Cognitive Engineering Laboratory

Research on the Characteristics of Long-Term Adaptation (III)

Gerard L. Torenvliet & Kim J. Vicente

CEL 99-03

Final Contract Report
September, 1999

Prepared for
Japan Atomic Energy Research Institute

Cognitive Engineering Laboratory, Department of Mechanical & Industrial Engineering, University of Toronto
5 King's College Road, Toronto, Ontario, Canada M5S 3G8
Phone: +1 (416) 978-7399 Fax: +1 (416) 978-3453
Email: benfica@mie.utoronto.ca URL: www.mie.utoronto.ca/labs/cel/

Director: Kim J. Vicente, B.A.Sc., M.S., Ph.D., P. Eng.

The Cognitive Engineering Laboratory (CEL) at the University of Toronto (U of T) is located in the Department of Mechanical & Industrial Engineering, and is one of three laboratories that comprise the U of T Human Factors Research Group. CEL was founded in 1992 and is primarily concerned with conducting basic and applied research on how to introduce information technology into complex work environments, with a particular emphasis on power plant control rooms. Professor Vicente's areas of expertise include advanced interface design principles, the study of expertise, and cognitive work analysis. Thus, the general mission of CEL is to conduct principled investigations of the impact of information technology on human work so as to develop research findings that are both relevant and useful to industries in which such issues arise.

Current CEL Research Topics

CEL has been funded by Atomic Energy Control Board of Canada, AECL Research, Alias|Wavefront, Asea Brown Boveri Corporate Research - Heidelberg, Defense and Civil Institute for Environmental Medicine, Honeywell Technology Center, Japan Atomic Energy Research Institute, Microsoft Corporation, Natural Sciences and Engineering Research Council of Canada, Nortel Networks, Nova Chemicals, Rotoflex International, Westinghouse Science & Technology Center, and Wright-Patterson Air Force Base. CEL also has collaborations and close contacts with the Mitsubishi Heavy Industries and Toshiba Nuclear Energy Laboratory. Recent CEL projects include:

- Developing advanced human-computer interfaces for the petrochemical industry to enhance plant safety and productivity.
- Understanding control strategy differences between people of various levels of expertise within the context of process control systems.
- Developing safer and more efficient interfaces for computer-based medical devices.
- Creating novel measure of human performance and adaptation that can be used in experimentation with interactive, real-time, dynamic systems.
- Investigating human-machine system coordination from a dynamical systems perspective.

CEL Technical Reports

For more information about CEL, CEL technical reports, or graduate school at the University of Toronto, please contact Dr. Kim J. Vicente at the address printed on the front of this technical report.

…instead of searching for mechanisms in the environment that turn organisms into trivial machines, we have to find the mechanism within the organisms that enable them to turn their environment into a trivial machine.

(van Foerster, 1984, p. 171)

ABSTRACT

While interacting with complex systems, operators of these systems frequently modify their interfaces in order to achieve task goals more effectively or more efficiently (what Rasmussen (1986) has called 'finishing the design). While there is a literature on operator modifications, this literature includes only field studies and theoretical treatments, but has not proceeded to experimentation. This research describes two preliminary experiments designed to answer four questions related to operator modifications: (1) Why do operators modify their interfaces? (2) How do these modifications develop over time? (3) How does this activity affect performance? and (4) How does this activity affect understanding? The results of these experiments pointed to answers to most of these questions, and most significantly, pointed to a taxonomy of the purposes behind operator modifications in complex systems.

ACKNOWLEDGMENTS

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, contract monitor).

A number of people were instrumental in completing this research. First thanks go to my supervisor, Dr. Kim Vicente, for supporting and guiding this research effort. Deeper than that, I would like to thank Kim both for introducing me to the field of cognitive engineering during my undergraduate education, for fostering my enthusiasm in the field, for putting up with my weaknesses, and for developing my research career. Thanks are also due to Dr. Cathy Burns, who has helped me to flesh out a number of the ideas presented in this thesis, and who has helped me out with other research leading up to this thesis. John Hajdukiewicz also deserves mention for putting up with my barging into his office to unload some new idea or insight, and for encouraging me in my work.

Finally, this thesis is dedicated to three important individuals: to my father who sparked my love of science, to my mother, who has nurtured it, and to my wife, who has persevered with it. I have been blessed to have each of you in my life.

To the glory of God alone.

G.T., July 1999

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

INTRODUCTION

In a sixteen-month period spanning the years 1975 to 1977, a group of three researchers spent a great deal of time studying five nuclear power plant control rooms. At this early date in the history of cognitive engineering, their stated purpose was to conduct a thorough evaluation of these control rooms "in response to technical criticisms regarding the lack of attention to human factors in nuclear power plants" (Seminara, Gonzalez, & Parsons, 1977, p. 1-1). Using techniques such as interviews, task analyses, procedure evaluations, and even basic anthropometrics, they were able to uncover a host of significant problems in contemporary control room design and management. As the investigators summed it up, "In general, the study findings paint a rather negative picture" (Seminara et al., 1977, p. 1-3).

With this picture in mind, the question arises as to why major nuclear power plant accidents like Three Mile Island were not happening with frequency. After all, if the design and management of nuclear power plants were so poor, task demands would surely outstrip operators' abilities from time to time. The answer to this question is that, despite the interfaces that operators had to use, the operators themselves were extremely creative, flexible, and adaptive. They were able to close the gap between poor design and safe, productive use.

How did they accomplish this? One phenomenon observed by Seminara et al. (1977) was operators' actions to modify their interfaces to make them more relevant to the tasks at hand. Modifications made by operators were not trivial, and included re-labelling of control panels, highlighting problematic elements of the interface design, and tracing out process flow diagrams to link the controls. In one instance, they even went to the effort of replacing two visually similar levers that performed different functions with a set of more easily discriminated controls – Heineken and Michelob beer taps (Figure 1)! In effect, operators started with the product left

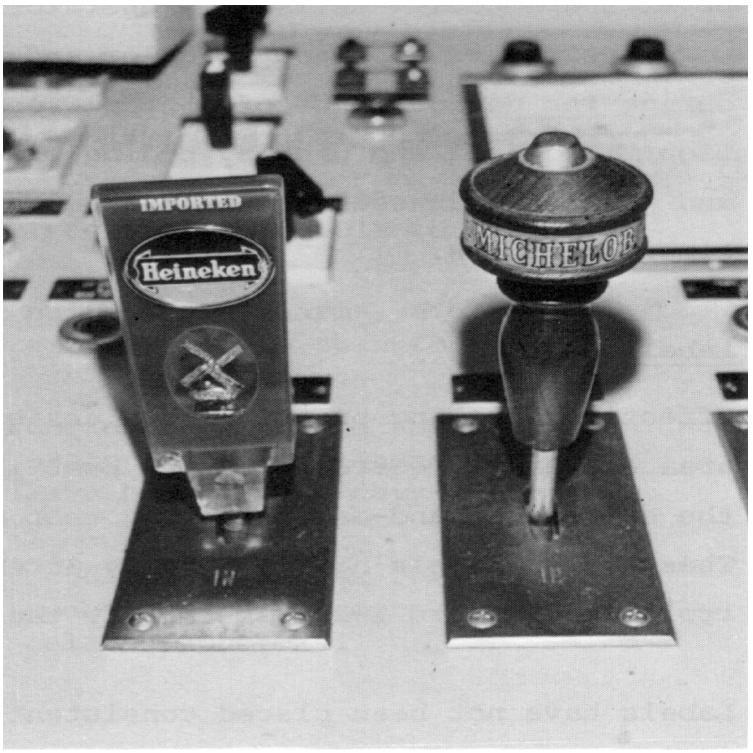


Figure 1. One of many operator efforts at finishing the design observed by Seminara et al.. (Reproduced from Seminara et al., 1977.)

to them by designers and proceeded to finish the design[1].

Even though these instances of operator modifications were observed already back in 1977, save for a few field studies, the human factors community has rarely explicitly studied this practice. In other communities, the most notable work has been performed by a number of researchers in artificial intelligence and cognitive science, who have contributed a number of general theoretical studies to the literature. These studies have been important, and have led to a number of significant insights about tool use in general and in the particular case of complex sociotechnical systems. However, since these studies were based on either informal or structured observation of people working in situ, questions about causality were not rigorously addressed. As Vicente (1997a) has observed, field studies are important, but they should be the point of

[1] Rasmussen (1986) is to be credited with coining this instructive term.

departure for a programme of more controlled research in the laboratory. It is precisely here where there is a gap in the literature on operator modifications.

This thesis is an effort to study operator modifications from the relatively more controlled environment of the laboratory. Since there does not seem to be any other laboratory research addressing operator modifications in complex sociotechnical systems, the research documented here is a preliminary effort and so was exploratory in nature. Still, even if it was not informed by any specific theory or guided by a set of rigorous hypotheses, this research was motivated by a number of broad questions about the nature and impact of operator modifications across different interfaces into the same work domain.

The pages that follow describe these questions and how they were addressed in a series of laboratory experiments using a process control micro-world as a testbed. To position this research within the greater body of human factors research, the literature on operator modifications is first reviewed, and the motivation for this study is more fully outlined. Since this research addresses a novel experimental problem, a pilot study that was used to help gain a better understanding of the phenomena at work, the conditions necessary to promote them, and how to measure them is described next. This leads to the description of a larger, more rigorous experiment, the design, procedure, and results of which are discussed in turn. Finally, a number of concluding remarks are made to place operator tool use within a broader framework, to address a number of limitations of this present work, and to point to some avenues for further study.

BACKGROUND AND MOTIVATION

The work in this thesis was motivated by a number of papers written in both the artificial intelligence and cognitive engineering communities on the topic of tool use and operator modifications. As already hinted at in the introduction, this work can be grouped under two broad categories. The first category includes a number of general theoretical treatments, and the second includes field studies that were carried out in a number of diverse domains including desktop computing, anaesthesiology, and nuclear power plant operations. After the work done in each of these categories is summarised, efforts by Vicente, Mumaw, and Roth (1997) and Vicente (1999) to construct a preliminary framework for operator tool use will be presented. The work from these diverse domains will then be summarised to motivate the present research.

General Theoretical Treatments of Tool Use

Researchers from the artificial intelligence (AI) community have been performing research to understand how humans are able to cope with the complex environments and task situations they encounter every day. Within the broader scope of AI research, one aim of this work is to understand how humans make their environments more predictable and less complex, so that simpler models of human activity can be generalized across contexts. Two papers from within the AI community will be presented here. The first (Hammond, Converse, & Grass, 1995) introduces stabilization, a useful concept that describes how agents modify their environments over the long-term. The second (Kirsh, 1995) places the concept of stabilization within a broader framework, and then works to flesh out areas of this framework that deal with environmental modifications that happen on the shorter-term.

Stabilization. Hammond et al. (1995) have as their starting point a concern with traditional ideas of "general purpose intelligence" in the AI community. Along with Lave (1988), they are sceptical of appeals to domain-independent intelligence. Instead they believe that context drives

human behaviour. How can human agents act in the face of so many different contexts without resorting to some sort of general-purpose intelligence? Hammond et al. argue that in order to be successful, one or more of the following must be true:

- The agent, the environment, or both, must be designed with the other in mind.
- The agent is able to learn to adapt to its environment.
- The agent is able to change the environment to suit itself, an activity that Hammond et al. call stabilization.

An example they give in support of this framework is of a person who moves into and successfully adapts to a new home. She is successfully able to live within this context for three reasons. In the first place, this is because the design of the home reflects the needs of the general class of occupants (e.g., children and adults). Second, as the person interacts with her new home, she will learn about its layout, uses, and quirks. Third, she will also effect considerable change to the home by adding appliances, furniture, decorations, and so on. Further, the last two reasons for her success are complementary. As she learns about her home, she gains ideas as to how to further stabilize it, and as she works in the context of her stabilizations, she will learn more about her home.

Hammond et al. then go on to describe an experiment they carried out with an AI model that implemented the concept of stabilization in a simulated actor working in the context of a virtual woodworking shop. While the particulars of this model will not be discussed here, it is important to note that Hammond et al. were able to observe performance improvements in this context as the actor stabilized its environment. Even more important, the general concept of stabilization outlined by Hammond et al. is a valuable perspective from which to consider operator tool use. Their example of a person moving into a new home can easily be extended to

an actor learning to work in the context of a new complex domain. While it is hoped that the actors' task support will be designed to reflect an understanding of the work domain and the actor's relevant capabilities and limitations, where this is not the case actors are able to both adapt to and adapt the task support in efforts to achieve the task goals. Also valuable is the idea that human adaptation to a given context complements their adapting of that context. If this is true, it is possible that operator tool use is the result of operator learning and adaptation, and that these tool uses can themselves be springboards for further learning and adaptation.

How stabilization happens. While Hammond et al. (1995) were interested in the relatively long-term activity of stabilization, Kirsh (1995) is more interested in understanding the shorter-term modifications that human actors make in the context of a well-structured environment. In other words, if Hammond et al. were interested in understanding how a person would structure the furniture and appliances in her apartment, Kirsh is interested in understanding how that same person would structure her cooking implements and ingredients while cooking a single meal in the kitchen of that apartment. Kirsh discusses this shorter-term activity under the general title of intelligent space management.

Kirsh strongly believes that the intelligent management of the space that we act in is an integral part of our ability to act over many different contexts. The manifestations of our space management can be complex and diverse, but they serve to simplify cognitive and physical tasks in a context-relevant way. Or, as Kirsh puts it:

> …whether we are aware of it or not, we are constantly organizing and re-organizing our workplace to enhance performance. Space is a resource that must be managed, much like time, memory, and energy. When we use space well, we can often bring the time and memory demands of our tasks down to workable levels. We can increase the reliability of execution and the number of jobs we can handle at once. The techniques we use are not always obvious, or universal. Some were taught to us, some naturally evolved as we improved our performance through practice, some are inevitable consequences of having the type of bodies, manipulators, and sensors we have. In every case, though, the reason

> space can be used to simplify our cognitive and physical tasks is because of the way we are embedded in the world. (Kirsh, 1995, p. 32)

In short, Kirsh's view is that spatial and informational modification of our environment is an efficacious and frequent occurrence. If we do not naturally notice these types of activities, Kirsh argues, it is because they are ubiquitous in the human experience.

Kirsh also stresses that spatial modification is a hallmark of expert behaviour. His view of expertise is consistent with those in the camp of situated activity (e.g. Lave, 1988; Suchman, 1987), and so he understands that experts generally do not have to construct plans prior to action execution. Instead, they use the ample information they have locally available to act in a flexible and context-tailored way. To help in this activity, experts jig their environments with informational artefacts that serve as cues for action by reducing the perceived degrees of freedom for action.

In this context, Kirsh's main interest was to build a classification of spatial modifications in terms of the perceptual and cognitive ends achieved. This classification included three elements:

- Spatial modifications to simplify choice. These modifications bound the size of an actor's decision space or highlight paths through it by highlighting or hiding the affordances of objects. For instance, Kirsh gives the example of preparing a salad along a kitchen assembly line. Although whole vegetables always afford being both washed and cut, in a kitchen assembly line objects close to the sink are more likely to be washed and those close to the cutting board are more likely to be cut.
- Spatial modifications to simplify perception. Instead of bounding the size of a decision space, this category includes spatial modifications that facilitate perception, making it easier to notice the relevant affordances of objects. For instance, Kirsh gives the example of veteran

jigsaw puzzlers who group similar puzzle pieces together to highlight the subtle differences between these pieces.

- Spatial modifications that save internal computation. Finally, this category includes uses of space that serve to offload computation from the head to the world. For instance, Kirsh observed that Tetris players rotate game pieces on the screen instead of in the head. On screen these rotations are quick and essentially error free; in the head they are costly and error-prone.

Two things are notable about the categories Kirsh describes. First of all, they reflect the tight coupling between human actors and their environment. Human actors can become sufficiently attuned to their environments to recognize opportunities for functional modifications, and then act on them. Second, the examples given by Kirsh highlight how effective each of these types of modifications can be in use. While the net impact of each modification may not be dramatic, what is dramatic is how easily these modifications reduce the burden on scarce cognitive and perceptual resources.

Moving From Theory to the Field

Before moving on to discuss a number of field studies of operator modifications, it is important to take stock and summarise the contribution of these studies from the AI community. First, it is notable that both of the papers above discuss tool use from an ecological perspective (Flach, Hancock, Caird, & Vicente, 1995; Gibson, 1979/1986; Hancock, Flach, Caird, & Vicente, 1995). Hammond et al. (1995) stress the mutuality of actor and environment and place their concept of stabilization within a perception-action loop, and Kirsh (1995) similarly emphasises the tight coupling between actor and environment. Kirsh's analysis also attempts to give an account for both the environment and the organism acting in that environment. Thus the

approach of these studies dovetails with continuing research applying an ecological perspective to the design of human-machine systems (see Hancock et al., 1995; Vicente, 1997b; Vicente, 1999). Second, although they are supported by mostly anecdotal evidence, these studies are at least a preliminary place to begin in understanding and designing experimental investigations to promote spatial modifications. Hammond et al.'s (1995) concept of stabilization seems to imply that spatial modifications are long-term adaptations that occur as agents learn about and become more closely coupled to their environments. Kirsh's (1995) preliminary classification re-emphasises this point (recall his stress on expertise) while also pointing to the broad categories of spatial modifications that might be seen in a laboratory investigation. Together, they point to the necessity of experience and expertise to the emergence of operator modifications. Third, Kirsh's study also presents the preliminary hypothesis that the net effect of spatial modifications on performance may not be large, but that each modification achieves its ends easily and effectively. Accordingly, it may be most beneficial to work on understanding how these modifications achieve their cognitive savings so easily rather than how much they improve performance.

Field Studies of Tool Use

To supplement the theoretical treatments of tool use above, members of the human factors community have carried out a number of field studies. These field studies, from the domains of desktop computing, anaesthesiology, and nuclear power plant operations, are described below.

Desktop computing: Tailoring and Use.

One domain in which user modifications are frequent is that of desktop computing. At one level, much software is designed with an inherent degree of tailorability, and competent users exploit these features to personalize their computing workplace. At a deeper level, expert users write macros and auxiliary programs in support of

their tasks. Henderson and Kyng (1991) have explored both of these types of user modifications in order to derive a number of implications for the design of software. A number of their findings are outlined below.

Since much software is designed to give the user opportunities for customization (for instance, window sizes and default fonts can be changed, styles can be created, and some interface elements can be added or deleted at will), Henderson and Kyng first make a distinction between the use and tailoring of software. Modifications that can be considered as use are those that are made to the subject of that tool. For example, documents are the subject matter of word processors, so any changes made to documents with that tool are considered use. Tailoring involves making modifications to the tool itself, and includes such things as the creation of new styles, the modification of button bars, and the programming of macros. Of course, this is not a crisp distinction, as one person's use could be another's tailoring, but the gist of the distinction is easily applied across contexts.

Having made this distinction between use and tailoring, Henderson and Kyng next make a classification of different types of tailoring. Note that their classification is somewhat different from that of Kirsh (1995). Whereas Kirsh's distinctions deal with the cognitive and perceptual ends of various spatial modifications (i.e., to simplify choice, aid in perception, or aid in computation), Henderson and Kyng classify user modifications by the magnitude of the change vis à vis the original design. Small changes are made when users modify various interface parameters to achieve a new behaviour while still staying well within the realm of the designers' intent. More complex changes are made when users combine existing and understood functionality in a novel way to achieve a new function. These changes achieve something beyond the intent of the system designers while still respecting the constraints of the interface.

Finally, changes can be made that go beyond the constraints of the interface and the designers' intent by modifying the source code of the software or by patching in new modules.

Anaesthesiology: Tool use and learning. Moving into the domain of complex sociotechnical systems, Cook and Woods (1996) performed a study of the effects of the introduction of new technology into the operating room for the support of anaesthesiologists during open-heart surgery. The authors observed anaesthesiologists' performance during 22 surgical procedures directly after the introduction of a new physiological monitoring device. This new device integrated the information from what were previously four devices into a single CRT with multiple windows and a large space of menu-based options. Since the physiological information sampled by these devices could not be fed into multiple machines, instead of having a phased switchover, one evening the old devices were replaced by the new.

The four previous displays were each devoted to a single sensor system, and so provided very few degrees of freedom in displaying information. While these devices did not provide a great deal of flexibility in information presentation, very little effort was required to extract information from them. The new device, on the other hand, allowed users to combine many different types of data in a variety of ways, and so provided a large number of degrees of freedom for information presentation. The cost of this flexibility came in the form of increased effort to extract information from the device. The anaesthesiologists now had a complex menu hierarchy to learn and navigate through to access the information provided.

In the face of this new technology, Cook and Woods present the anaesthesiologists' task as one of bridging the gulf between the tasks inherently supported by the new device and the actual tasks in the operating room. The anaesthesiologists achieved this by a combination of what Cook and Woods call system tailoring and task tailoring. System tailoring involves modifying

the new device to suit the tasks at hand while task tailoring involves the modification of work practices to fit the new device (compare to Hammond et al.'s (1995) twin concepts of agent adaptation and stabilization). System tailoring primarily took the form of an extensive procedure of window manipulations used to obtain an overview display of various blood pressures. "Users expended substantial effort to develop this representation, force its appearance at the beginning of each case, and maintain it in the face of automatic window management conducted by the computer system. … It was clear that this representation…supported the [intended] task" (Cook & Woods, 1996, pp. 600-601). Over the 14 weeks spanned by this investigation, many of these efforts at system tailoring became constant fixtures of the anaesthesiologists' interaction with the device. In most cases, task tailoring took the form of interface navigation where previously this effort was not needed. Task tailoring also appeared when the anaesthesiologists took to interpolating rough estimates of a number of values from graphs because the interface delay in displaying the values in a digital form was unacceptably long. Between these dual efforts of system and task tailoring, the anaesthesiologists were able to successfully bridge the gulf between the new device and their actual tasks. Fortunately, this new device did not result in any patient deaths over the course of the procedures observed; unfortunately, the necessary efforts of system and task tailoring placed an additional (and largely avoidable) burden on the anaesthesiologists.

