors the off-line guidance technique.

Measure	n	Mean	Std. dev.
Difference in time	20	-9.1	48.9
Difference in % failures	20	5.0	13.9
Difference in rate	20	8.3	16.7

Table 8: Results for effect of guidance technique (off-line – on-line)

Closer examination of Figure 2 reveals that four of the six subjects who were Ph.D. candidates worked more efficiently when using on-line guidance, but only five of the fourteen subjects who were M.S. candidates worked more efficiently when using on-line guidance. Although there are too few data points to support inferences, these results suggest a correlation between performance and experience level for on-line guidance that deserves further investigation.

Independent variables 3 and 4: Ordering. Table 9 summarizes the data used for analyzing the effect of ordering of trials, and Figure 3 plots the differences in failure-detection rate attained by each subject (trial 1 minus trial 2). The figure reveals no clear trend, and statistical analysis of the rate data do not indicate a significant difference.

Measure	n	Mean	Std. dev.
Difference in time	20	13.3	47.9
Difference in % failures	20	7.0	13.0
Difference in rate	20	2.2	18.6

Table 9: Results for effect of order (trial 1 – trial 2)

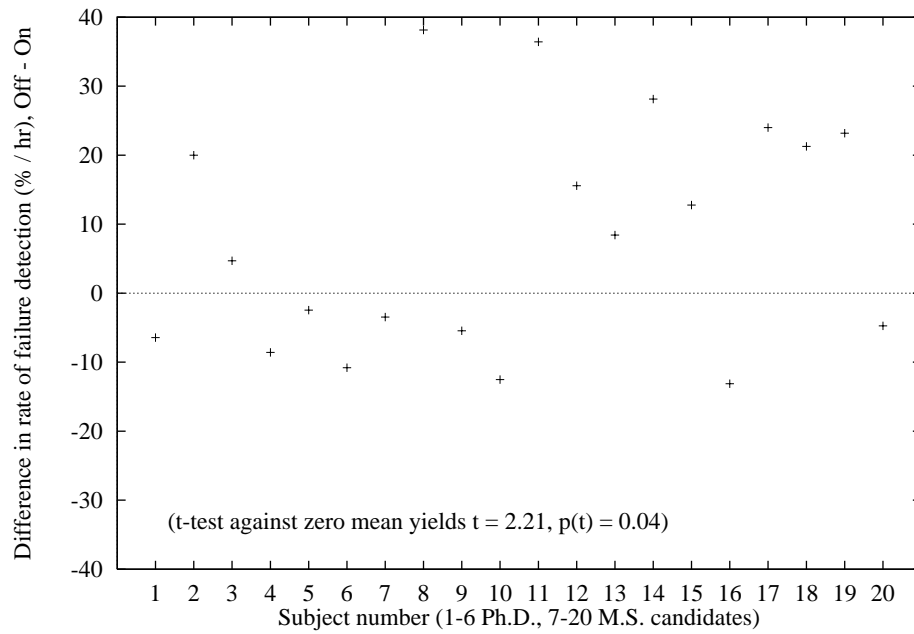


Figure 2: Difference in failure-detection rate, off-line – on-line

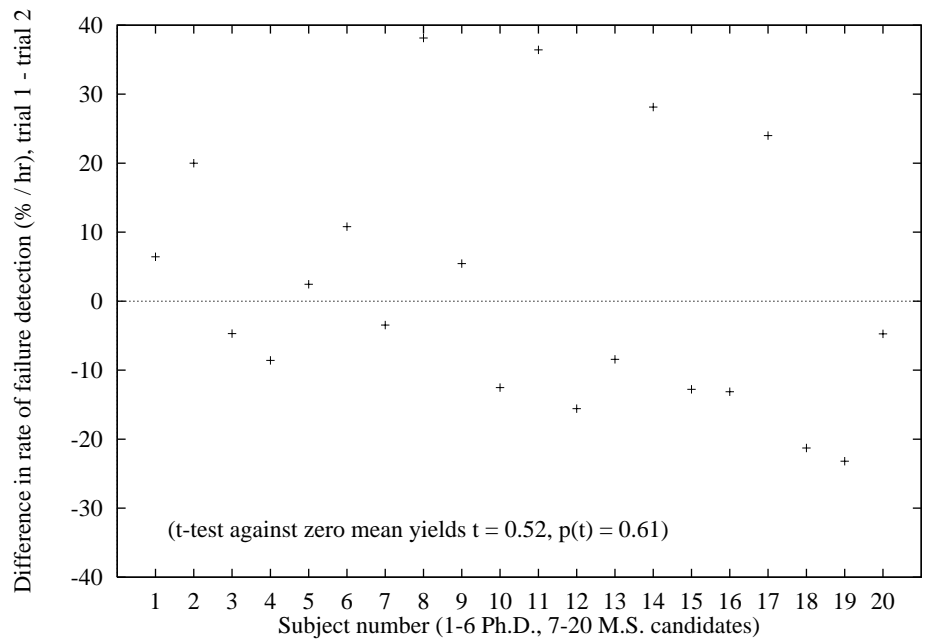


Figure 3: Difference in failure-detection rate, trial 1 – trial 2

Possible interaction effects. The data suggest superior performance by subjects who used off-line guidance during exercise 1 as compared to the subjects who used on-line guidance. No such interaction was suggested for exercise 2. The experimental design precludes the use of a statistical test for this effect. If this effect is real, a mechanism that would explain it is that subjects work better if the work medium and guidance medium are the same. During the experiment, subjects were asked to do considerably more off-line work in exercise 1 as compared to exercise 2. The use of on-line guidance for the off-line work activity may have been detrimental.

Summary of hypothesis 1. The original hypothesis was not confirmed. In fact, the rate data suggests that subjects worked more efficiently on average when using off-line guidance. These data also show that the quantitative quality models for total time were approximately 10–20 minutes too low, and suggest a noticeable difference between the M.S. and Ph.D. candidates. The quantitative quality models could have been improved if the median value for total time from prior observations had been used instead of the mean value.

6.1.2 Hypothesis 2: Acceptance

Hypothesis 2 focuses on the acceptance that a process-centered system finds among subjects. Table 10 summarizes the data collected to evaluate acceptance of the on-line guidance technique among subjects.

Measure	Data		
M3: Guidance tech.	Prefer on-line: 8	Prefer off-line: 10	Don't care: 2
M4: Restrictiveness	Not restrictive: 17	Too restrictive: 3	
M5: Data collection tech.	Prefer on-line: 7	Prefer off-line: 3	Don't care: 10

Table 10: Subjective data concerning acceptance (hypothesis 2)