Nuclear power plant operations: Hard-wired control panels. Vicente and Burns (1995) performed a field study of the cognitive demands placed upon nuclear power plant operators under normal operating conditions. This study was performed in an older, hard-wired control room. Instead of having a number of flexible computer-based displays, this control room generally had one indicator per sensor. Consequently, each indicator was always visible and did

not have to be called up from within a hierarchy of computer 'pages'.

Vicente and Burns were able to compile a lengthy list of design deficiencies that made monitoring in this plant a difficult task, but they also found that the operators had an entire toolbox of strategies to aid them in monitoring that went well beyond any formal training they had received. A number of these strategies are highlighted below:

1. Using historical logs for context. During shift turnover, operators were observed reading through the past few days' entries in the plant's log to understand the context of current operations. In a similar effort to learn the context of the plant, some operators would walk around the control panels writing down the exact numerical values on the displays (procedures only require them to check off that the parameters are within normal ranges). This helped the operators to identify the parameters that were close to their boundary values and thus that needed close monitoring.
2. Monitoring alarms to offload memory. If an alarm was triggered and there was a reason to suspect that the alarmed parameter would continue to rise, operators sometimes recorded the value of the parameter. Since the alarm would not sound again if the parameter continued to increase, this activity offloaded some of the mental effort of monitoring these parameters from the head to the world.
3. Changing alarm setpoints across contexts. If the displays allowed it, operators would achieve the same end as writing down the values of alarmed parameters by changing the setpoints on an alarm after it had sounded. If the parameter continued to rise, it would re-trigger the alarm. This technique was applied across various task contexts, and saw the operators turning the alarms into cues for specific actions. For instance, if the operators were emptying a large tank, rather than constantly monitoring the level of the tank, they would change the

alarm setpoints to indicate when the tank was empty. This strategy obviated continuous visual monitoring.

4. Using post-its to mark exceptions. Operators would sometimes flag unusual indications with post-its. These markers served to remind the operators not to react as usual to those indicators.
5. Opening strip-chart recorder doors to make them stand out. The control room studied by Vicente and Burns had a bank of strip-chart recorders to display trend information on a number of work domain parameters. If operators had to monitor one of these strip-charts more closely, they would leave its cover open. This worked to distinguish the one strip-chart recorder from the others and so lessen the cognitive burden on the operators.

In short, the operators have devised a great number of ingenious strategies to aid in monitoring the plant. Some techniques worked to offload memory and others to create new information on the display. Each modification addressed an issue that was not fully considered during the design of the interface. In sum, operators at this plant were actively engaged in efforts to finish the design.

Nuclear power plant operations: Computer-based displays. Vicente, Mumaw, and Roth (1997) performed a second study of operator monitoring under normal conditions, but this time in a plant equipped with more modern and flexible computer based displays. Vicente et al. summarise the design of this control room as follows:

> ...the computer-based display set at [the plant] is both relatively comprehensive and flexible. It consists of many different display screens, many more than can possibly be viewed at any one time, even with the large number of CRTs available. Also, the same display can be brought up on many different CRTs, and the same variable can be viewed in different ways. (Vicente et al., 1997, p. 14)

The types of operator modifications observed at this plant were quite different in form from those observed at the hard-wired plant discussed above. Most of the tailoring activities in this

plant were directed at managing navigation through the extensive computer support system. As an example, each of the units were equipped with four CRTs in close proximity to the control panels as well as four desktop CRTs. Each of these were able to display many different 'pages' of information, as selected by the operators. However, instead of constantly changing the display pages to view different types of information, the operators adopted a stable configuration for the four control panel CRTs that essentially formed an overview display. With these four CRTs dedicated to this overview configuration, operators used the four remaining CRTs on their desk to call up context-specific information that would help to flesh out the picture painted by the overview displays. The selection of screens for the overview display was not ad hoc, but rather, each of the screens were well integrated with the nearby controls. In terms of a design solution, Vicente et al. emphasise that this arrangement "obeys human factors design principles" (p.32).

The investigators also observed a number of tailoring activities that were similar to those at the hard-wired plant. Operators changed alarm setpoints across contexts, and wrote values, cues and reminders on post-its or other media. One operator even kept a nuisance alarm up on the screen as a reminder to call maintenance to fix the problem!

Motivation: Moving From the Field to the Laboratory

Again at this point, now before moving on to discuss the motivation for this present research, it is useful to summarise what has been learned so far:

- Henderson and Kyng's (1991) studies of software design introduced the distinction between use and tailoring, and then built a classification of users' tailoring activities along the axis of distance from the software designers' intentions. In this classification, tailoring can range from changes to parameters that were well within the designers' intent, to additions to the

content of the interface that go beyond the designers' intent.

- Cook and Woods' (1996) field study of the impact of computer-based devices on anaesthesiologists' work introduced the distinction between task tailoring and system tailoring. An important form of system tailoring done by the anaesthesiologists was the creation of overview displays that reduced the amount of interface management necessary to perform their tasks.
- Vicente and Burns' (1995) and Vicente et al.'s (1997) field studies of operator monitoring in nuclear power plant control rooms demonstrated the variety of operator modifications that were made to both hard-wired and computer-based interfaces. These modifications helped the operators in their task of monitoring both by helping to offload their memory and by creating new information on the display. The large number of modifications observed demonstrates that operators at these plants were actively involved in efforts to finish the design.

Each of these studies is very important to an understanding of the nature and impact of operator modifications. Surprisingly, however, most of these studies do not present any controlled laboratory experiments in support of their findings. The studies that do present experimental evidence (i.e. Hammond et al., 1995; Kirsh, 1995) limit their scope to the consideration of what everyday experience, and do not focus on expert behaviour in complex sociotechnical systems. Kirsh (1995) only has experimental evidence based on observation of participants playing Tetris, and Hammond et al. (1995) built and tested a simulation. Obviously, there is a gap in the literature on operator modifications.

The most likely reason for this gap in the literature is the real difficulty in trying to conduct controlled experiments on operator modifications. First of all, operator modifications happen within a rich social environment, and it is likely that social interactions are foundational in the

creation of and agreement upon functional modifications (e.g. Hutchins, 1995). Second, operator modifications seem to be most frequent when operators are involved with complex tasks. Third, expertise seems to be the cornerstone of the making of modifications. Trying to create each of these conditions in the laboratory at the same time is surely a difficult task!

There is, however, another approach. The important contribution of the research reviewed above is that it has identified some of the relevant dimensions for tool use. The approach taken in this current research is to pare off the dimension of social interactions, to hold constant the dimension of expertise, and to vary the dimension of task complexity. It is hoped that this reduction will make the experimental problem more tractable so that it can provide conclusions relevant to the design of more complex and representative experiments. This move from the field to the laboratory is consistent with Vicente's (1997a) model of research (that has been derived from the work of Meister, Brunswik, and Gibson):

> Field studies can be used to observe behaviour in situ to identify phenomena that are worth studying under more controlled conditions. Laboratory studies can then be conducted under more controlled conditions in order to develop causal explanations for the observed phenomena. The generalisability of these causal explanations can then be tested under more representative conditions by conducting experiments that are more complex in nature. Finally, a theory or design intervention can then be evaluated in high-fidelity simulators or in the field in the presence of a wide range of factors that had been controlled for or eliminated in the laboratory…
>
> (Vicente, 1997a, p. 326)

The research presented here follows on from the field studies that have been documented, and so is a second step down this road. It should be stressed that this step is a preliminary and exploratory effort into understanding the phenomenon of operator tool use more thoroughly. At the outset, the experimenter did not know with certainty how to create the conditions necessary to bootstrap operator modifications. Accordingly, what follows is the description of a pilot experiment that was carried out to get a more thorough understanding of operator modifications in a laboratory context. The lessons learned from that effort were applied to the design of a

larger, more thorough experiment that is also documented in turn.

Since these experiments were exploratory, they were not guided by a well-defined theory or set of hypotheses. However, they were designed to answer a number of questions about operator modifications that were not addressed in the research reviewed. These are:

- Why do operators modify their interfaces?
- How do operator modifications develop over time?
- Do operator modifications affect task performance?
- Do operator modifications affect understanding?

In short, they are questions about the reasons for, and the performance implications and temporal dynamics of operator modifications. The remainder of this thesis documents research conducted in an effort to answer them.

Research Context

This study was conducted in the context of a larger programme of research that has been investigating the effects of ecological interface design (EID, Vicente & Rasmussen, 1992) on operator adaptation to complex work domains. Much of this research has been performed using DURESS II, an interactive thermal-hydraulic process control microworld driven by a simplified yet representative real-time computer model. Since the DURESS II microworld and its three interfaces have been described elsewhere (Christoffersen, Hunter, & Vicente, 1996; Janzen & Vicente, 1998), readers unfamiliar with DURESS II and its three interfaces should consult these references before continuing.

PILOT STUDY

Purpose

Since this research deals with a novel experimental problem, the first step taken in trying to answer the motivating questions (above) was to conduct a pilot study. The purpose of this pilot study was to gain an understanding of how to create the right conditions to bootstrap operator modifications in the laboratory, while at the same time providing some preliminary clues in the search for answers to the four questions motivating this research.

Experimental Design

For this preliminary investigation of tool use, a repeated measures, single factor experiment with two interface conditions (P and P+F) was used. Members from each pair of participants were randomly assigned to one of the two interface conditions, and participated for 87 trials each (approximately four weeks).

Participants

Previous work (Torenvliet, Jamieson, & Vicente, in press) has shown that the holist/serialist cognitive style distinction as measured by the Spy Ring History Test (SRT, Pask & Scott, 1972) is correlated with performance on the various interfaces of the DURESS II microworld. Accordingly, the SRT was administered to a pool of ten volunteer candidates from second to fourth year mechanical and industrial engineering classes (a copy of the SRT can be found in Appendix G). The six most closely matched candidates[2] were selected to participate in the experiment. Members of each pair of participants were randomly assigned to one of the two interface conditions. Each participant completed a demographic questionnaire to determine their

[2] Participant matching was done using a minimum distance algorithm. The SRT outputs three scores corresponding to an individual's holist and serialist cognitive style and overall cognitive ability. Distance scores were generated for each possible pairing of the ten candidates by summing the absolute differences between their scores on each of the three SRT components. The three unique pairs with the lowest total distance constituted the overall optimal solution, and were thus selected for the experiment.

Table 1. Summary of participant groupings.

		Demographics			*Education*			*SRT Data*		
Group	**Participant***	**Gender**	**Age**	**Year**	**Discipline**	**Thermo-dynamics†**	**Physics†**	**Holist (%)**	**Serialist (%)**	**Neutral (%)**
P	Britten	M	21	3	Mechanical	1	5	85.6	87.5	100.0
P+F	Bruckner	F	21	3	Industrial	0	3	68.9	82.5	92.9
P	Hammerstein	F	25	4	Industrial	0	1	66.7	67.5	50.0
P+F	Beethoven	F	20	2	Industrial	0	2	56.7	65.0	50.0
P	Respighi*	M	20	2	Industrial	0	2	82.2	97.5	35.7
P+F	Rossini**	M	28	2	Industrial	0	2	73.3	82.5	28.6
Means	P Group		22	3.0		0.3	2.7	78.1	84.2	61.9
	P+F Group		23	2.3		0	2.3	66.3	76.7	57.1

*Participant pairings are indicated by shading (e.g., Britten and Bruckner are cohorts).
†Number of half-year courses taken.
*Respighi did not complete the experiment, but dropped out after trial 7.
**Rossini did not complete the experiment, but dropped out after trial 16.

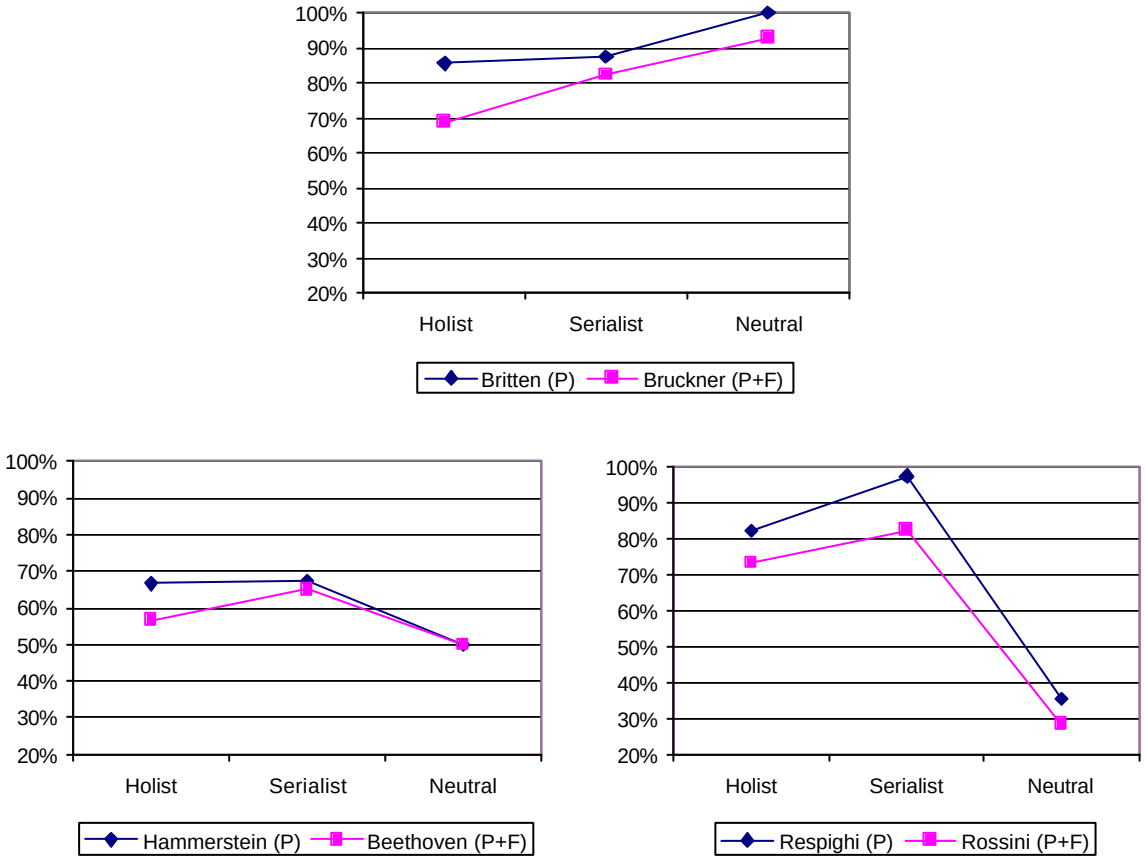


Figure 2. SRT data for participant pairs.

age, the number of courses they had taken in physics and thermodynamics, and their overall level of education. These data are summarised in Table 1, and SRT scores for each participant pair are graphed in Figure 2.

Across groups, there were no significant differences in serialist and neutral scores on the SRT as calculated by a paired *t*-test ($t_2 = 1.96$, n.s., and $t_2 = -2.00$, n.s., respectively), but there was a significant difference in holist scores across groups ($t_2 = 4.89$, $p < 0.04$) in favour of the P interface group. This difference was an unfortunate artefact of the random assignment procedure. However, given that performance on the P+F interface is positively correlated with participants' holist scores (Torenvliet et al., in press), this is a conservative error. Since it was only discovered post-hoc, participants could not be reassigned to different groups.

Participants were paid at the rate of $6 per hourly session, with an additional bonus of $2 per session for completing the experiment. An extra bonus of $2 per session was also offered for 'good' performance. This bonus was designed to maintain the participants' interest and motivation, and there were no criteria for awarding it. Of course, this was not told to the participants. In the end, all participants received this bonus.

Two participants dropped out of the experiment: Respighi quit after trial 7 and Rossini quit after trial 16. Both participants cited course workload as the reason for quitting. Perhaps coincidentally, the two were also poor performers who seemed to have difficulty controlling and understanding the simulation. Nonetheless, their participation was not a complete loss as some of the data collected from their first trials was still useful. Fortunately, the two were cohorts and so did not disturb the matching of the other participants.

Apparatus

The DURESS II simulation runs on a Silicon Graphics Iris Indigo R4000 computer

workstation. The simulation code was written in C, while the two interfaces were constructed using a graphical construction set called FORMS. Verbal protocols were collected using a Sony CCD-FX330 8mm Handycam with a Shure SM10A head-worn unidirectional microphone. Photographs were taken using a Ricoh RDC-4300 digital camera with a resolution of 1280 × 960 pixels. Data analysis was performed on a Pentium II class PC using SAS, SPSS, and Excel.

In addition, participants were given many different types of supplies that they could use to construct 'tools' for modifying their interfaces. In particular, they were provided with:

- post-it Notes
- post-it tape flags
- grease markers (to make non-permanent markings on the monitor)
- a notebook with both graph paper and lined paper
- a stopwatch
- a calculator
- pens and pencils of various colours
- an eraser
- a highlighter
- a magnifying glass
- scissors
- paper clips

Procedure[3]

Participants devoted one hour per weekday to the experiment over the four weeks of its duration. Before being assigned to groups, the participants read a description of the experimental

[3] This description follows the procedure description of Howie, Janzen, & Vicente (1996) closely.

procedure, completed a demographic questionnaire and the SRT, and signed a consent form. On the first day of the experiment, participants read a technical description of DURESS II, completed a short test to confirm their understanding, and set a schedule for the duration of the experiment with the experimenter. On the second day of the experiment, participants were introduced to the interface that they would be using for the experiment (P or P+F), and answered a number of questions to confirm their comprehension of the material. They were then asked to complete a control recipe (Irmer & Reason, 1991), a short set of instructions describing how to control DURESS II. The participants were requested to make these recipes detailed enough that a person unfamiliar with DURESS II would be able to operate it in the same way that they themselves did using their instructions alone. They were also introduced to the tools they had at their disposal, and were told that during the experiment they were to try and use the tools in any way that would make the task less difficult or their control more efficient. On the third day of the experiment, participants were introduced to the procedure for verbal protocols, and then commenced running trials with DURESS II.

The remainder of the experiment was divided into two phases. During phase one (67 trials), participants were encouraged to use the tools to make modifications to their interfaces. During phase two (20 trials), these tools were taken away from the participants. During both phases, the participants' task was to bring the system from a shutdown to a steady state, where steady state is defined as having the temperature and demand goals for both reservoirs satisfied for five consecutive minutes. Participants gave verbal protocols during all trials, both to habituate them with this practice and to prevent them from associating the verbal protocol trials with any other experimental event.

The simulation configuration files for phase one were the same as those used in two

previous research efforts (Howie et al., 1996; Hunter, Janzen, & Vicente, 1995). Randomizing the last twenty trials from phase one generated the sequence of trials followed by all participants in phase two. This was done to facilitate performance comparisons between the last twenty trials of phase one and the trials of phase two. Other than this, each trial had different steady state demand pairs to prevent the participants from adopting excessively simplified control methods.

Faults were distributed randomly and infrequently throughout the trials, to simulate the unanticipated character of faults in field settings. Two types of faults could occur. Routine faults were designed to be representative of recurring failures that occur in process control plants. These consisted of valve blockages, heater failures, and reservoir leaks. Non-routine faults were designed to simulate unfamiliar, unanticipated faults that operators would rarely encounter. These faults were implemented as two routine faults that could interact perversely. Non-routine faults in this investigation included a reservoir leak that interacted with a heater failure, one reservoir leaking into another, and a heater failure coupled with an increase in the input water temperature. There were seven routine and three non-routine faults in phase one, and three routine and two non-routine faults in phase two. The fault trials in phase two were repeats of the final five fault trials in phase one so that performance on fault trials with and without tool support could be investigated.

To elicit information on participants' understanding of the system, they were asked to complete four control recipes over the course of the experiment (after phase one trials 8, 33, and 67, and phase two trial 20) in addition to the one prepared before performing any trials. Some may argue that this caused participants to think about the system at a deep level more than they would otherwise. While this is true, the effect is the same for all participants, so these recipes may still be diagnostic of differences in system understanding.

Participants were not informed of the results of their trials, although they could receive feedback at any point during the trial by observing the elapsed time indicated on the simulation timer. They could also receive feedback from the status message displayed on the screen at the termination of all trials. This would inform them if they had reached steady state or had somehow violated the work domain constraints (i.e., "Reservoir 1 heated empty").

Finally, participants were interviewed after the completion of all trials and were asked for any comments they may have had about the experiment. They were also asked to explain the reason behind their tool uses.

A copy of the experimental schedule (including demand pairs and fault characteristics) is included in Appendix A. All experimental protocols can be found in Appendix G.

Data Sources

Four types of data were generated for each participant:

- Time stamped log files. During each trial, for each control action, DURESS II recorded the control action, the time, and the current values of all simulated variables. This information was compiled into one log file per participant per trial.
- Verbal protocols. Verbal protocols were recorded on videotape so that the participants' comments could be viewed within the context of the corresponding control actions and the overall state of DURESS II. Verbal protocols were used to understand both fault performance and participants' tool uses and their contexts.
- Control recipes. As outlined above, each participant completed five control recipes. These were compiled for later analysis.
- Artefacts. All tools created by participants were recorded. Physical artefacts (notebooks, post-it notes, etc.) were collected, and photographs were taken of representative non-

permanent modifications (i.e., on-screen modifications).

Results: Tool Use

Logs. Early into the experiment, it was recognized that participants were using tools less than was expected. While a few participants experimented with a variety of tools (such as the magnifier, stopwatch, and ruler) as they were waiting for the simulation to enter steady state, the majority of tool use was confined to writing in the notebooks (or, logs) that they had been given. A brief description of the participants' logs is given below:

- Britten (P): Britten only used his log for the first six trials of the experiment, and his notes were confined to the first half of the first page of his notebook (see Figure 3). On this page he made a table with four columns: 'Flow', 'Temp', 'Heater', and 'Volume in Tank ~'. The first six entries in this table were devoted to entries for reservoir 1 (i.e., Temp = 40°C) and the remaining entries were for reservoir 2 (i.e., Temp = 20°C). There is a fixed relationship between the heater setting and output flow for both reservoirs: for reservoir 1, the steady-state heater setting is equal to the desired output flowrate, and for reservoir two the steady-state heater setting is equal to the desired output flowrate divided by three. Britten used the information in his table to derive these ratios, and wrote, "Flow ≐ Heater" and "Flow / 3 ≐ Heater" beside the entries in his table devoted to reservoirs 1 and 2, respectively. After this, he did not make any more entries in his log.

  Two things are notable about this log. First, it seems to have acted as an external memory aid to help Britten compile the information necessary to derive the heater relationships quickly and efficiently. Second, Britten seems to have set out purposefully to derive these relationships. Since he wrote the first entry for reservoir 1 on line 1 and the first entry for reservoir 2 on line 6, it looks like he expected to make only a few entries and then use that

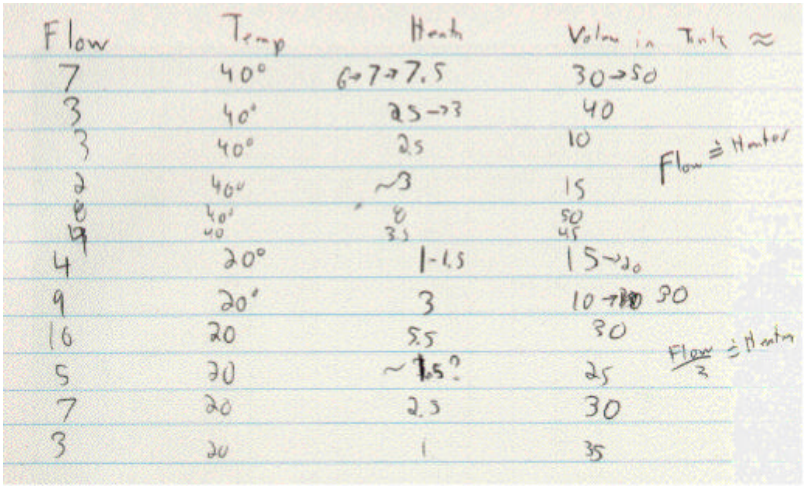


Figure 3. Britten's log.

information to derive any relationships driving the system (if he were planning on exploring, it is more likely that he would have written his entries sequentially, using lines 1, 3, 5, etc. for reservoir 1 and lines 2, 4, 6, etc. for reservoir 2). Since this participant had a deep knowledge of physics (he recorded that he had taken 5 courses in physics and 1 in thermodynamics) his knowledge that this relationship should exist and his intent to find it are not surprising.

- Hammerstein (P): Hammerstein used her log more than any other participant. In fact, midway through the experiment she found it necessary to add a bookmark to her log so that she could find her place more quickly. The information she recorded was remarkably consistent in form over the experiment: Beginning at trial 4, she recorded the valve settings required to achieve the desired output flows for each trial. Over the next number of trials she would record two lines per trial. In order, the first line indicated the settings for VO1, VA, VA1, and VA2, and the second line indicated the settings for VO2, VB, VB1, and VB2. There were a number of variations on this theme over the next 15 trials, but by trial 21 she settled into a stable pattern. At the beginning of every day she would estimate the number of trials that she would be finishing, and would then write down a template for recording the

valve settings information before beginning her trials (see Figure 4). She would then fill in the information specified by these templates at the beginning of each trial. Later, she began to write templates for many days ahead while she was waiting for DURESS II to enter steady state. By the end of phase one (trial 67), she had templates prepared up to trial 80.

Figure 4. One of the templates used by Hammerstein for recording her valve settings. Note that this is one of the templates that she prepared in advance (for trial 71) but was never able to use because her notebook was taken away at the beginning of phase 2.

As the organization of Hammerstein's notation makes apparent, she used these notes to derive the feedwater valve settings necessary to achieve the required output flows. Starting from the values for VO1 and VO2 (i.e., the flow goals), she would work out the flows that needed to be carried through the secondary and primary valves. This order of reasoning is made graphic by her entry for trial 5 (Figure 5). There she first wrote down the values for VO1 (2) and VO2 (16), and then used a set of sums to satisfy these demands.

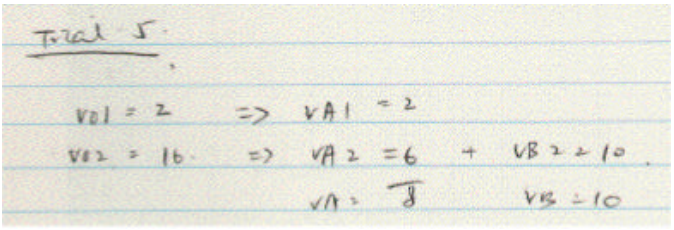


Figure 5. Hammerstein's log entry for trial 5, showing how she reasoned from the required output to the settings for the secondary valves, and from there to the settings for the primary valves. Note especially the line between VA2 and VA indicating that she was calculating a sum.

Hammerstein's log also included two other elements. First, each time that she encountered some problem with DURESS II, she noted her diagnosis of it (e.g., "Suspect that VB is a little bit blocked, [therefore] more water is provided for PA to Reservoir 2 than it should be.").

Second, starting with trial 11 she also noted her completion times beside the trial number in her notebook.

In sum, Hammerstein used her log mainly to derive the valve settings necessary to achieve the output demands, but also recorded her completion times as well as fault diagnoses. Over the long term, her log assumed the most stable form of all participants in this pilot investigation.

- Respighi (P): Respighi only participated in the experiment for 7 trials, so his log cannot be used to make any conclusions about long-term tool usage. However, in the space of the first seven trials, his log is noteworthy (Figure 6). He did not start using his log until the first fault trial (trial 5 – VA1 blockage) when he found that he did not understand how the simulation was functioning. Respighi's verbal protocol reveals that he did not even consider the possibility that this trial contained a fault. Instead, under the assumption that the simulation was operating normally, he tried to figure out why he had to add so much more water to R1 than to R2. In the end he concluded that much of the water in R1 vaporizes because it is at a much higher temperature than the water in R2, and that the extra water was needed to compensate for the loss of vapour. He recorded this process of reasoning as one long log entry. For his remaining two trials, he wrote down heater and valve settings in an effort to understand why the settings for these trials did not match those that his understanding of trial 5 would have predicted!
- Bruckner (P+F): Bruckner's habits in using her log had an interesting evolution. During the first 20 trials, her entries were procedural constraints that she thought were useful in operating DURESS II and that could be applied across trials (i.e., "open all valves before turning pumps on"; "cannot heat an empty reservoir"; "keep HTR slope & water input slope the same

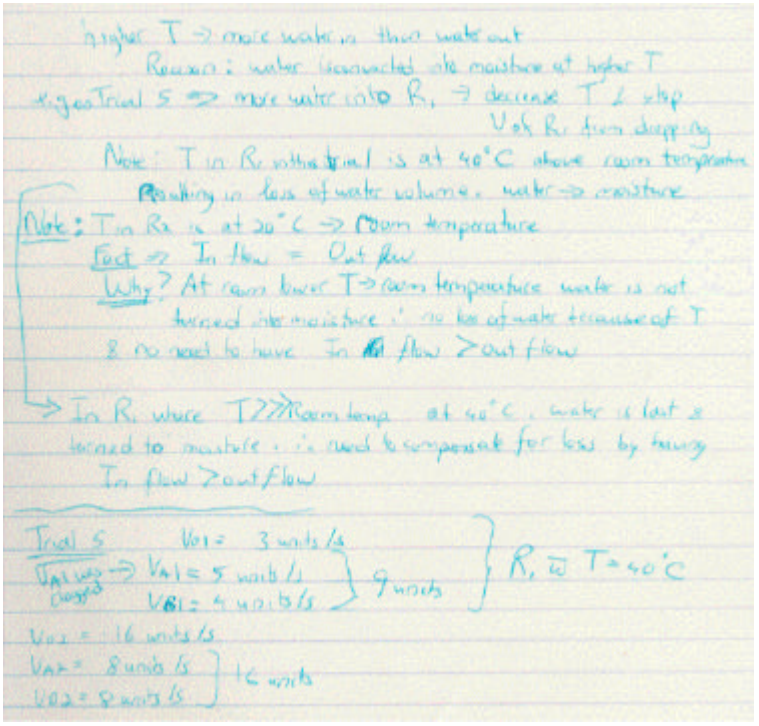


Figure 6. The first page of Respighi's log, showing his reasoning around the valve blockage that occurred in trial 5.

(regardless of whether it's steady, it'll balance out)" – see Figure 7a). Gradually, the nature of her log switched to being a report of abnormalities (Figure 7b). Finally, at trial 29 she began to bvadd a record of her completion times to the mix (Figure 7c), along with notes about her performance (i.e., "best time to date!"). In only once instance (trial 4) did she write down a list of component settings. In sum, Bruckner's log mainly discussed her experiences with DURESS II, including work-domain and procedural constraints. In addition, she also used her log to benchmark her performance.

- Beethoven (P+F): Beethoven used her log less than all other participants, and made only 5 separate entries over two trials (see Figure 8). The first entries were made after her fourth trial, and included one statement about the task goals, and three procedural constraints on her action that could be applied across trials. Her final entry was made after trial 8, and contains a reminder that during fault trials she should "detect error, diagnose, solve" as well as a hint to

"accumulate volume first – so that Reservoir doesn't dry up." Though Beethoven only made a small number of notes, it is striking that nowhere does she record any component settings, but instead seemed to focus on more procedural information.

- Rossini (P+F): Since Rossini only participated in 16 trials, like Respighi's log, his log cannot be used to make any conclusions about long-term tool use. Still, even in these early trials, Rossini used his log frequently. On the first day, he made quick scrawls about component settings. On the second day he adopted a neater, chart-based format that he used to keep track of the settings for all components as well as the steady-state reservoir volumes (Figure 9). He maintained this format for the duration of the time that he remained in the study.

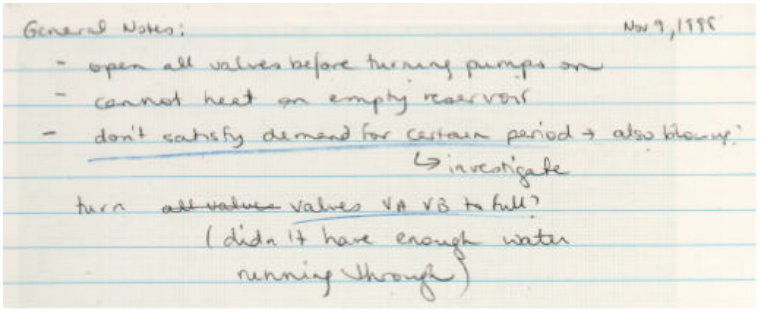


Figure 7a

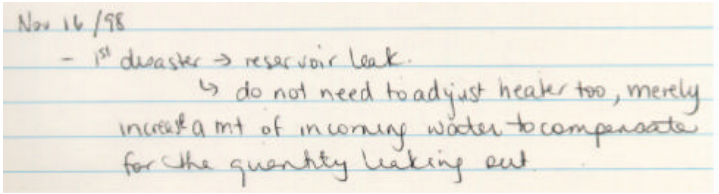


Figure 7b

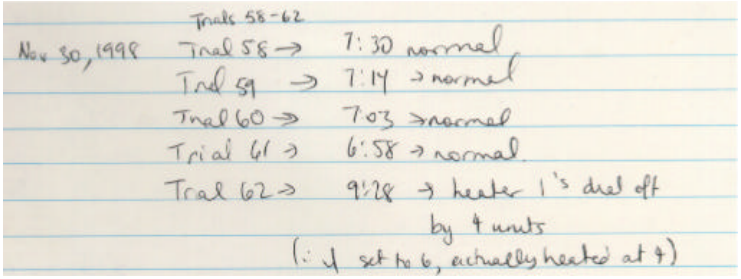


Figure 7c

Figure 7. The stages in Bruckner's log use. (a) Shows an example of the procedural constraints that she wrote early on. (b) Shows an example of one of her abnormality reports. Finally, (c) shows an example of how she recorded her trial times while continuing to report abnormalities.

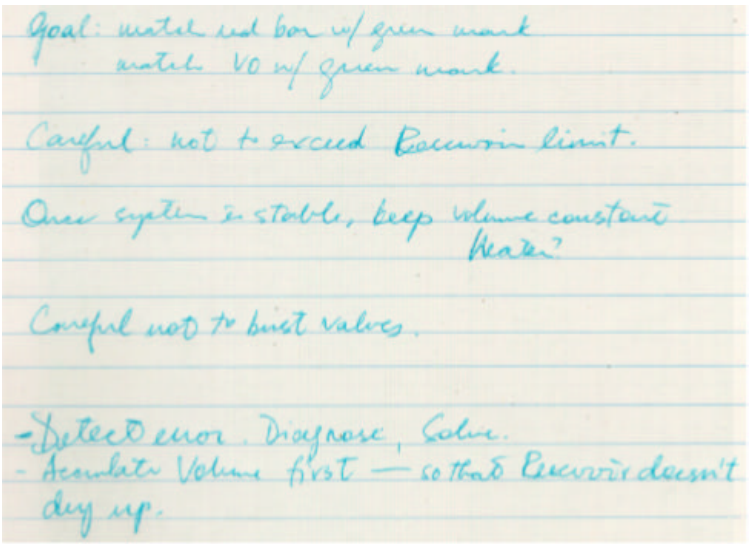


Figure 8. Entries from Beethoven's log.

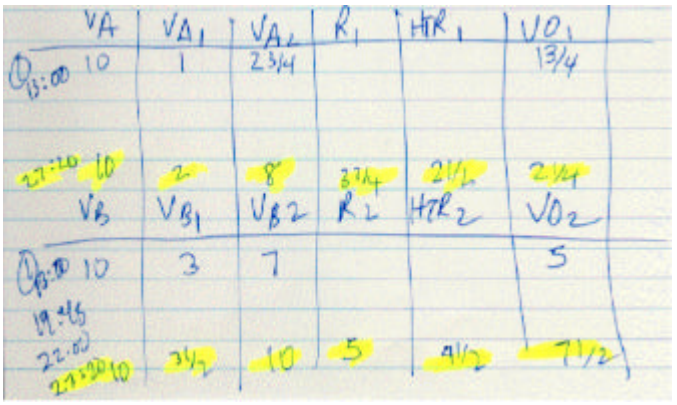


Figure 9. An example of the chart used by Rossini to keep track of component settings.

Summary. The general trend that can be seen in these logs is that participants using the P+F interface tended to jot down procedural constraints relevant to their operation of DURESS II that easily generalized across contexts, while participants using the P interface focused more on recording quantitative information about component settings that less easily afforded application across contexts. It is possible that the higher-order information contained in the P+F interface better afforded the observation and recording of procedural constraints while the lower-order information in the P interface limited participants to considering the quantitative settings that

they were using.

Rossini and Respighi are notable exceptions to these trends (recall that Respighi (P) recorded a great deal of reasoning in his log while Rossini (P+F) concentrated on setting information). Unfortunately, since these participants left the experiment early, it is difficult to make any statements about the form that their logs would have eventually taken.

An important preliminary conclusion can be made from these data. While it cannot be said that participants on either interface used their logs more than the others, the different types of information on the two interfaces seems to have affected the nature of the logs kept by the participants: the logs kept by participants on the P+F interface generally reflected the higher-order information contained in their interface, while the logs of the P participants reflected the relatively lower-order information contained in their interface. This finding indicates that the informational and visual content of an interface may affect the ways that operators use tools.

Other tool uses. Unfortunately, participants used few tools other than the logs with any frequency. The two tools that were used are the magnifier and the stopwatch. The magnifier was used early in the experiment, mostly by participants on the P interface (Britten and Hammerstein) to check the actual setting of the output valves in relation to the goal area. On the P+F interface, Bruckner never used the magnifier, and Beethoven only used it once (though this looked to be more of an effort to pass the time than for any strict purpose). Since the components for changing the output setting have similar visual forms on both interfaces, it is likely that this tool use was based on individual preferences. However, it is possible that since the P+F interface has a number of different indicators that are affected by output setting, the actual output setting is more obvious on this interface than on the P.

Bruckner was the only participant to use the stopwatch, and she used this tool for the

purpose of keeping track of the time elapsed since all goal variables had entered the goal region. Using the stopwatch, all she had to do was start the timer when the goal variables entered the goal region. After this, she could figure out the time remaining until steady state by subtracting the time displayed on the stopwatch from 5 minutes. Using the simulation timer, this task is much more complex. First, she would have to remember the time that the goal variables all entered the goal region. Second, to figure out the time remaining until steady state, she would have to subtract the remembered time from the actual time, and then subtract that result from 5 minutes. In short, her use of the stopwatch instead of the simulation timer offloaded a memory task from her head to the stopwatch, and also reduced the number of computations required by half.

One other instance of tool use deserves mention. During one non-routine fault trial that included a leak from reservoir 1 and an overheat of the heater for reservoir 1, Bruckner asked if there was a protractor available for her use. The logic behind this request is worth noting. Under normal steady-state conditions, the gradient lines for the mass and energy inventories would be perpendicular to the base of the reservoir. Under conditions of a reservoir leak, a portion of the output mass and energy is unaccounted for by the interface which would cause the mass and energy input to the reservoir to appear to be greater than the output at steady-state. Thus, the gradients would no longer be perpendicular to the base of the reservoir at steady-state. However, since the energy inventory is scaled relative to the mass inventory, given that the relationship between mass and energy is linear, the mass and energy gradients will be deflected by a constant amount at steady-state for a leaking reservoir. In other words, the mass and energy gradients would still be parallel to each other at steady state. Bruckner wanted to use a protractor to make sure this was the case. Since no protractor was available, this does not strictly qualify as a tool

use. However, it was a highly creative solution, indicating information that was not included in the interface but that was needed to address a specific context.

Summary. There was little tool use other than the logs. Use of the magnifier may point to a deficiency in the visual form of the P interface, and Bruckner's use of the stopwatch was a sensible offloading of her memory. Bruckner's request for a protractor, on the other hand, represents a creative solution that indicates the information needed to approach a unique context not considered by the system designers. In casual terms, the first two tool uses were simple and the third was much more complex.

These results also help in classifying the participants according to their tool uses. Overall, Britten and Beethoven tended not to use tools (their logs were brief, and they exhibited few other tool uses) while Bruckner and Hammerstein tended to use tools (their logs were more extensive, and Bruckner made much use of the stopwatch). Rossini and Respighi are not included in this classification as they dropped out of the experiment too early to allow them to be categorized on this basis. This classification is informal, but it is a useful distinction for considering the impact of tool use in the results that follow.

Results: Control Recipes

An analysis of participants' control recipes was performed to discover the impact of tool use on understanding. Analyses were performed on the length and content of these recipes.

Length. Table 2 shows the results of an analysis on the length of control recipes, both by amount of tool use and by interface. The most important insight from this analysis is while interface seems to have little effect on the length of control recipes, amount of tool use has a noticeable impact. Participants who tended to use tools wrote control recipes that were longer than those who tended not to use tools.

Table 2. Control recipe length (number of words).

	Tended not to use tools		*Tended to use tools*	
Trial	**Britten (P)**	**Beethoven (P+F)**	**Bruckner (P+F)**	**Hammerstein (P)**
0	323	201	169	325
8	175	211	280	171
33	189	173	214	244
67	182	211	382	256
87	177	185	260	319
Average, 8-87	181	195	284	248

Number of warning statements. There were also systematic differences in the content of the control recipes between those who tended to use tools and those who did not that can at least partially explain the differences noticed above. During the analysis, it was noticed that some participants tended to add words of warning that went beyond the purely procedural content at that base of all participants' control recipes. These comments were generally indicated by the text "Warning!" or "DO NOT…" and included advice about how to avoid the work domain constraints. The number of statements and total words per recipe dedicated to these statements are documented in Table 3.

Quite clearly, participants who tended not to use tools did not write any of these comments, while those who did use tools did make these comments. This could indicate a predisposition on

Table 3. Number of statements and total words in control recipes devoted to warnings of work domain constraints.

	Tended not to use tools				*Tended to use tools*			
Trial	**Britten (P)**		**Beethoven (P+F)**		**Bruckner (P+F)**		**Hammerstein (P)**	
	Statements	Words	Statements	Words	Statements	Words	Statements	Words
0	0	0	0	0	0	0	0	0
8	0	0	0	0	2	27	0	0
33	0	0	0	0	1	16	1	11
67	0	0	0	0	3	202	2	23
87	0	0	0	0	1	9	2	21
Average, 8-87	0	0	0	0	1.8	63.5	1.3	14.1

the part of those who tended to use tools to engage in self-explanation about DURESS II (Howie & Vicente, 1998). Notice, however, that this does not entirely explain the differences in length between the two sets of control recipes. Excluding these comments, participants who tended to use tools still tended to write longer control recipes.

Diagrams. Also noticed in an analysis of these control recipes was the inclusion of diagrams and other graphical features. Table 4 summarises the number of diagrams observed in participants' control recipes. The data in this table indicate that participants who tended to use tools also tended to include diagrams in their control recipes.

Table 4. Number of diagrams included in control recipes.

	Tended not to use tools		*Tended to use tools*	
Trial	**Britten (P)**	**Beethoven (P+F)**	**Bruckner (P+F)**	**Hammerstein (P)**
0	0	0	2	2
8	0	0	1	0
33	0	0	1	4
67	0	0	0	3
87	0	0	1	4

Correlation to SRT scores. Finally, it is possible that the differences between participants on these measures of tool use stem from differences between participants as measured by the SRT. Unfortunately, there was no correlation between the tool use measures and SRT scores. The two cohort pairs (Britten/Bruckner and Beethoven/Hammerstein) were well matched on the basis of SRT scores, but yet exhibited many differences in their control recipes.