No subject complained of being treated as a “subroutine” or otherwise felt subjugated to the control of a computer. The results are unclear with regard to data collection due to the large number of “don’t care” responses.

Overall, more subjects preferred off-line guidance than on-line guidance. However, the Ph.D. candidates (i.e., more experienced subjects) who participated were unanimous in preferring on-line guidance. A possible confounding factor was the density of the information that can be presented on a computer screen versus on paper using 1996 technology. These results are partially explained by the following two work styles articulated by the subjects:

- The subjects who preferred on-line over off-line process information indicated that they liked seeing precisely that information that was relevant to a single step (a focused view). They also enjoyed being rid of a pile of paper through which they had to search frequently.
- The subjects who preferred off-line over on-line process information indicated that they liked seeing all the necessary information at once. In other words, they preferred an overview, and didn't like the small doses of information that the system offered to them. Some also mentioned that they needed to write on the instruction sheets to work effectively, which was not possible in the on-line guidance technique.

6.1.3 Hypothesis 3: Comprehension

Hypothesis 3 focuses on the subject's comprehension of off-line versus on-line guidance materials. The results are summarized in Table 11.

Measure	Data	
M. 6: Guidance materials were...	Comprehensible: 18	Incomprehensible: 2
M. 7: Information was...	Sufficient: 18	Insufficient: 2
M. 8: Obs. opinion of comprehension:	Good: 16	Fair: 4

Table 11: Subjective data concerning comprehension (hypothesis 3)

Most subjects stated that the guidance materials were comprehensible, and the comprehensibility was neither improved nor worsened by offering them on paper or on screen. Most felt that the guidance and other instruments offered sufficient information to accomplish the exercises. Those who felt that they had insufficient information mainly complained about needing more help in developing equivalence classes in the functional testing exercise. Finally, most subjects responded very well when interrupted and queried about the step that they were currently performing.

6.1.4 Hypothesis 4: Use of models

Hypothesis 4 focuses on the use of quantitative quality models. The question of whether subjects adjusted their behavior was answered by examining the results and by querying the subjects. Because all subjects received the quantitative quality models for the suggested amount of time, number of equivalence classes (FT), and target coverage values (ST), no tracking or comparison was done to evaluate this

hypothesis in a controlled way. Of the 20 subjects, 9 reported that they felt the models were useful and they adjusted their behavior, 4 subjects reported that the models were harmful and they did not adjust their behavior, and 7 reported that the models were both helpful and harmful.

Subjects adjusted their behavior in response to the quantitative quality models in large part based on their own performance:

- If a subject's own performance matched the quantitative quality models fairly well, the subject felt the value of his or her work was confirmed, saw the model as a goal that had been attained, and consequently felt favorably towards the quality models. After attaining that goal, they were highly likely to end that step and move on. The difference between more and less-experienced subjects was again apparent; five of the six Ph.D. candidates found these models helpful and indicated that they adjusted their behavior.
- If a subject's own performance deviated from the quality models sharply, the subject was likely to report feeling pressured and unpleasantly influenced. Subjects did not abandon the exercise and move on simply because they had exceeded the estimated time or generated more than the estimated number of test cases, to name two examples. Most of the less experienced subjects fell into this category.

It was possible for both situations described above to arise during a single exercise. The conclusion is that the quantitative quality models had a significant influence, but it was not always positive.

6.1.5 Hypothesis 5: Reporting speed

Hypothesis 5 involves the time lapse between completing a step in the exercise and reporting data. This refers both to reporting the process state (i.e., start or completion of a step) and reporting data about some step such as its duration. Of the 20 subjects, 12 reported with no noticeable delay (i.e., under 5 minutes), 7 reported data with a short delay (i.e., within 5-10 minutes), and 1 subject reported data after a long delay (i.e., over 10 minutes). The subject's awareness that they were taking part in an experiment was probably a strong influence.

6.2 Implications of the results

The debriefing sessions revealed that differences in subject's acceptance of on-line process guidance may have depended heavily on their understanding of the goals of the exercises. For example, the goal of the functional testing exercise was to test the

program thoroughly against the specification. We assume that inexperienced people lacked an understanding of the overall goals behind functional and structural testing. Therefore, they preferred a comprehensive overview that helped them understand the goal, which was best supported by off-line guidance. In contrast, the experienced people probably understood the goals, and therefore preferred seeing only what they needed at each step, which was best supported by on-line guidance. These results suggest providing inexperienced people with books or other comprehensive overviews, and providing experienced people with filtered, directed views of the necessary process information.

6.3 Lessons learned about MVP-S

No subject reported any real difficulty with using MVP-S in the experiment. However, nearly all subjects complained that the interface was awkward to use and needed work. Many offered ideas about possible improvements. This experience suggested the following changes to the interface:

- Use immediately accessible buttons or other convenient interface facility for frequently accessed features “start process” and “complete process.”
- Offer a graphic overview of all steps, possibly a view of just product flow.
- Offer a “print” feature so that users who prefer a plain-paper version of the information could gain an overview easily and could scribble notes on that output.

6.4 Critique of the experiment

Researchers who are interested in repeating this experiment should pay special attention to the following issues. First, the size of the exercise was kept small to help recruitment; a larger, more difficult exercise would help the external validity of the results. Second, the experience of the subjects with the techniques used in the exercises and with on-line process guidance systems should be assessed very carefully. Third, the decision to offer no training might not be appropriate for other environments. Fourth, although no fatigue effect was confirmed, several subjects stated that they should not have attempted both exercises in a single morning or afternoon. Fifth, the measures taken of a subject’s opinions should use something like a five-point scale. Sixth, it may be possible to measure the amount of time spent seeking guidance separately from the amount of time spent performing the exercise, if this can be done in a manner that is not too intrusive. Finally, a second debriefing session that is conducted *after* the data have been analyzed may gain the experimenter additional insights into the results.

6.5 Other issues

The following questions were raised by these results and deserve investigation.

Subjects overlook failures. Can testers be trained to search more diligently for failures? Subjects overlooked a significant number of failures revealed by their test cases. It appears that testers could significantly improve their performance simply by searching for failures more diligently.

Unfamiliar tasks. How do people react when confronted with unfamiliar tasks? Very different reactions were observed in the experiment. Some subjects slowed down while they attempted to learn how to accomplish the task properly. Others raced through without much thought.

Context switching. Can the cost of interruptions be reduced by keeping developers better informed about the current status of their activities? Many subjects stated that they liked the list of activities and status values for each activity maintained by the on-line guidance system, and that they found it easy to resume work after an interruption.