Summary. The results from these analyses of participants' control recipes present an unclear picture. Although the control recipes of participants who tended to use tools were longer than those of the others, additional length is not necessarily an indication of deeper understanding. In fact, Christoffersen, Hunter, and Vicente (1998) have made the observation that while longer control recipes might be indicative of a greater amount of knowledge, shorter

control recipes are indicative of knowledge that has been tied together under some suitable abstraction. However, given the small sample size here, statements about the relative quality or quantity of knowledge between participants cannot be made.

Part of the difference in length can be attributed to the inclusion in the control recipes of warnings about the work domain constraints. Since Hammerstein and Bruckner discussed these constraints in their notebooks (in direct terms, but also in the context of faults), this may have triggered their remembrance of these constraints in their control recipes. It is also possible that participants who tended to use tools were predisposed towards a self-explanation of their actions (e.g. Howie & Vicente, 1998). Unfortunately (as will be seen below) this did not mean that these participants were better at avoiding the work domain constraints. Both Hammerstein and Bruckner violated the work-domain constraints and so blew up DURESS II later in the experiment than did Britten and Beethoven.

The results on diagrams are a little more difficult to understand. Although it is tempting to attribute these diagrams to a deeper familiarity with DURESS II that may have been caused by their use of tools, it is equally (if not more) likely that this result is due to some differences in spatial ability between participants. Given that both Hammerstein and Bruckner included diagrams in their first (pre-trial) control recipe, this is a more likely explanation.

In sum, these results indicate that participants who tend to use tools might have a greater quantity of knowledge about DURESS II, but that this does not necessarily represent a deeper knowledge. They also indicate that participants may tend to internalize the types of information that they include in their logbooks, and thus reproduced that information on their control recipes. Alternatively, this could reflect a predisposition on the part of those who used tools to engage in self-explanation.

Results: Normal Trial Completion Time

The results of an analysis of participants' trial completion times are shown below. These results have been divided into three groups of trials to gain an accurate picture of post-transfer performance. The first group, trials 1-47, represent the learning portion of the experiment. The second group, trials 48-67, represent quasi-expert performance with tools, while trials 68-87 represent quasi-expert performance with no tools (recall that trials 68-87 used the same demand pairs as trials 48-67, randomized). Table 5 presents the means and standard deviations for these data. Figure 10 represents these data graphically, with 95% confidence intervals about the means. Figure 11 presents participants' individual practice curves, an unaggregated view of the data.

A number of trends identified in previous research (i.e. Christoffersen et al., 1996; Hunter et al., 1995) were also noticed here. That is, while the performance of participants using the P+F interface was only slightly faster than those on the P interface, it was generally less variable. A Cochran test for homogeneity of variance (Kirk, 1995) revealed that the variances in trial completion time between interface groups for trials 1-47 were homogeneous ($C = .368 < C_{.05}* = .372$, $p < .05$), but that they were not homogeneous for trials 48-67 ($C = .592 > C_{.05}* = .437$, n.s.) and 68-87 ($C = .495 > C_{.05}* = .437$, n.s.). Figure 10 and Table 5 illustrate these results further, and indicate that participants on the P+F interface had less variable trial completion times in trials 48-67 and 68-87 than participants on the P interface.

Since these later variances were not homogeneous (even for this low sample size), it can be concluded that long-term performance on the P+F interface was less variable than on the P interface, both with tools and without. To confirm that trial completion times on both interfaces

were similar, a general linear model (GLM)[4] was constructed on trial completion times for trials 1-47, and is reproduced in Table 6. As expected, there is no interface effect for trial completion time, although there is a significant trial (i.e., learning) effect. Since the variances in trial completion times for trials 48-67 and 68-87 were not homogeneous, GLMs could not be constructed for these data. Nonetheless, Figure 10 indicates that a similar result holds true for them as well.

Table 5. Trial completion times by participant.

	P Interface				*P+F Interface*			
	Britten		**Hammerstein**		**Bruckner**		**Beethoven**	
Trials	$\bar{x}$ (s)	s^2 (s^2)	$\bar{x}$ (s)	s^2 (s^2)	$\bar{x}$ (s)	s^2 (s^2)	$\bar{x}$ (s)	s^2 (s^2)
1-47	579	264	900	380	500	146	627	241
48-67	447	52	550	189	434	25	481	53
68-87	429	61	478	148	423	22	459	47

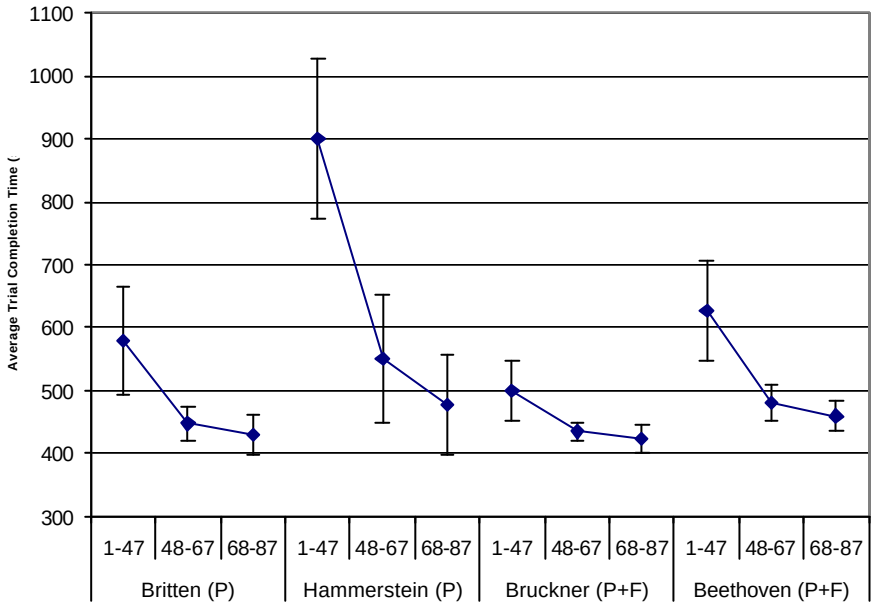


Figure 10. Trial completion time by participant.

[4] In what follows, analyses of variance (ANOVA) were used for balanced sets of data, and general linear models (GLMs) were used for unbalanced data sets. The results of both of these analyses are interpreted in the same way, but GLMs are more robust for designs with missing data.

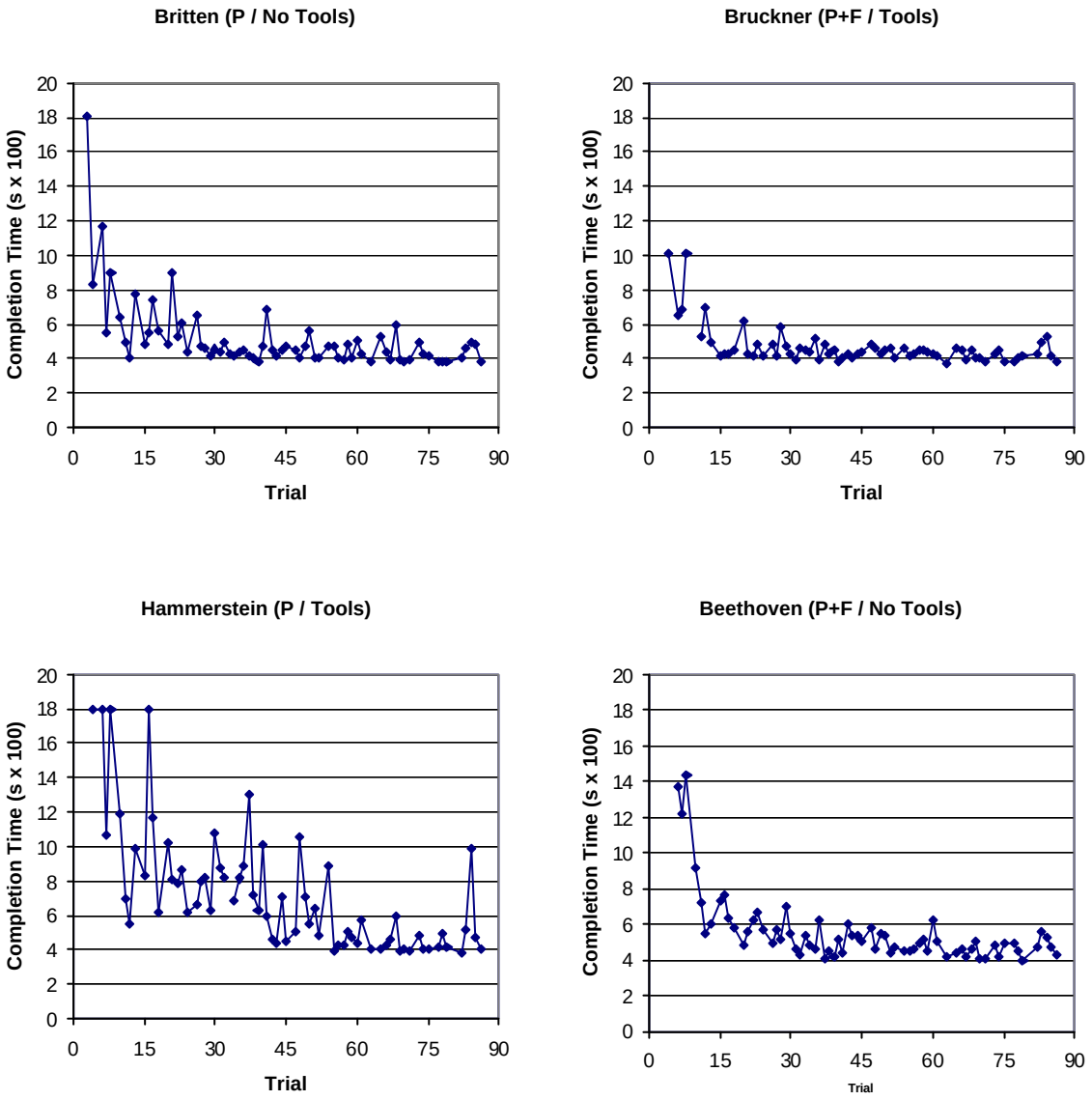


Figure 11. Individual practice curves.

Note also that performance for all participants improved at about the same rate in phase two (trials 68-87) for all participants. Since Bruckner and Hammerstein's performance did not degrade drastically in this period, their tool use was not a cognitive crutch. Since their performance did not improve at a greater rate than the other participants, tool use also did not hold back their performance.

Finally, note that there is no correlation between tool use and trial completion time. Among those who tended to use tools, Bruckner had the fastest and least variable completion times overall while Hammerstein had the slowest and most variable completion times. Britten and Beethoven, who both tended not to use tools, had performance in between these two extremes.

Table 6. GLM on trial completion time for trials 1-47.

Source	DF	SS	MS	F	p
Interface	1	888453.757	888453.757	0.67	0.500
Participant(Interface)	2	2669829.232	1334914.616		
Trial	38	7240302.007	190534.263	6.22	< 0.001
Interface × Trial	37	929472.559	25391.150	0.83	0.730
Participant × Trial (Interface)	68	2084477.661	30654.083		

Results: Normal Trial Blowups

Still on the topic of normal performance, a final performance metric is the number of blowups. DURESS II has a number of constraints that, if violated, cause the simulation to terminate abnormally (or, blowup). Table 7 shows the results of an analysis of on the number of these blowups per participant. Given the small sample size, it is difficult to make any conclusions from these data. It is worth noting, however, that Bruckner and Hammerstein each experienced a blowup later into their trials than other participants. These participants still blew up the simulation at trial 10 and 33, respectively, while Britten and Beethoven experienced their last blowups at trial 2 and 4, respectively.

Table 7. Number of normal trial blowups.

	Tended not to use tools		*Tended to use tools*	
Trials	**Britten (P)**	**Beethoven (P+F)**	**Bruckner (P+F)**	**Hammerstein (P)**
1-47	1	4	4	4
48-67	0	0	0	0
68-87	0	0	0	0

Results: Fault trials

Finally, fault data are presented below on three measures: detection, diagnosis, and compensation. At the beginning of the experiment, all participants were told that in the event that they encountered a fault, their task was to verbalise their detection and diagnosis, and then to compensate for the fault and reach steady-state. Accordingly, the data for detection and diagnosis comes from participants' verbal protocols. Participants' level of diagnosis was scored based on their level of accuracy, and were categorized into four levels (Pawlak & Vicente, 1996):

0 - the participant says nothing relevant to the fault or nothing at all

1 - the participant states that the system is not behaving as expected and describes the symptoms of the fault at a general level (i.e., "The level of reservoir 1 is dropping.")

2 - the participant describes fault symptoms at a more functional level, but still fails to localize the fault (i.e., "I'm losing flow into reservoir one somehow.")

3 - the participant correctly localizes the root cause of the fault (i.e., "VA1 is blocked.")

Using this scheme, higher diagnosis scores indicate a higher understanding of the functioning of DURESS II.

Compensation time is the time taken to successfully complete a fault trial. Compensation times were counted regardless of whether or not a fault was detected.

Table 8 presents a summary of participants' average fault detection times and number of faults detected, Table 9 summarises participants' average time to reach an accurate diagnosis of the fault and the number of faults diagnosed, Table 10 details the average highest detection scores reached, and Table 11 presents the average fault compensation times and number of faults compensated for.

Table 8. Average fault detection time and number of faults detected.

	Tended not to use tools				*Tended to use tools*			
	Britten (P)		**Beethoven (P+F)**		**Bruckner (P+ F)**		**Hammerstein (P)**	
Trials	Time	Number	Time	Number	Time	Number	Time	Number
1-47	0:43	5/7	1:19	6/7	0:38	5/7	0:50	5/7
48-67	0:42	5/6	0:49	6/6	0:31	6/6	1:26	6/6
68-87	0:58	6/6	0:50	5/6	0:14	6/6	2:26	6/6

Table 9. Average fault diagnosis time and number of faults diagnosed.

	Tended not to use tools				*Tended to use tools*			
	Britten (P)		**Beethoven (P+F)**		**Bruckner (P+F)**		**Hammerstein (P)**	
Trials	Time	Number	Time	Number	Time	Number	Time	Number
1-47	2:26	1/7	2:44	4/7	0:51	4/7	—	0/7
48-67	3:16	2/6	2:18	3/6	3:04	5/6	4:58	2/6
68-87	2:30	2/6	1:13	3/6	0:44	6/6	4:08	1/6

Table 10. Average highest diagnosis score reached.

	Tended not to use tools		*Tended to use tools*	
Trials	**Britten (P)**	**Beethoven (P+F)**	**Bruckner (P+F)**	**Hammerstein (P)**
1-47	1.0	0.7	2.0	2.3
48-67	1.5	1.7	2.8	2.5
68-87	1.7	1.5	3.0	1.5

Table 11. Average fault compensation time and number of faults trials completed successfully.

	Tended not to use tools				*Tended to use tools*			
	Britten (P)		**Beethoven (P+F)**		**Bruckner (P+F)**		**Hammerstein (P)**	
Trials	Time	Number	Time	Number	Time	Number	Time	Number
1-47	15:29	4/6	16:03	5/6	10:16	4/6	21:48	3/6
48-67	14:17	4/4	12:58	3/4	12:15	3/4	19:32	4/4
68-87	11:27	3/4	11:33	4/4	9:49	3/4	12:52	4/4

In each of these tables, a similar pattern can be noticed. Bruckner was generally the best performer on faults, Beethoven and Britten followed next and were generally quite closely matched, and Hammerstein performed most poorly. The only measure on which an interface effect is clear is average highest diagnosis score, on which the P+F group performed markedly better than the P group in trials 1-47 and 48-67. This effect was less pronounced in trials 68-87, but it is still apparent. In sum, there does not seem to be a correlation between tool use and

performance on fault trials.

Discussion: Implications for a more thorough study

Despite the small sample size of this pilot study, it has presented a number of trends that are worth investigating further. First, an analysis of the contents of participants' logs gave a clear indication that interface content affects the type of information recorded. This hints that operator modifications might be sensitive not just to the underlying task, but also to the interface through which that task is to be performed. Further, these logs were used for tasks including remembering (e.g., Britten's base data for calculating the heater-output ratios), reinforcement (e.g., Hammerstein and Bruckner's recording of procedural hints and system events), and derivation (e.g., Britten's derivation of the heater-output ratios, and Hammerstein's derivation of the feedwater valve settings). Second, the analysis of control recipes demonstrated that the participants who tended to use tools regarded a discussion of work-domain constraints important enough to add to their control recipes. While this did not have an impact on their performance, it may indicate that items recorded in logs (or, addressed by tools) achieve importance in the mind of a participant. This analysis also revealed that participants who tend to use tools might assemble a greater amount of knowledge about a system, but that that knowledge is not necessarily abstracted to reveal deeper constructs. Third, the analysis of performance under both normal and fault conditions did not reveal a noticeable effect of tool use (although all of the expected interface effects were present).

It should be stressed that the above results are based on an extremely small sample size, and so are preliminary at best. Nonetheless, they do make a contribution toward both answering the motivating questions and designing a more sensitive experiment that might elicit more tool use.

Preliminary answers to motivating questions. These results make a few valuable first steps toward answering the questions motivating this study. First, even in the context of this small study, the effect of interface on tool use was pronounced. Unfortunately, this finding is only in the context of one type of tool use, and it remains to be seen how this would generalize across different types of tool uses. Second, the issue of stability was partially addressed. While all participants tested out the tools at the start of the experiment, once a pattern — either of tool use or no tool use — was established, it was not broken. After a brief exploratory phase, tool use was stable in the long-term. Third, these findings indicate that tool use does not affect performance. Moreover, tool use is neither a cognitive crutch, nor does it impair performance. Finally, this study seems to indicate that people who use tools might also be more reflective of their operating knowledge, but that this does not necessarily translate into performance enhancements. It is likely that this tendency for self-explanation co-exists with tool use, but is not caused by tool use.

Two major issues stood in the way of posing stronger answers to these questions. First is the issue of sample size. This limitation was expected, as this study was deliberately done on a small scale. Overcoming this limitation does take some effort, but is conceptually easy to do. Second is the more basic issue of tool use: how are the conditions necessary to promote a greater amount and variety of tool use to be created? In the next section a larger experiment, designed to overcome these problems, is described.

EXPERIMENT

Purpose

The pilot experiment described above provided at least tentative answers to the four questions motivating this research, but it did not generate a great deal of tool use. While this could have been a result of the individual participants used in the experiment (after all, it was a small sample), it could be that the conditions necessary to elicit tool use were not created. This section describes a larger experiment that was designed to elicit more tool use. A manipulation was designed in an attempt to induce a greater variety of tool use, and a larger sample was used in an aim to bolster statistical power. The larger purpose of this experiment was to provide a set of more thorough answers to the four questions motivating this research than were provided by the pilot study.

Experimental Design

To create an experimental manipulation that would do a better job of inducing a larger amount and variety of tool use, the results of the literature review were revisited. Of particular interest is the work of Cook and Woods (1996) who observed operator modifications in the context of an interface transfer. Many of these modifications were directed at correcting interface deficiencies of the new interface in relation to the old. An experimental manipulation was designed in an attempt to replicate these conditions in the laboratory. Recall that DURESS II has three interfaces (see Appendices A and B):

- The P+F interface has been designed using the principles of ecological interface design, and is rich in relevant information to support operators in both normal and abnormal situations.
- The P interface is representative of the class of interfaces that are based on piping and instrumentation diagrams, and provides only a subset of the information available on the P+F interface.

- The divided P+F interface provides the same information as the P+F interface, but divides this information over four windows that are only available one at a time in a serial fashion. This introduces the task of interface navigation and also requires operators to remember critical system parameters while navigating between windows.

The approach taken in this experiment was to give participants about 7 hours of experience (23 trials) with the P+F interface, and then to transfer them to either the P or the divided P+F interface for an additional 7 hours (29 trials). The interface transfer was designed to reduce the amount of task support available in the interface. Those transferring to the P interface would lose the higher-order functional information provided by the P+F interface, while those transferring to the divided P+F interface would keep that information while gaining the tasks of interface navigation and information integration across windows. The trials performed on the P+F interface were designed to establish a referent for participants' tool use so that the reasons for any differential tool uses that occurred across interfaces in the post-transfer phase could be inferred. Further, the experience gained by all participants both before and after the interface transfer will provide a venue for observing the evolution of their tool use. Finally, faults were distributed randomly and infrequently throughout the experiment so that any correlations between tool use and performance could be observed in both normal and fault conditions.

This design was inspired by the work of Giraudo and Pailhous (in press) who used a similar type of manipulation in their studies on human memory. They characterized this type of manipulation as a perturbation of the system, and were interested in observing the types of behaviour that developed as a result of being perturbed. It is instructive to view the current experiment from this perspective as well. Over the first introductory trials with the P+F interface, participants will develop strategies based on the information contained in the P+F

interface. After the system is perturbed by changing the participants' interfaces, it will be interesting to see how they make modifications and use tools to cope with this perturbation.

To sum up the above discussion in the language of experimental design, this study used a repeated measures, single factor experiment with two interface conditions (P and divided P+F). All participants completed an initial 23 trials on the P+F interface, and then were transferred to either the P or the divided interface, on which they completed an additional 29 trials. Each participant participated for 16 days, 2 days of which were devoted to pre-experimental activities.

Participants

Just as in the pilot study, participants were selected on the basis of their scores on the SRT. The SRT was administered to a pool of twenty volunteer candidates from second to fourth year mechanical and industrial engineering classes, and the twelve most closely matched candidates were selected to participate in the experiment. Members of each pair were randomly assigned to one of the two interface conditions. Each participant completed a demographic questionnaire to determine their age, the number of courses they had taken in thermodynamics and physics, and their overall level of education.

Because it was impractical to test all of the participants in parallel, participants completed their trials in two groups of six participants each. Górecki (divided), Prokofiev (divided), Rachmaninov (P), Schubert (divided), Telemann (P), and Willan (P) participated in the first group, and Bach (divided), Bartók (P), Boccherini (divided), Mozart (P), Schoenberg (divided) and Wagner (P) participated in the second group.

These data are summarised in Table 12 and SRT scores for each participant pair are graphed in Figure 12.

Table 12. Summary of participant groupings.