Activities not involving the computer. Is automated support helpful for off-line constructive activities? For example, what are the consequences of putting a definition of an off-line process such as code reading on the computer?

Iteration. What intelligent support can be provided for iteration? After completing a task, personnel frequently realize how they should have done the task, and in some cases will need to iterate through that task at least once more. An on-line guidance system should support this naturally.

7 Conclusion

We conducted a controlled experiment that compared the use of off-line and on-line measurement-based process guidance for individuals who performed small exercises. The experiment contributed an initial understanding of several variation factors that seem to influence how users interact with process guidance systems and quantitative quality models, specifically a user's experience level, personal work style, and performance on the exercise. Post hoc analysis identified strong correlations between subject experience level and preference for the type of guidance, a result that deserves further investigation.

A kit of materials is available in both German and English-language versions for other empiricists who would like to repeat this experiment. Replication of this experiment will permit checking whether the variation factors identified in this study can be confirmed when using other process guidance systems, exercises, and subjects.

Acknowledgements

Many thanks to Victor R. Basili for his assistance with all aspects of this experiment as well as with Table 1, to H. Dieter Rombach for supporting and supervising this research, and to all the subjects who participated in the experiment. Thanks also to Lionel Briand, Alfred Bröckers, Allen McIntosh, Adam Porter, Larry Votta, Stuart Zweben, and the anonymous reviewers for their suggestions and helpful criticism of the paper.

References

- [Basili, 1994] Basili, V. R. (1994). A research agenda for ISERN: Validating the Quality Improvement Paradigm. Presented at the 1994 ISERN annual meeting.
- [Basili and Perricone, 1984] Basili, V. R. and Perricone, B. T. (1984). Software errors and complexity: An empirical investigation. *Communications of the ACM*, 27(1):42–52.
- [Box et al., 1978] Box, G. E. P., Hunter, W. G., and Hunter, J. S. (1978). *Statistics for Experimenters*. John Wiley & Sons, New York.
- [Brooks, 1980] Brooks, R. E. (1980). Studying programmer behavior experimentally: The problems of proper methodology. *Communications of the ACM*, 23(4):207–213.
- [Campbell and Stanley, 1966] Campbell, D. T. and Stanley, J. C. (1966). *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin, Boston. ISBN 0-395-30787-2.
- [Christie, 1995] Christie, A. M. (1995). *Software Process Automation, The Technology and Its Adoption*. Springer-Verlag.
- [Dart et al., 1987] Dart, S. A., Ellison, R. J., Feiler, P. H., and Habermann, A. N. (1987). Software development environments. *IEEE Computer*, pages 18–28.

- [Kamsties and Lott, 1995] Kamsties, E. and Lott, C. M. (1995). An empirical evaluation of three defect-detection techniques. In Schäfer, W. and Botella, P., editors, *Proceedings of the Fifth European Software Engineering Conference*, pages 362–383. Lecture Notes in Computer Science Nr. 989, Springer–Verlag.
- [Lott, 1994] Lott, C. M. (1994). Measurement support in software engineering environments. *International Journal of Software Engineering & Knowledge Engineering*, 4(3):409–426.
- [Lott, 1996] Lott, C. M. (1996). *Measurement-based feedback in a process-centered software engineering environment*. PhD thesis, Department of Computer Science, The University of Maryland, College Park, Maryland 20742. Available online at <http://www.cs.umd.edu/users/cml/>.
- [Lott et al., 1995] Lott, C. M., Hoisl, B., and Rombach, H. D. (1995). The use of roles and measurement to enact project plans in MVP-S. In Schäfer, W., editor, *Proceedings of the Fourth European Workshop on Software Process Technology*, pages 30–48, Noordwijkerhout, The Netherlands. Lecture Notes in Computer Science Nr. 913, Springer–Verlag.
- [Lott and Rombach, 1993] Lott, C. M. and Rombach, H. D. (1993). Measurement-based guidance of software projects using explicit project plans. *Information and Software Technology*, 35(6/7):407–419.
- [Perry and Kaiser, 1991] Perry, D. E. and Kaiser, G. E. (1991). Models of software development environments. *IEEE Transactions on Software Engineering*, 17(3):283–295.
- [Sørumgård, 1996] Sørumgård, S. (1996). An empirical study of process conformance. In *Proceedings of the 21st Annual Software Engineering Workshop*. NASA Goddard Space Flight Center, Greenbelt MD 20771.