	Demographics			*Education*				*SRT Data*		
Group	**Participant***	**Gender**	**Age**	**Year**	**Discipline**	**Thermo-dynamics†**	**Physics†**	**Holist (%)**	**Serialist (%)**	**Neutral (%)**
P	Willan[1]	M	21	2	Mechanical	0	2	85.7	80.0	97.5
Divided	Bach[2]	F	22	3	Industrial	0	3	100.0	91.1	100.0
P	Telemann[1]	F	23	4	Industrial	0	2	71.4	61.1	77.5
Divided	Boccherini[2]	F	19	2	Mechanical	0	3	71.4	50.0	90.0
P	Bartók[2]	M	20	2	Industrial	0	2	57.0	41.0	45.0
Divided	Schoenberg[2]	M	20	2	Mechanical	0	4	64.2	42.2	55.0
P	Wagner[2]	M	21	2	Industrial	0	3	78.6	71.1	72.5
Divided	Górecki[1]	M	20	2	Mechanical	0	3	71.4	72.2	80.0
P	Mozart[2]	M	19	2	Mechanical	0	4	78.5	95.5	95.0
Divided	Schubert[1]	M	20	2	Industrial	0	3	78.6	91.1	97.5
P	Rachmaninov[1]	M	19	2	Mechanical	0	3	57.1	65.6	45.0
Divided	Prokofiev[1]	F	21	2	Mechanical	0	3	42.9	60.0	65.0
Means	P Group		20.5	2.3		0	2.7	71.4	69.1	72.1
	Divided Group		20.3	2.2		0	3.2	71.4	67.8	81.3

*Participant pairings are indicated by shading (e.g., Willan and Bach are cohorts)
†Number of half-year courses taken
[1] Participated in the first group / [2] Participated in second group

Across groups, there were no significant differences in holist and serialist scores on the SRT as calculated by a paired *t*-test ($t_5 = .54$, n.s., and $t_5 = .46$, respectively). There was a significant difference in neutral SRT scores across groups ($t_5 = 3.38$, $p < 0.02$). This difference was an artefact of the random assignment procedure, and was not deemed to be critical as neutral SRT scores have not been shown to be correlated with performance on either the P or P+F interface of DURESS II (Torenvliet et al., in press).

As in the pilot study, participants were paid at the rate of $6 per hourly session, with an additional bonus of $2 per session for completing the experiment. A motivational bonus of $2 per session for 'good' performance was again also offered.

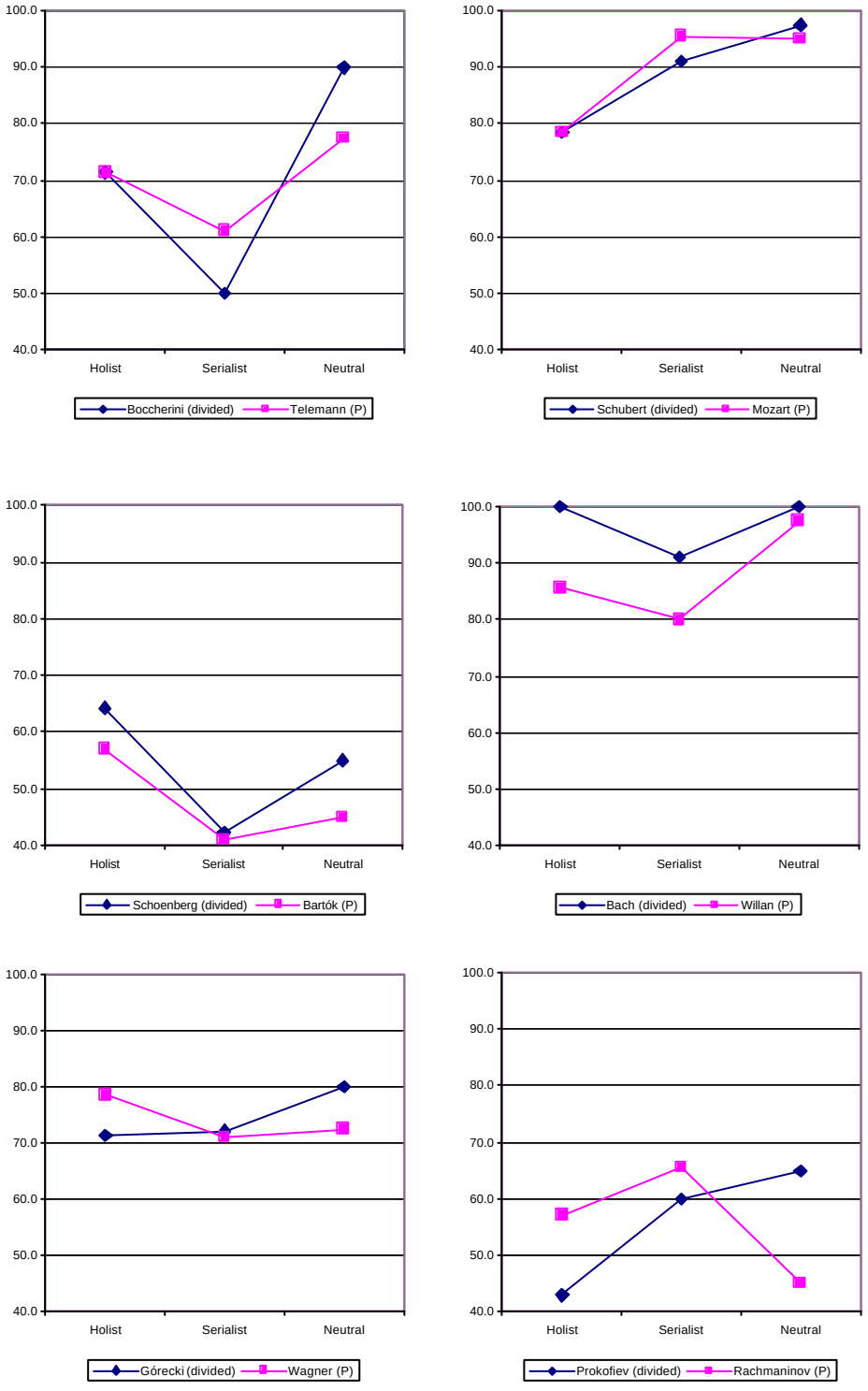


Figure 12. SRT data for participant pairs.

Apparatus

The apparatus for this experiment was the same as that used in the pilot study. Participants were given the same array of tools, with the addition of a protractor (as prompted by Bruckner's request for this tool during the pilot study).

Procedure

Participants devoted one hour per weekday to the experiment over the three weeks of its duration. Before being assigned to groups, the participants read a description of the experimental procedure, completed a demographic questionnaire and the SRT, and signed a consent form. On the first day of the experiment, participants read a technical description of DURESS II, completed a short test to confirm their understanding, and set a schedule for the duration of the experiment with the experimenter. On the second day of the experiment, participants were introduced to the P+F interface, and answered a number of questions to confirm their understanding. They were then asked to write a control recipe that would describe to a novice user how to operate DURESS II. They were also introduced to the tools they had at their disposal, and were told that during the experiment they were to try and use the tools in any way that would make the task less difficult or their control more efficient. On the third day of the experiment, participants were introduced to the procedure for verbal protocols, and then commenced running trials with DURESS II.

For the first 23 trials of the experiment, participants operated DURESS II using the P+F interface. At the beginning of trial 24 they were introduced to the alternate interface that they would be using (either the P or the divided interface), and then used this interface until the end of the experiment (trial 52). Each trial had different steady state demand pairs to prevent participants from adopting excessively simple control methods. During both phases, the

participants' task was to bring DURESS II from a shutdown to a steady state, where steady state is defined as having the temperature and demand goals for both reservoirs satisfied for five consecutive minutes. Participants gave verbal protocols during all trials, both to habituate them with this practice and to prevent them from associating the verbal protocol trials with any other event.

Faults were distributed randomly and infrequently throughout the trials, to simulate the unanticipated character of faults in field settings. Two types of faults could occur. Routine faults were designed to be representative of recurring failures that occur in process control plants. These consisted of valve blockages, heater failures, and reservoir leaks. Non-routine faults were designed to simulate unfamiliar, unanticipated faults that operators would rarely encounter. These faults were implemented as two routine faults that could interact perversely. Non-routine faults in this investigation included one reservoir leaking into another and a heat failure coupled with an increase in the input water temperature. There were two routine faults and one non-routine fault in phase one, and three routine and one non-routine fault in phase two.

To elicit information on participants' understanding of DURESS II, they were asked to complete three control recipes over the course of the experiment (after trials 12, 34, and 52) in addition to the one prepared before performing any trials.

Participants were not informed of the results of their trials, although they could receive feedback at any point during the trial by observing the elapsed time indicated on the simulation timer. They could also receive feedback from the status message displayed on the screen at the termination of all trials. This would inform them if they had reached steady state or had somehow violated the work domain constraints (i.e., "Reservoir 1 heated empty").

Finally, participants were interviewed after the completion of all trials and were asked for

any comments they may have had about the experiment. They were also asked to explain the reason behind their tool uses.

A copy of the experimental schedule (including demand pairs and fault characteristics) is included in Appendix B. All experimental protocols are included in Appendix G.

Data Sources

Four types of data were generated for each participant:

- Time stamped log files. During each trial, for each control action, DURESS II recorded the control action, the time, and the current values of all simulator variables. This information was compiled into one log file per participant per trial.
- Verbal protocols. Verbal protocols were recorded on videotape so that the participants' comments could be viewed within the context of the corresponding control actions and the overall state of DURESS II. Verbal protocols were used to understand both fault performance and participants' tool uses and their contexts.
- Control recipes. As outlined above, each participant completed four control recipes. These were compiled for later analysis.
- Artefacts. All tools created by participants were recorded. Physical artefacts (notebooks, post-it notes, etc.) were collected, and photographs were taken of representative non-permanent modifications (i.e., on-screen modifications).

Results: Tool Use

Overall tool use. Many more instances of interface modifications and tool use were observed in this study than in the pilot study. To describe this activity, it was necessary to supplement the qualitative descriptions used in the pilot study with more quantitative measures

of tool use. While qualitative descriptions do a good job of capturing the different types of tool use, quantitative methods are needed to capture the development and long-term patterns of tool-use that were observed so that reliable comparisons can be made between participants and groups.

To express participants' tool use in quantitative terms, the tool uses observed over the course of the experiment were broken down into a number of categories by inferred purpose of each tool use. These categories served as the basis for generating tabular profiles of tool use for each participant. Quantitative counts of tool use and tool use variability were then generated from these profiles. This development is described below.

Categories of tool use. Participants' comments, both from their verbal protocols and from the post-experiment interview, along with written comments in their notebooks and control recipes, helped in inferring a purpose for each of the tool uses observed. Below, seventeen categories (by inferred purpose) that encompass all of the participants' tool uses, are described.

- Notebook use. As in the pilot study, participants used their notebooks frequently, and for a number of different purposes. To help in understanding how their notebooks were used, the information written in them has been categorised under six groups. The first three groups attempt to improve on the casual categories of high- and low-level information introduced in the pilot study by categorizing the information by intent as inferred from the context of the trial in which they were made. Three different intents were noted. First, many participants used their notebooks to discuss the different properties of the work domain, their interface, and the task they had to perform. Since participants rarely referred to these notes again after writing them, it seems that their function was to reinforce their knowledge of DURESS II by writing down what they had learned. Thus, these notes have been categorised under the intent

of learning. Second, some participants were observed using their notebooks to help them derive higher order functional information about DURESS II (e.g., the steady-state heater ratios). Third, participants using the divided P+F interface were observed marking down the information contained in goals window to help them integrate information across the levels of the interface. Their notebooks became proxies for the goal window, and helped these participants to remember the information in the goals window while working with other windows.

Some entries did not fit under this classification. This included data about system events (such as records of faults, e.g., "Valve VA1 stopped working so I had to compensate…") and noting of trial completion times (i.e. "Trial 4: 15:32"). These have been included as separate categories. Finally, participants sometimes used the data recorded in their notebooks to reason about deeper properties of the work-domain (for example, recording output demands and their corresponding heater settings with a view to deriving the steady state heater ratios). If these activities were noted in participants' verbal protocols, they were grouped under the category of reasoning.

- Use of stickers to mark active portions of the feedwater stream. At the beginning of many trials, after settling on a valve configuration, one of the participants (Rachmaninov) would use round stickers to mark what he called the 'active' portions of the feedwater streams (see Figure 13). At first, he adopted a colour-coding scheme to help discriminate between valves that were to be used for making adjustments to DURESS II, and those that could be set to the maximum and then forgotten. This colour-coding did not persist, but the use of the stickers to mark the active portions of the feedwater stream did. According to him, this helped in focusing his attention only on the valves that he was using.

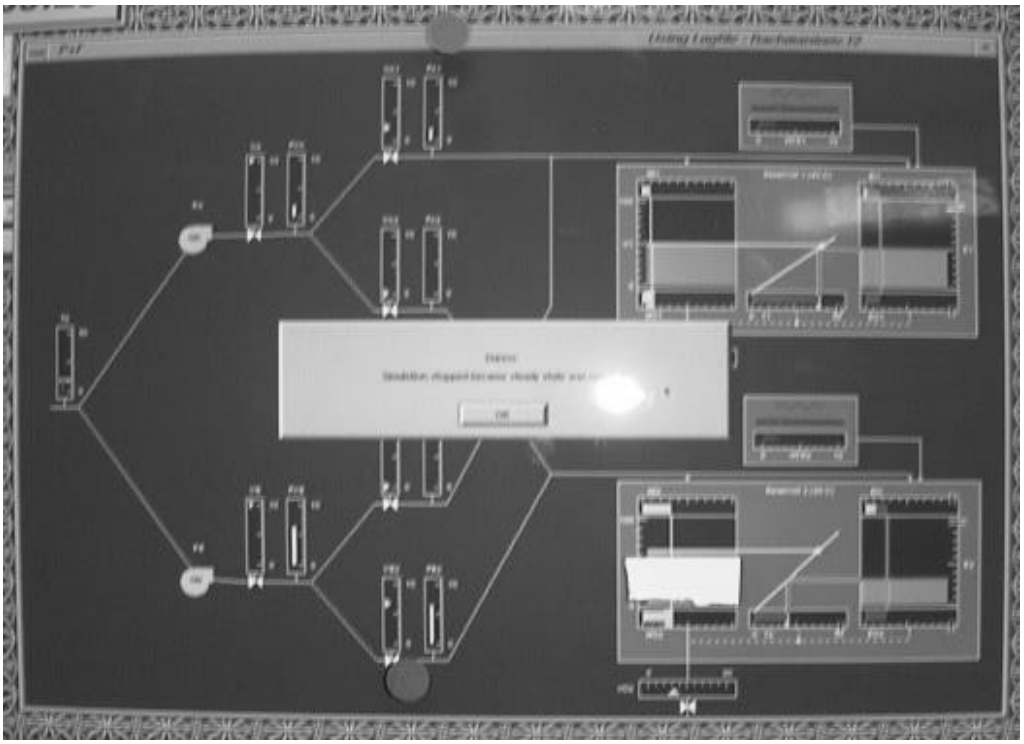


Figure 13. Rachmaninov's use of stickers to mark the active portions of the feedwater stream. Here he has marked that valves VA1 and VB2 are active. Note also the use of a post-it note to aid in monitoring the volume of R2 (see description on p. 57).

- Use of stopwatch to time steady state. Just as was observed in the pilot study, a number of participants in this study used the stopwatch to help in maintaining an estimate of how long DURESS II had been in steady state. Rather than using the simulation timer to keep track of the time that the simulation entered steady state, they offloaded this task to the stopwatch, which they started when they determined that the necessary task goals had been achieved (for a description of the cognitive benefits of using the stopwatch, see p.33).
- Use of the magnifier to focus on screen details. A number of participants used the magnifier to help them line up various controls or to check that the system parameters were indeed within their goal regions.
- Use of post-it notes to keep track of heater manipulations. Two participants would sometimes place a post-it note on the screen beside the reservoirs. On these post-its, they made a record

of the relative manipulations they had made to the heaters (see Figure 14), presumably to aid in retracing their steps in control.

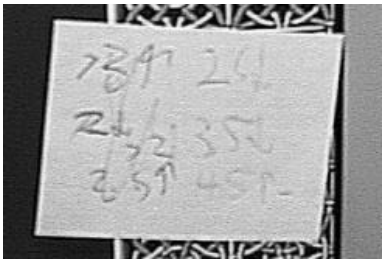


Figure 14. One of Wagner's post-it notes which he used to keep track of his heater manipulations.

- Use of post-it notes or grease marker as an alarm on reservoir volume. Unless the input and output valves have been configured to exactly the same value, reservoir volumes will creep up or down slowly. Even though the P+F interface has been designed to support operators in matching the mass inflow to outflow for each of the reservoirs by providing the emergent feature of a straight line to indicate equal mass or energy input to and from a reservoir, the actual implementation is lacking in supporting this feature. Due to display resolution limitations, a straight line does not necessarily indicate that the input and output to a reservoir are exactly matched, but only that they are nearly matched (see Figure 15). The P interface, on the other hand, lacks any explicit support for comparing the flow into a reservoir to its output. On both interfaces, participants were observed bringing the reservoir volumes to what they thought was steady state, and then marking the reservoir volume (either with a grease marker or a post-it note) so that they could detect even slow changes in the reservoir volumes (for an example of grease marker markings, see Figure 16; for an example of using a post-it note for the same purpose, see Figure 13).

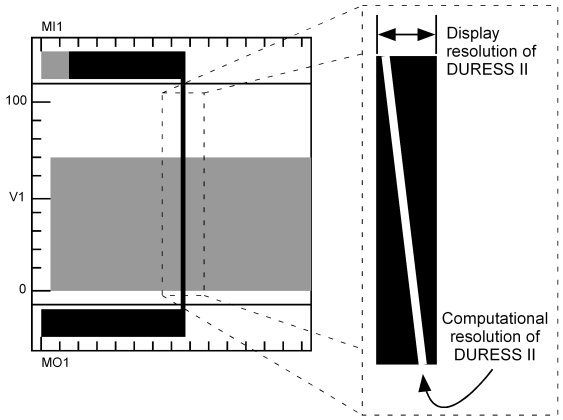


Figure 15. The mismatch between the display and computational resolution of DURESS II. The callout illustrates the coarseness of the display resolution (the black rectangle) in comparison to the relatively finer computation resolution of DURESS II. Due to this mismatch, a straight line between input and output does not necessarily mean that input and output are exactly matched. (Note that the callout is not drawn to scale. The actual computational resolution of DURESS II is finer than the display resolution by many orders of magnitude.)

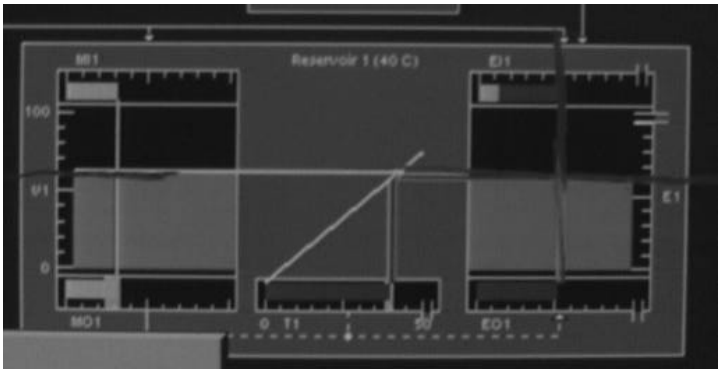


Figure 16. A representative example of the use of grease marker markings to keep track of reservoir volumes. In this case, this participant (Telemann) has made lines across both the mass and energy reservoirs.

- Grease marker to keep track of component names. While learning how to operate DURESS II, two participants wrote various notes about the DURESS II components on the screen with the grease marker to help in recalling what the various graphical elements on the screen represented (see Figure 17). This tool use was only observed during two trials, and seems to have outlived its usefulness as participants learned more about DURESS II.

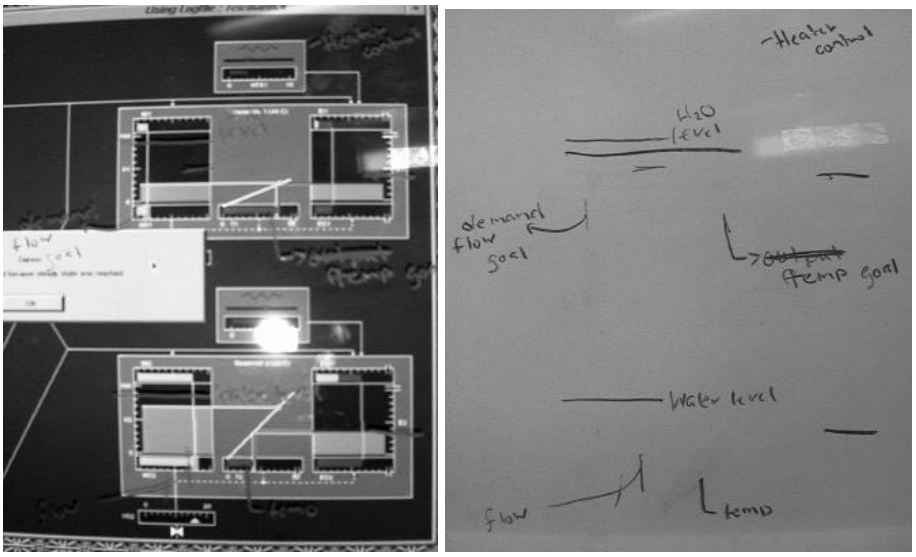


Figure 17. Telemann's efforts to write the names of the components on-screen. The picture on the left shows a snapshot of the screen as she saw it, and the picture on the left uses a white background to make her notes easier to read.

- Grease marker to keep track of flows. After transferring to the P interface, one participant (Wagner) was observed using his grease marker to fashion proxies for the flow meters that he had grown accustomed to on the P+F interface but that the P interface lacked. After performing his initial control actions to move the simulation toward steady state, he would write flow values beside each of the valves. This propagated down the feedwater streams to the top of the reservoirs, where he would write the amount of flow that he expected to be entering each reservoir. He completed his actions by writing the amount of water that he expected should be leaving the reservoir according to his valve settings (see Figure 18). Or, as he wrote in his log:

> …while the heaters 'warm up' use the grease pen to write the input and output levels for each valve beside the valves and put the sum of the inputs by the appropriate tank (i.e., $A_1 + B_1 = In_1$) this will allow you to make any adjustments to the input flows quickly and efficiently. (Wagner, Trial 33 control recipe)

Both these remarks and the fact that Wagner waited until the system was heading toward steady state before making these markings indicate that this action was not simply used to

figure out the valve settings (cf. Hammerstein's use of her log in the pilot study). Rather, this was a more sophisticated behaviour designed to help in understanding the pattern of flows through DURESS II to facilitate adjustment and tweaking.

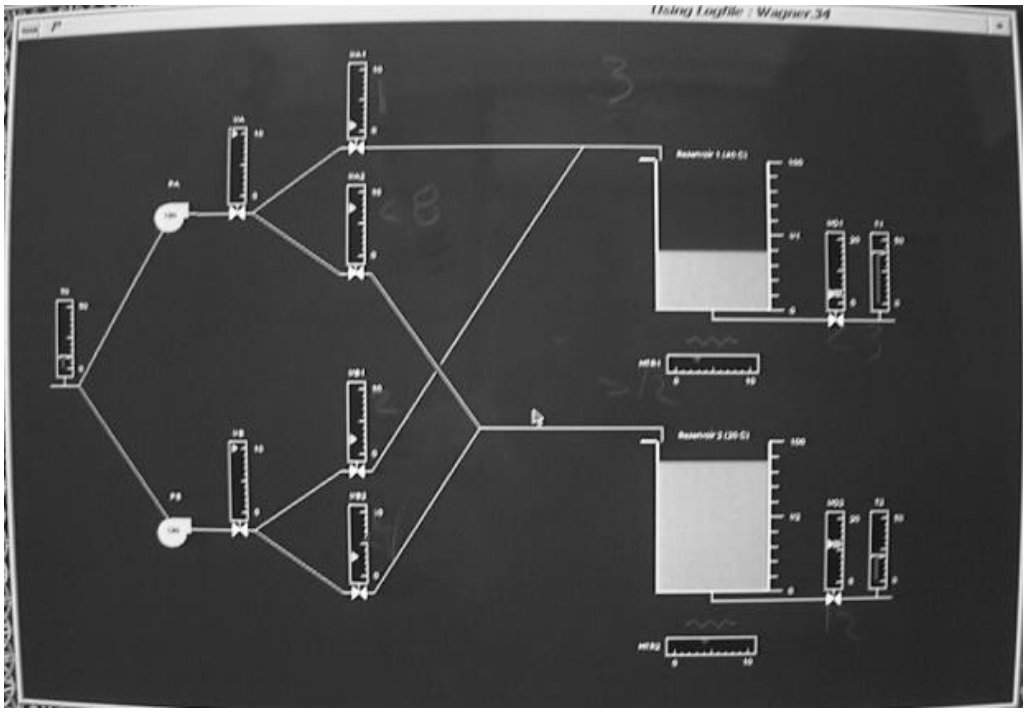


Figure 18. Wagner's use of the grease marker to keep track of flows. The values written beside the valves are: VA1 → 1, VA2 → < 8, VB1 → 2, VB2 → <4, R1 input → 3, R2 input → >12, R1 output → <3, R2 output → 12.

- Use of a straightedge to line up interface elements. In early trials with all three interfaces, some participants were observed using a post-it note, ruler, or some other straightedge to line up controls and their meters or scales. Górecki exhibited an extreme example of this in trial 10. His verbal protocol for trial 9 indicates that since there was a distance between the settings for VO1 and VO2 and their associated flow meters, he was having trouble lining up his setting with the output value he intended. To get around this problem, he fashioned a new gauge out of two post-it notes which he first cut down to size using the scissors (see Figure 19). Unfortunately, since the monitor was built with thick glass, this gauge only worked when

looked at from one angle, and Górecki discarded it quickly. Nonetheless, because these tool uses were discarded quickly, it seems that they lost their usefulness as participants became more attuned to the perceptual features of the display.

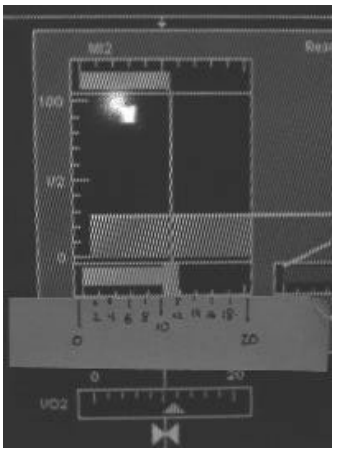


Figure 19. The gauge fashioned by Górecki during trial 10 to line up the setting of VO2 to its corresponding flow meter.

- Use of post-it note or grease marker as a setting memory. After all the parameters had entered their goal regions and participants were waiting for the simulation to reach steady state, they often had to make minor adjustments to keep all variables within their goal regions. To help in this task, a number of participants used either post-it notes or grease marker markings to keep track of any settings that they wanted to reproduce after remedying some small problem. An example of this is shown in Figure 20.

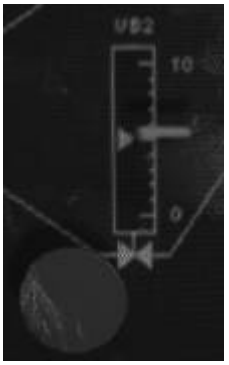


Figure 20. An example of Rachmaninov's use of a grease marker mark as a valve memory.

- Use of tape flags or grease marker to monitor output temperature. Two participants made modifications to help in monitoring the output temperatures. These participants would place either a tape flag or a grease marker marking on the desired output temperature so that they could more easily monitor this indicator for changes (Figure 21).

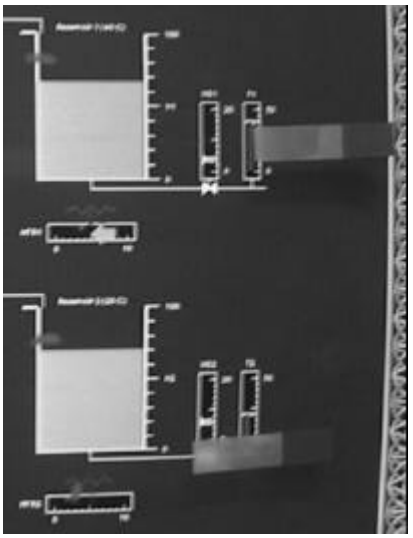


Figure 21. Rachmaninov's use of tape flags to keep track of the output temperatures. The tape flags do not seem to line up with the output temperatures due to the angle at which this picture was taken. Note also the marks on the reservoir volumes and the setting memories on H1 and H2.

- Moving the divided P+F interface control panel for faster screen switching. Finally, two of the participants (Bach and Schoenberg) using the divided P+F interface tended to move the control panel used to navigate between screens from the bottom left-hand corner of the screen (where it is automatically positioned at the beginning of each trial) to the middle of the screen. This modification reduced the distance that these participants had to travel from changing a control setting to switching screens, and so made for faster screen changes.

Tool use profiles. No participant exhibited tool use in all of the seventeen categories described above. Instead, each individual participant used different subsets of these tools and displayed unique trajectories as they explored the possible tool uses and interface modifications

available to them. To capture these trajectories, each participants' tool use has been summarised into a tabular tool use profile that breaks down their tool use in each of these categories over time. An example of a profile is shown in Table 14, and the profiles for each participant can be found in Appendix C. Explanations of the abbreviations in each tool profile are given in Table 13. Each of the profiles found in Appendix C are explained in the text that follows.

Table 13. Abbreviations used in tool use profiles.

Tool use category	Abbreviation
Notebook: integration	NB:Integration
Notebook: derivation	NB:Derivation
Notebook: learning	NB:Learning
Notebook: events	NB:Events
Notebook: times	NB:Times
Notebook: reasoning	NB:Reason
Use of stickers to mark active portions of the feedwater stream	Stickers:FWS
Use of stopwatch to time steady state	SWatch:Timing
Use of the magnifier to focus on screen details	Mag:Details
Use of post-it notes to keep track of heater manipulations	Post-it:Heater
Use of post-it notes or grease marker as an alarm on reservoir volume	Post/Gr:ResAlarm
Grease marker to keep track of component names	Gr:CompNames
Grease marker to keep track of flows	Gr:Flows
Use of a straightedge to line up interface elements	Edge:Lineup
Use of a post-it note or grease marker as a setting memory	Post/Gr:Setting
Use of tape flags or grease marker to monitor output temperature	Flag:OutputT
Moving the divided P+F interface control panel for faster screen switching	Control Panel

Table 14. Example of a tool use profile (for Górecki, divided interface).

Górecki (Div)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																															•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
NB:Derivation											•	•	•	•	•																																					
NB:Learning		•																																																	
NB:Events																																																				
NB:Times																																																			
NB:Reason			•												•																																			
Stickers:FWS																																																		
SWatch:Timing																•	•				•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Mag:Details					•																																													
Post-it:Heater																																																		
Post/Gr:ResAlarm																																																	
Gr:CompNames																																																	
Gr:Flows																																																		
Edge:Lineup					•				•	•	•	•																																					
Post/Gr:Setting						•																			•	•	•	•	•	•																			
Flag:OutputT																																																	
Control Panel																																																	

Descriptions of tool use profiles. The following text describes the tool use profiles of each participant, both in general terms and in terms of the differences between pre- and post-transfer tool use.

- Bartók (P interface). Bartók did not use that many tools, and the uses that he did make were in the pre-transfer phase. Early in this phase, he used his notebook to describe his experience — in prose — under the headings of "Objective", "Techniques", "Observations", and "Update". His notes under these headings mainly discussed DURESS II configuration strategies.
- Mozart (P interface). Mozart used quite a few tools in both the pre- and post-transfer phases. Pre-transfer, Mozart's tool uses included a few quick jottings in his notebook and a few uses of the magnifier and straightedge. Later in the pre-transfer stage, he began to use the grease marker as an alarm on the reservoir volume, and did this consistently from trial 16 to 24. He also used the stopwatch, which was a feature of both his pre- and post-transfer tool use. Post-transfer, during trials 26-31, he again turned to his notebook where he began to write down the demands and their corresponding heater settings for each trial. In trial 31 he used this information to derive the steady state heater ratios (Figure 22). In trial 35, he was confused about the application of these ratios, so he went back to his notebook to re-learn their application. By trial 37, he had internalized the ratios, and never referred to his notebook again.
- Rachmaninov (P interface). Rachmaninov used more tools and made more interface modifications than any other participant. Early in the pre-transfer phase, he began to use stickers to mark the active feedwater stream (he was the only participant to do so), and also used the stopwatch to time steady state. While both of these behaviours persisted for the

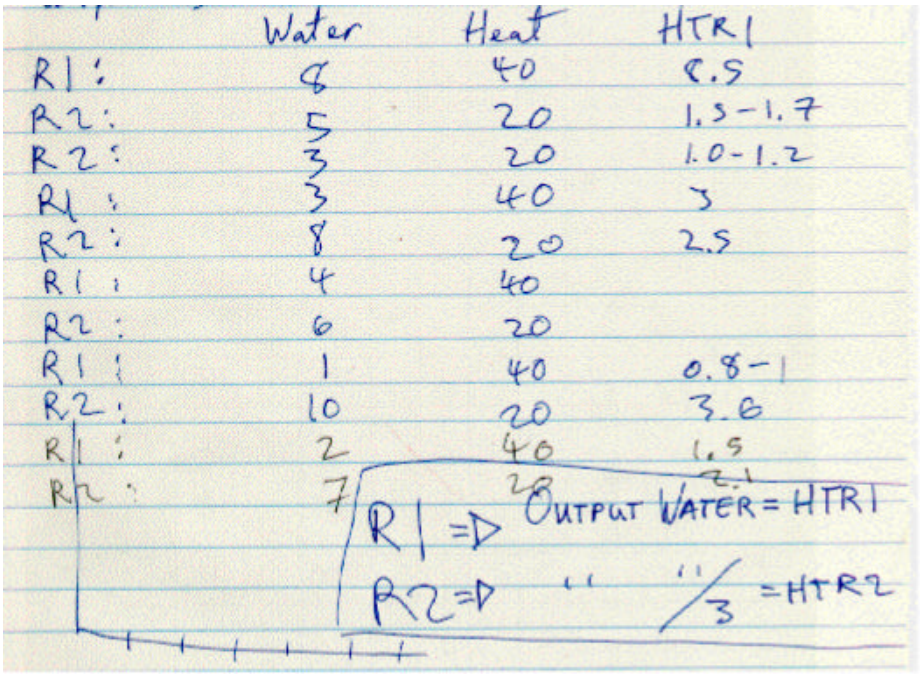


Figure 22. Excerpt from Mozart's log, showing his derivation of the steady state heater ratios.

duration of the experiment, the persistence of his feedwater stream markings is difficult to explain. It is likely that this tool use was helpful as Rachmaninov learned how to operate DURESS II, but that he kept it up either because it became a routine or because he felt that the experimenter wanted to see tool use. This topic will be revisited later.

Other tool uses in the pre-transfer phase included a few uses of the magnifier and straightedge, as well as the use of a number of post-its to make reservoir volume alarms. Late in this phase, he began to make grease marker markings as setting memories.

In the pre-transfer phase, Rachmaninov only made use of his notebook to record the completion times for a number of trials, but in the post-transfer phase this changed. In the second post-transfer trial (trial 24), he began writing down the output demands and temperatures, reservoir volumes, and heater settings for both of the reservoirs, and on trial 29 he plotted all of this data on a graph (Figure 23). He used this graph at the beginning of each trial to find the necessary heater settings, and in one trial even extrapolated to accommodate a

demand pair not within the bounds of his graph. In the post-transfer phase, he also continued to use grease marker markings as setting memories, and for a brief time experimented with using tape-flags to help in monitoring the output temperature.

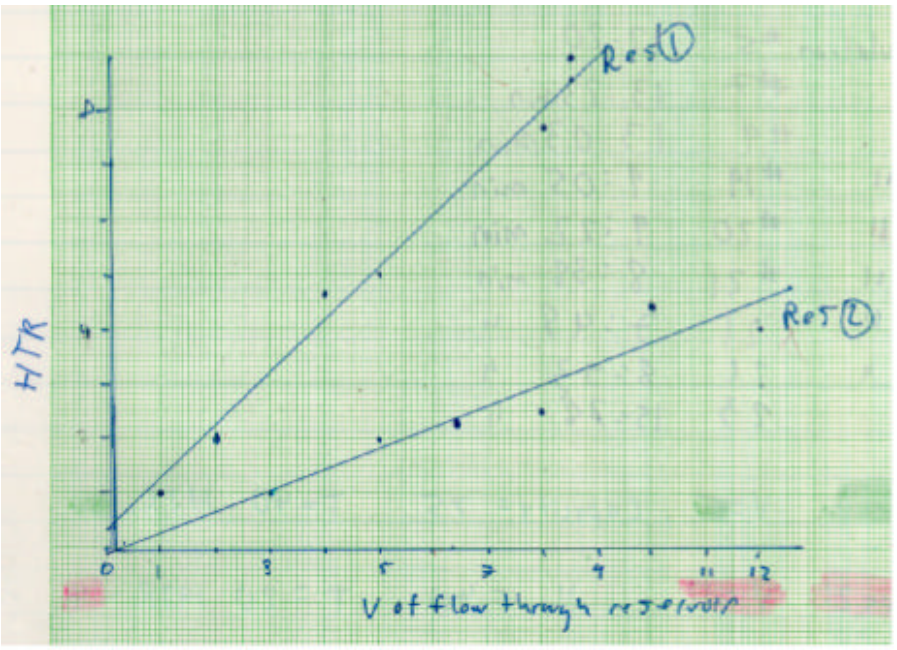


Figure 23. The graph constructed by Rachmaninov to help in finding the steady-state heater settings.

- Telemann (P interface). Telemann engaged in a moderate amount of tool use. Early in the first phase she made a few preliminary jottings in her notebook as she learned about DURESS II. Also, in trials 3 and 4, she used the grease marker to write component names on the screen, presumably in an effort to learn the meanings of the graphical forms displayed on-screen. Early in the pre-transfer phase she also started using the grease marker to make an alarm on the reservoir volume, and kept this up fairly consistently for the duration of the experiment. During one trial she used a post-it note to help keep track of heater manipulations. Early in the post-transfer phase she made a few more jottings in her notebook and also used the straightedge, but the only tool use that she kept up over the long-term was

making reservoir volume alarms.

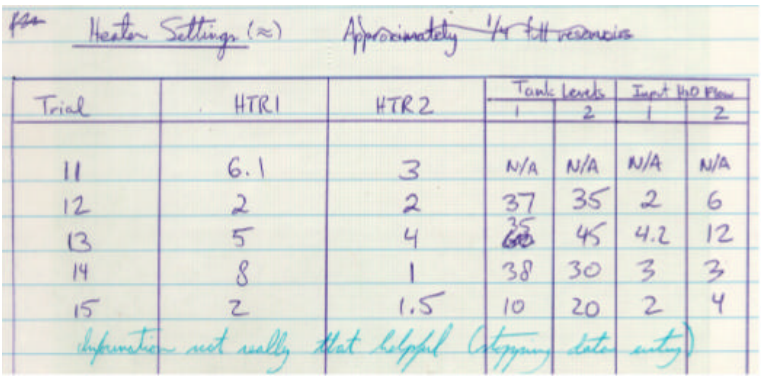
- Wagner (P interface). Wagner did not use his tools very much, but the tool uses that he did make are notable. In the pre-transfer stage, he explored the use of tools but did not settle on anything for more than one use. Post-transfer, at trial 28, he began to use the grease marker to derive the flows through the valves (see Figure 18 and the accompanying description of this tool use on p. 60 ff.). He was the only participant observed using tools in this way. Late in the experiment he also began to use post-it notes to keep track of the heater settings.
- Willan (P interface). Willan did not use any tools over the course of the experiment.
- Bach (divided interface). Bach did not use many tools. In the pre-transfer stage she made a number of notes in her notebook that included tips on DURESS II configuration as well as notes to herself (or to the experimenter?), like "I don't ever pay intention [sic] to T0 (is that important?)." She also tried to make setting memories on the heaters to persist across trials, but soon realised that since the output demands change each trial, these marks were ineffective. In the post-transfer phase, she used her notebook once to make some final comments about the dynamics of DURESS II. These comments do not seem to be influenced by the change in interface, but rather reflect a growing understanding of the work-domain and her task. In trial 28, she began moving the control panel to the centre of the screen to make switching between screens faster.
- Boccherini (divided interface). Boccherini also did not engage in a great deal of tool use. In the pre-transfer stage, she made a number of markings in her notebook mainly relating to the task of configuring DURESS II. This continued briefly into the post-transfer phase. One isolated, yet notable, tool use was her use of the grease marker to monitor the output temperature. While on the goals screen, Boccherini drew a line to indicate the temperature

goal regions. Since the temperature indicators are in the same place on all screens, this modification replicated the information from the goals level across all levels. Probably due to the parallax caused by the thick glass of the monitor, she only made this modification once.

- Górecki (divided interface). Górecki used a moderate amount of tools. His activity in the pre-transfer phase was mainly exploratory, and included the use of the stopwatch and the fashioning of a supplementary gauge to make the correspondence betweenthe output flow meter for the reservoirs and the setting of the output valves more clear (see Figure 19). His use of the notebook in the pre-transfer phase is notable: during trials 11-15, he constructed a chart to record the heater settings, volume levels, and input flows to the reservoirs (Figure 24). The title to this chart ("Heater Settings") makes it clear that he was searching for some relationship that would predict the steady state heater settings. Unfortunately, he was not able to find a relationship, and wrote "Information not really that helpful (stopping data entry)".

  In the post-transfer phase, Górecki's use of the stopwatch became more consistent. More notable were his efforts to make the information contained in the goals level available across all levels of the interface. Starting in trial 25, Górecki began recording the output demand goals on post-it notes that he would place on the screen near between and just to the right of the two reservoirs (Figure 25). In trial 31, instead of using post-it notes, he recorded this information in his notebook (Figure 26).

- Prokofiev (divided interface). Prokofiev used a moderate amount of tools. She used the stopwatch in both pre- and post-transfer trials, but most of her tool use was restricted to her notebook. Her use of this tool had an interesting development. In the pre-transfer trials, she tended to make a notebook entry for each day (i.e., they had the titles "Day 1", "Day 2",


Heater Settings (≈) Approximately ¼ full reservoirs

Trial	HTR1	HTR2	Tank Levels 1	Tank Levels 2	Input H2O Flow 1	Input H2O Flow 2
11	6.1	3	N/A	N/A	N/A	N/A
12	2	2	37	35	2	6
13	5	4	35	45	4.2	12
14	8	1	38	30	3	3
15	2	1.5	10	20	2	4

Information not really that helpful (typing, later entry)

Figure 24. The chart used by Górecki to (unsuccessfully) derive the steady state heater ratios.

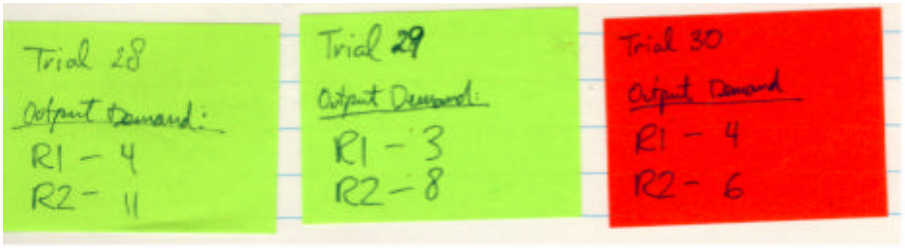
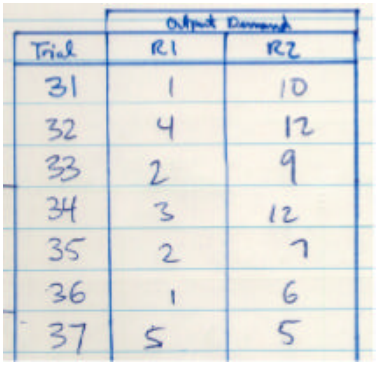


Figure 25. Three of the post-it notes used by Górecki to record the output demands, as he placed them in his notebook after each trial.



Trial	Output Demand R1	Output Demand R2
31	1	10
32	4	12
33	2	9
34	3	12
35	2	7
36	1	6
37	5	5

Figure 26. An excerpt of the chart used by Górecki to record the output demands for each trial.

etc.). These entries were used for recording hints about configuring DURESS II as well as system events and were mainly written in prose. At the interface transfer (her "Day 8"), this format changed dramatically. Where before she wrote prose to help in configuring DURESS II, she now wrote down numerical information about the goals for each reservoir (see the contrast in Figure 27).

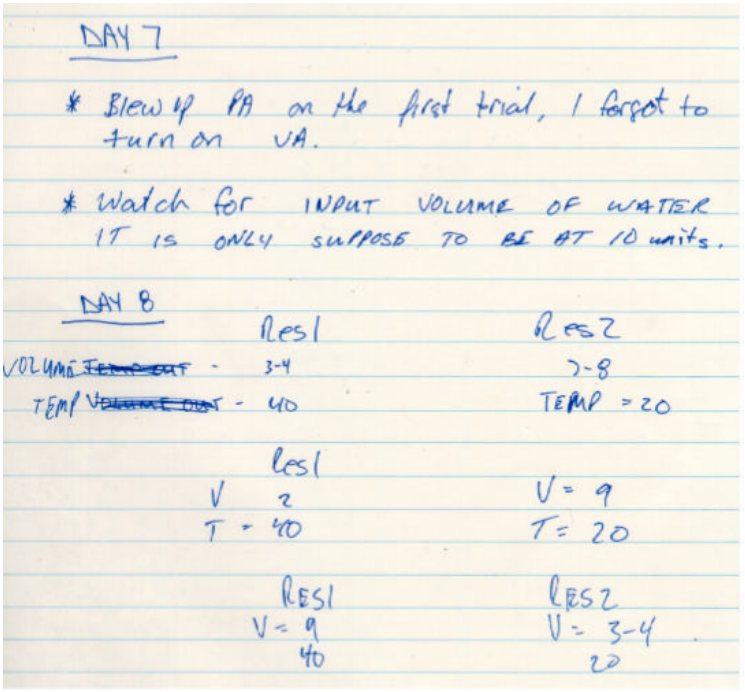


Figure 27. Entries in Prokofiev's log for the day before and the day after the interface transfer.

- Schoenberg (divided interface). Schoenberg did not engage in a great deal of tool use. He used the stopwatch for trials 15-23 of the pre-transfer phase, and then only once again in the post-transfer phase. He also adjusted the position of the control panel during a number of trials late in the post-transfer phase.
- Schubert (divided interface). Schubert's only tool use was confined to his notebook, which he used to record times and system events in both the pre- and post-transfer phases of the experiment. He also recorded a number of humorous messages on post-it notes and made small sculptures out of paper clips, but since these had no relevance to operating DURESS II, they were not recorded on the tool use profile and will not be included in subsequent analyses.

Summary of qualitative analyses of tool use. A number of patterns motivating these 18 categories of tool use were found, and help to tie these categories into a more coherent framework. These motivations for tool use are:

- Tool use as an aid to, or artefact of, learning. Many participants were observed using tools in connection with learning how to operate DURESS II. Tool uses that fit in this category include:
  - The use of notebooks to record information about the operation of DURESS II. Bach, Bartók, Boccherini, Górecki, Mozart, Prokofiev, Telemann, and Wagner were observed doing this during the pre-transfer phase, and Bach (divided), Boccherini (divided), and Telemann (P) did this post-transfer. Each of these participants wrote logs containing information on the characteristics of DURESS II, but rarely referred to their notes after having written them. The act of writing them seemed to be important of itself.
  - Writing component names on the screen with the grease marker. Telemann and Wagner each exhibited this type of tool use, and each for only one or two trials in the pre-transfer phase. After they had learned the meaning of the graphical forms on the interface, this tool use was no longer necessary.
  - Using a straightedge to line up controls and display elements. This tool use was exhibited by Górecki, Rachmaninov, and Wagner in pre-transfer trials, and by Telemann (P) in post-transfer trials. Each participant seemed to use these types of tools as they learned how the various controls and indicators of the various interfaces corresponded to one another, and only for a short while (Górecki used a straightedge most often, and this was only during four trials).
  - Using the magnifier to focus on screen details. This tool use is related to the use of the straightedge, and was exhibited by Górecki, Rachmaninov, and Mozart pre-transfer and by Mozart (P) post-transfer. Again, this tool use helped these participants to learn the features of the interface, and was only exhibited over the short-term (a maximum of three

trials by any one participant).

- Marking the active portions of the feedwater stream. Although this tool use, exhibited only by Rachmaninov (P), persisted over the long-term, it seems that its purpose was to help Rachmaninov learn about the interface. It is likely that it persisted either because it became a rote action or because Rachmaninov thought that the experimenter would be more pleased if he exhibited this tool use.

Although varied, each of these tool uses were connected to participants' learning the DURESS II work-domain and its interfaces. It is difficult, however, to draw a causal link between learning and tool use. Tool use might have a function in learning, but it is also possible that it is an artefact of learning. If this is the case, these instances of tool use might be a form of self-explanation that has resulted from learning (Howie & Vicente, 1998).

- On the divided P+F interface, tool use to help in integrating information across levels. The main design deficiency of the divided P+F interface in relation to the P+F interface is that it decomposes information about the work domain over a number of levels, and so forces the user to integrate the information over these levels in order to understand how the plant is functioning. Three out of six participants engaged in tool uses to help them in this task. As soon as they switched to the divided interface, two participants (Prokofiev and Górecki) began to use post-it notes and their notebooks to record the information represented on the goals level of the interface (cf. Figure 25). One other participant, Boccherini, tried using her grease marker to mark the goal temperatures and so replicate this information across levels. Both of these types of tool uses were attempts to represent the goal information so that it could be more easily used across levels.

- On the P interface, tool use to help in deriving higher-order functional information about the system state. The main design deficiency of the P interface in relation to the P+F interface is that it does not display all of the goal relevant relationships and constraints that govern the system. Christoffersen, Hunter, and Vicente (1997) observe that in the absence of these types of information, users are left to either (a) operate the system by trial and error, (b) construct rules and look for violations of them, or (c) derive this information themselves. Participants on the P interface engaged in tool uses directed to the last two of these. First of all, after transferring to the P interface, Wagner began his practice of noting the flows that should be passing through each valve beside that valve (cf. Figure 18). These values were derived using a rule, and were proxies for the information that he was used to seeing on the P+F interface. Second, both Mozart and Rachmaninov used their notebooks to derive the steady state heater ratios soon after transferring to the P interface (cf. Figure 22). In switching to the P interface, they lost the support provided by the P+F interface to facilitate reasoning at the level of first principles (i.e., abstract function) and so derived information that partially obviated the need for that type of reasoning. Both of these types of tool use helped these participants to deal with the new demands placed on them by the P interface.
- On all interfaces, tool use to extend the functionality of the interfaces in ways not considered by the designers. A final class of tool use involves modifications directed at extending the functionality of the interface to either address issues not considered by the designer or to suit individual needs. Examples of these types of tool use included:
  - Reservoir volume 'alarm'. Three participants (Mozart, Rachmaninov, and Telemann) used the grease marker or post-it notes to make reservoir volume alarms in both the pre- and post-transfer phases (P interface). The task of monitoring the reservoir volume is

supported imperfectly on the P+F interface, and not at all on the P interface that these three participants transferred to.

- Setting memories. To help in making adjustments to DURESS II, a number of participants used grease marker markings to add setting memories to the various interfaces. This included Górecki, Rachmaninov, and Telemann in the pre-transfer trials, and Rachmaninov (P), Wagner (P), Boccherini (divided) and Górecki (divided) in post-transfer trials.
- Use of the stopwatch to time steady state. Although the design of DURESS II includes a simulation timer, the stopwatch is a better design that was used by many participants in both the pre- and post-transfer phases: Górecki, Mozart, Prokofiev, Schoenberg, Rachmaninov, and Wagner in the pre-transfer phase, and Mozart (P), Rachmaninov (P), Telemann (P), Górecki (divided), Prokofiev (divided), and Schoenberg (divided) post-transfer.
- Notebook for times and events. A number of participants used their notebooks to record histories of their performance and of system events. This included Boccherini, Rachmaninov, and Schubert in the pre-transfer trials, and Telemann (P), Prokofiev (divided) and Schubert (divided) post-transfer.

While the individual profiles of tool use may at first seem to be based on little more than personal preferences, these categories demonstrate participants' tool use was motivated by three types of concerns: learning about the system, replacing information lost during the interface transfer, and adding functionality not available in the first place. Although some of the manifestations were idiosyncratic, all tool uses were driven by common concerns as defined by each of these four categories.

Quantitative measures of tool use. The qualitative descriptions of tool use covered above give a good indication of the variety of tool uses that participants engaged in, the motive behind these tool uses, and how they developed and were affected by the interface transfer at trial 24. Below, these qualitative descriptions are supplemented by two quantitative measures derived from the tool use profiles, total tool use and tool use variability.

Total tool use. A count of total tool use for each participant was made by counting the number of different categories of tools used in each trial. Under this scheme, multiple uses of one type of tool only counted as one tool use. For instance, participants who used the stopwatch tended to refer to it many times over the course of the trial; nonetheless, this was counted as one tool use. An alternative method to calculate this measure would have been to count the number of uses of or references to each type of tool within each trial, but this was not done for two reasons. First, different tools naturally afford different frequencies of use. For instance, setting memories might only be referred to once or twice within the context of making adjustments to system parameters, while volume alarms might be referred to frequently as participants sample the system state. Second, in the absence of sophisticated measuring systems it is difficult to count the number of references to some tools reliably (e.g., the setting memories and volume alarms). The method of counting the number of types of tool uses within each trial, on the other hand, treats all tools equally.

Tool use counts for all participants are given in Table 15. These data are graphed in Figure 28.

Table 15. Tool use counts by participant and interface. Participants are ordered within interface from highest to lowest total tool uses.

		Total Tool Uses			*Average per Trial*		
	Participant	**Pre-Trans**	**Post-Trans**	**Total**	**Pre-Trans**	**Post-Trans**	**Total**
P	Rachmaninov	62	120	182	2.70	4.14	3.50
	Mozart	34	43	77	1.48	1.48	1.48
	Telemann	18	31	49	0.78	1.07	0.94
	Wagner	3	28	31	0.13	0.97	0.60
	Bartók	4	0	4	0.17	0.00	0.08
	Willan	0	0	0	0.00	0.00	0.00
Divided	Prokofiev	20	57	77	0.87	1.97	1.48
	Górecki	19	56	75	0.83	1.93	1.44
	Bach	10	26	36	0.43	0.90	0.69
	Schubert	19	15	34	0.83	0.52	0.65
	Schoenberg	9	10	19	0.39	0.34	0.37
	Boccherini	5	2	7	0.22	0.07	0.13
Averages	P	20.0	37.0	57.0	0.87	1.28	1.10
	Divided	14.3	27.7	42.0	0.62	0.95	0.81

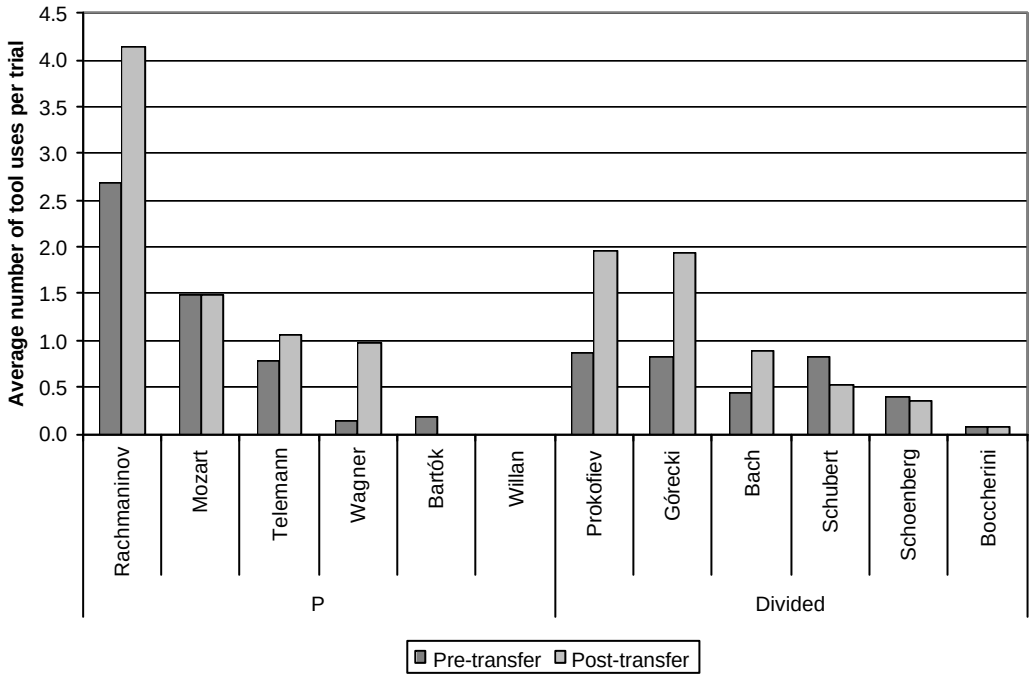


Figure 28. Average number of tool uses per trial, by participant and interface.

The first issue to address in analysing these data is whether or not there is an interface effect on the number of tools used, as this determines whether the data will with analysed as one or two populations. Two ANOVAs[5] were performed on these data. Table 16 contains the results of an ANOVA on the pre-transfer data and Table 17 contains the results for the post-transfer data. These ANOVAs show that there is no interface effect. When coupled with Figure 28, this result indicates that the divided group used less tools than the P group overall, but that this difference was not reliable.

The ANOVA on pre-transfer tool use also indicates a significant trial effect. The average number of tool uses per trial has been plotted in Figure 29 and demonstrates that tool use increased steadily over the pre-transfer phase. However, tool use stabilized in the post-transfer period (cf. the lack of an trial effect in the post-transfer trial).

Since there was no significant interface effect in the pre-transfer phase, the P and divided groups can be considered as coming from similar populations. Accordingly, an independent samples *t*-test was performed on the average number of tool uses per trial for all participants in the pre- and post-transfer phases. This test confirmed that there was not a significant increase in tool use between the pre- and post-transfer phases ($t_{22} = 0.94$, $p = 0.36$).

In summary, the two groups of participants engaged in roughly the same number of tool uses in the pre-transfer phase, and there was no interface effect on the number of tool uses post-transfer. In addition, even though Figure 28 gives an indication that tool use increased in the post-transfer phase, this increase was not reliable enough to be considered significant.

---

[5] These data were balanced, and so regular ANOVA techniques could be applied. In what follows, ANOVA will be applied to balanced data, and general linear models (GLMs) will be applied to unbalanced data. See footnote 4 on p. 40 for a further explanation.

Table 16. ANOVA on pre-transfer tool use.[6]

Source	DF	SS	MS	F	p
Interface	1	5.797	5.797	0.43	0.527
Participant(Interface)	10	134.971	13.497		
Trial	22	15.493	0.704	1.59	0.050
Interface × Trial	22	8.536	0.388	0.88	0.626
Participant × Trial (Interface)	220	97.362	0.443		

Table 17. ANOVA on post-transfer tool use.

Source	DF	SS	MS	F	p
Interface	1	9.011	9.011	0.21	0.658
Participant(Interface)	10	433.701	43.370		
Trial	28	1.736	0.062	0.27	0.999
Interface × Trial	28	5.655	0.202	0.89	0.625
Participant × Trial (Interface)	280	633.299	0.226		

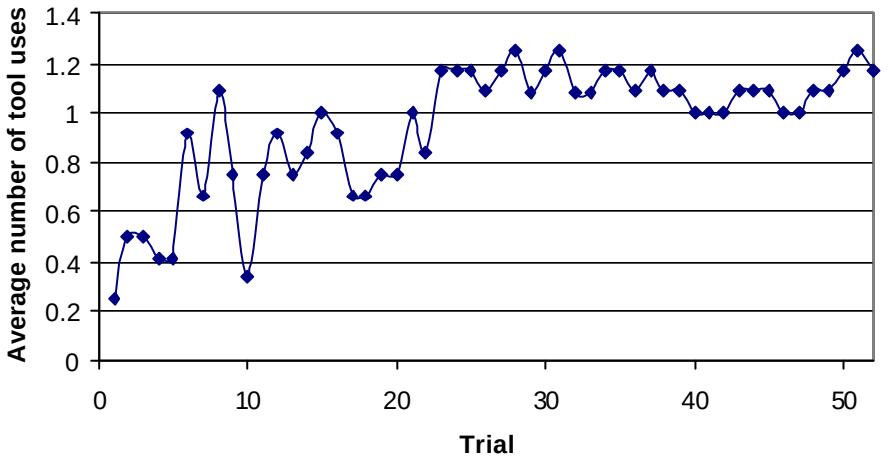


Figure 29. Average number of tool uses per trial across participants.

[6] Although it may seem counterintuitive to perform an ANOVA with interface as a primary effect when all subjects were using the same interface in the pre-transfer phase, the purpose of this analysis was to see if there was a significant difference between the tool use of the two groups before the interface transfer.

Tool use by category. To supplement the categories of tool use developed from the qualitative analyses of tool use, the number of tool uses in each category for both the pre- and post-transfer phases were counted. These results are shown in Table 18 and Figure 30.

Table 18. Total number of tool uses by category.

		Pre-transfer		*Post-Transfer*			
		Learning	**Addition**	**Integration**	**Derivation**	**Learning**	**Addition**
P	Rachmaninov	27	35		29	29	62
	Mozart	11	23		13	1	29
	Telemann	4	14			5	26
	Wagner	2	1		25		3
	Bartók	4					
	Willan						
Divided	Prokofiev	8	12	27			30
	Górecki	12	7	22			34
	Bach	8	2			1	25
	Schubert		19				15
	Schoenberg		9				10
	Boccherini	4	1			1	1
	Totals	80	123	49	67	37	235
	Percentages	39.4%	60.6%	12.6%	17.3%	9.5%	60.6%

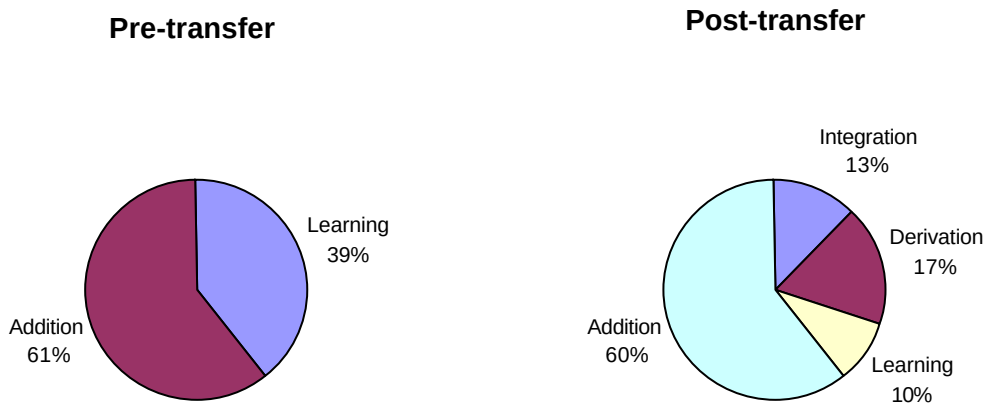


Figure 30. Percentages of total tool use, by category.

These results indicate that the majority of tool uses fell under the category of adding to or extending the functionality of the interfaces being used. The amount of tool uses directed at learning decreased in the post-transfer phase, and was replaced by tool uses directed at integration (divided) and derivation (P). It is notable that the tool uses in these two categories made up over one-quarter of the post-transfer tool use.

Tool use variability. A second quantitative measure was designed to capture the stability (or lack thereof) in the patterns of participants' tool use. The approach taken was to count the net number of variations in tool use for each participant. For each type of tool use that a participant either adopted or dropped in a given trial, they would be scored one point. For instance, if a participant used tools A, B, and C in trial *n* and tools A, D, and E in trial *n* + 1, his variability would be scored as follows:

- since there was no change in his use of tool A over the two trials, this would be given 0 points
- since he dropped tool B and C in the transition to trial *n* + 1, this would receive 2 points
- since he added tools D and E in trial *n* + 1, this would receive 2 points

In total, trial *n* + 1 would be given a score of 5 variability points.

These variability points calculated for each participant and summed over the pre- and post-transfer phases are shown in Table 19. Aggregate data is graphed in Figure 31, individual data for the P group is graphed in Figure 32, and for the Divided group is graphed in Figure 33.

Just as with the counts of tool uses, the first issue is to determine if there is an interface effect on tool use variability. To achieve this, ANOVAs were performed on tool use variability[7] for both pre- and post-transfer performance (Tables 20 and 21, respectively).

---

[7] An ANOVA can be performed on tool use variability because this statistic is counted within trials, as opposed to statistics like s^2 which are calculated across trials.

Table 19. Tool use variability by participant and interface.

		Totals			*Per Trial Averages*		
		Pre Trans	**Post Trans**	**Overall**	**Pre Trans**	**Post Trans**	**Overall**
P	Rachmaninov	32	9	41	1.39	0.31	0.79
	Mozart	27	15	42	1.17	0.52	0.81
	Telemann	13	22	35	0.57	0.76	0.67
	Wagner	6	3	9	0.26	0.10	0.17
	Bartók	5	0	5	0.22	0.00	0.10
	Willan	0	0	0	0.00	0.00	0.00
Divided	Prokofiev*	11	7	18	0.48	0.24	0.35
	Górecki	19	5	24	0.83	0.17	0.46
	Bach	10	2	12	0.43	0.07	0.23
	Schubert	8	2	10	0.35	0.07	0.19
	Schoenberg	1	10	11	0.04	0.34	0.21
	Boccherini	9	5	14	0.39	0.17	0.27

*Prokofiev's pre-transfer notebook uses do not add to her tool variability, as she used her notebook consistently at the end of each day. Thus no variability points were scored for her notebook use notebook in trials 3, 6, 9, 12, 15, 19, and 23.

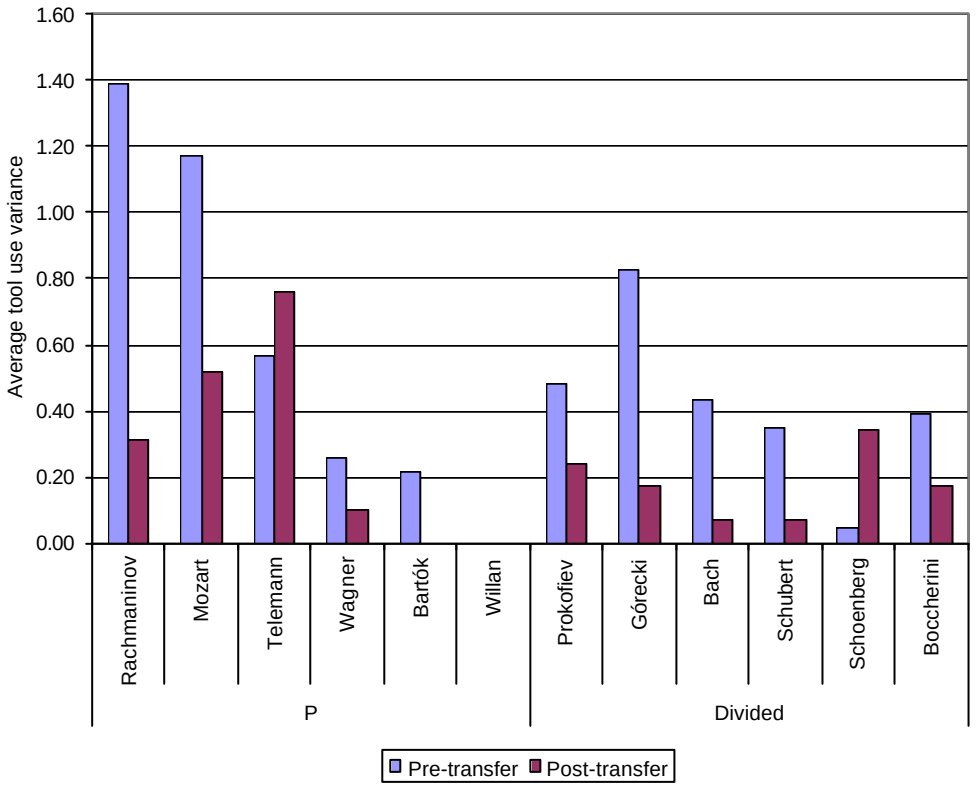


Figure 31. Tool use variability by participant and interface.

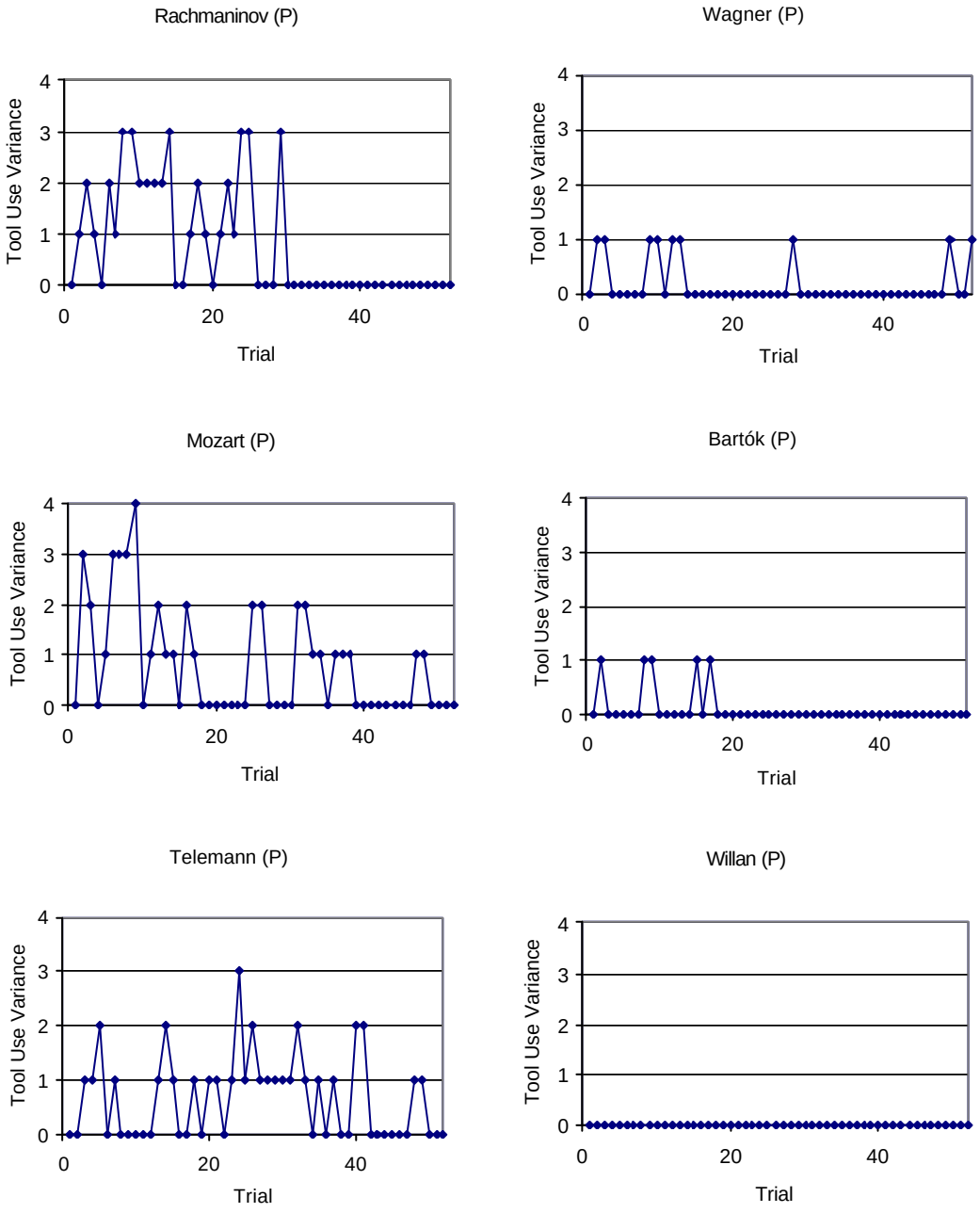


Figure 32. Tool use variability for the P interface group.

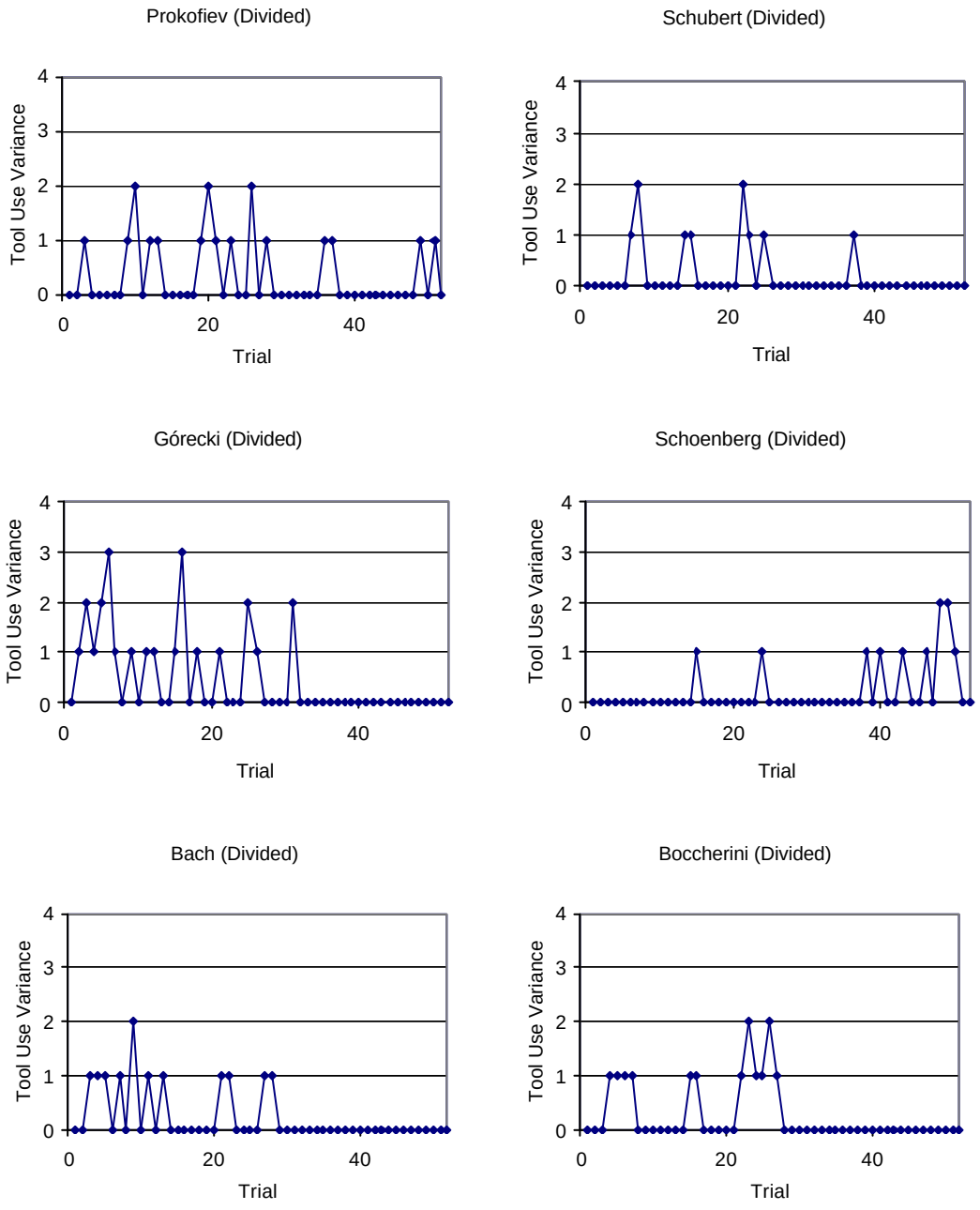


Figure 33. Tool use variability for the Divided interface group.

Table 20. ANOVA on pre-transfer tool use variability.

Source	DF	SS	MS	F	p
Interface	1	2.560	2.560	0.56	0.47
Participant(Interface)	10	45.515	4.552		
Trial	21	11.955	0.569	1.14	0.30
Interface × Trial	21	11.106	0.529	1.06	0.39
Participant × Trial (Interface)	210	104.485	0.498		

Table 21. ANOVA on post-transfer tool use variability.

Source	DF	SS	MS	F	p
Interface	1	0.931	0.931	0.61	0.454
Participant(Interface)	10	15.368	1.537		
Trial	28	16.443	0.587	2.17	< 0.001
Interface × Trial	28	5.236	0.187	0.69	0.879
Participant × Trial (Interface)	280	75.632	0.270		

In the first place, these ANOVAs confirm that the two groups were similarly matched during pre-transfer trials, and that no interface effect existed during post-transfer trials. Second, there was no trial effect during the pre-transfer trials, but a significant trial effect during post-transfer trials. The graphs of tool variability for the individual participants confirm that this trial effect in the post-transfer phase was a reduction in tool use variability. In other words, tool use became more stable over the post-transfer phase. Finally, to supplement these analyses, an independent samples *t*-test was performed to compare the pre- to post-transfer tool use variability for all participants. This test revealed that the difference in tool use variability between the pre- and post-transfer trials was almost significant ($t_{16.8} = -2.02$, $p = .06$); a glance at Figure 31 confirms that tool use variability decreased for 9 out of 12 participants.

Summary of quantitative analyses. Pulling together the data on tool use counts and tool use variability begins to draw a picture of the development of tool use over time. First of all, the data on tool use counts indicates that the amount of tool use at least remained constant, and perhaps even increased over time. At the same time, the amount of tool use variability decreased

over time, both between the pre- and post-transfer phases and over the course of the post-transfer phase. Together, these results indicate that even if the net amount of tool increased over time, more stable patterns of tool use also developed over time. Participants engaged in a great deal of exploration in the pre-transfer phase and early in the post-transfer phase, but then generally settled on a consistent pattern of tool use during later trials.

More difficult to explain is the fact that there was no trial effect in the pre-transfer phase. It is most likely that this phase was not long enough to bring participants beyond exploring the possible tool uses. However, there is also an interesting alternate explanation. It is possible that lower task demands on the P+F interface gave participants the opportunity to explore possible tool uses, but that the P and divided P+F increased task demands to the point that participants looked for a useful set of tools, settled on them, and then worked on honing their task performance. In other words, there could be an interaction between interface and experience on the development of tool use. More thorough experimentation is needed to address these issues.

Results: Normal Trials

Trial completion time. Results on normal trial completion times for the pre- and post-transfer phases are presented below. Tables 22 and 23 contain the average trial completion times and standard deviations for the P and divided groups, respectively. These data are presented graphically, with 95% confidence intervals in Figure 34 (for the P interface group) and Figure 35 (for the divided interface group). Individual practice curves for the P group are presented in Figure 36 and for the divided group in Figure 37.

Table 22. Average trial completion time (in seconds) for the P interface group, ordered from fastest to slowest.

		Willan	Wagner	Mozart	Bartók	Rachmaninov	Telemann
Pre-transfer	$\bar{x}$	581	623	630	683	695	939
	s^2	155	246	200	325	245	306
Post-transfer	$\bar{x}$	646	617	565	660	796	979
	s^2	166	246	186	191	224	404

Table 23. Average trial completion time (in seconds) for the divided interface group, ordered from fastest to slowest.

		Schubert	Schoenberg	Bach	Górecki	Prokofiev	Boccherini
Pre-transfer	$\bar{x}$	480	591	599	631	644	1048
	s^2	90	132	102	206	203	534
Post-transfer	$\bar{x}$	637	741	766	655	718	872
	s^2	128	189	313	136	204	236

Figures 34 and 35 illustrate that the overall performance of the P group remained roughly the same in the pre- and post-transfer periods (3 participants' completion times increased slightly, and the other three decreased slightly), but that the completion times of all participants in the divided group increased in the post-transfer trials. That the divided interface led to higher trial completion times than the P interface is most likely a product of the time penalty incurred while switching screens to navigate between levels.

Note also that the individual practice curves generally spike at around trial 23, indicating the effect of the interface transfer. Figure 38 shows a plot of the variability in trial completion time for the 10 trials immediately prior to and immediate after the interface transfer for each participant. Variability increased over this period for 11 out of 12 participants, demonstrating that the interface transfer had a pronounced effect on performance consistency, especially for the P group.

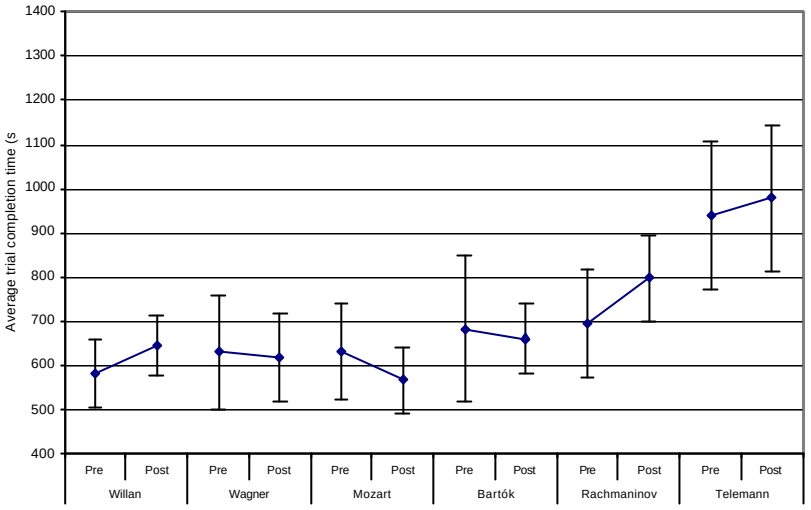


Figure 34. Average trial completion times for the P interface group.

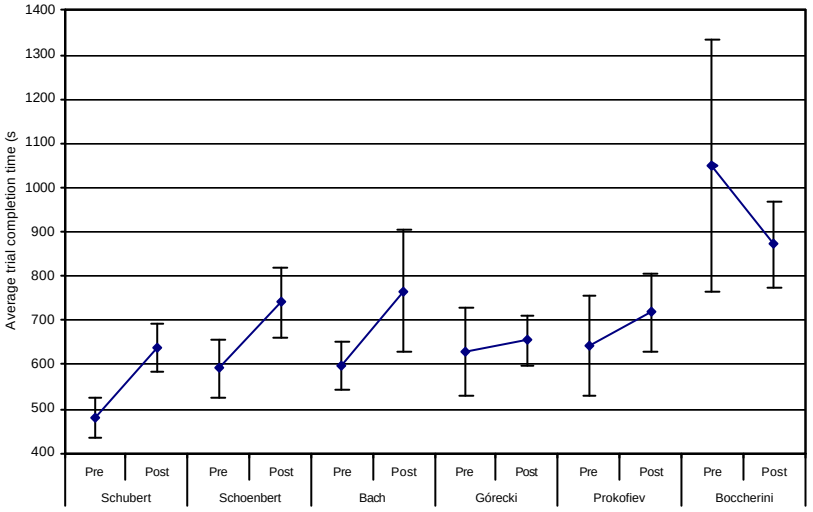


Figure 35. Average trial completion times for the divided interface group.

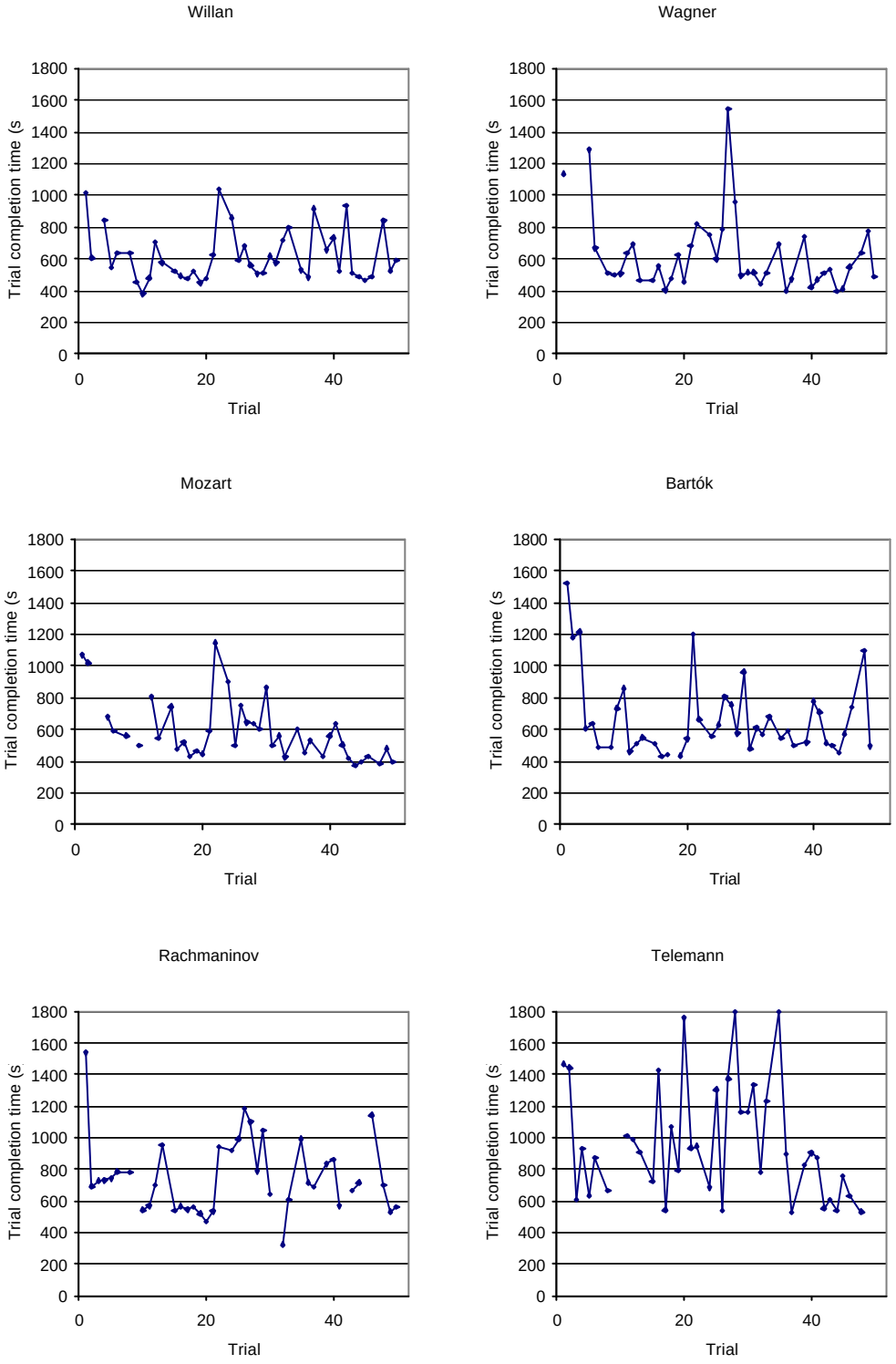


Figure 36. Individual practice curves for the P interface group.

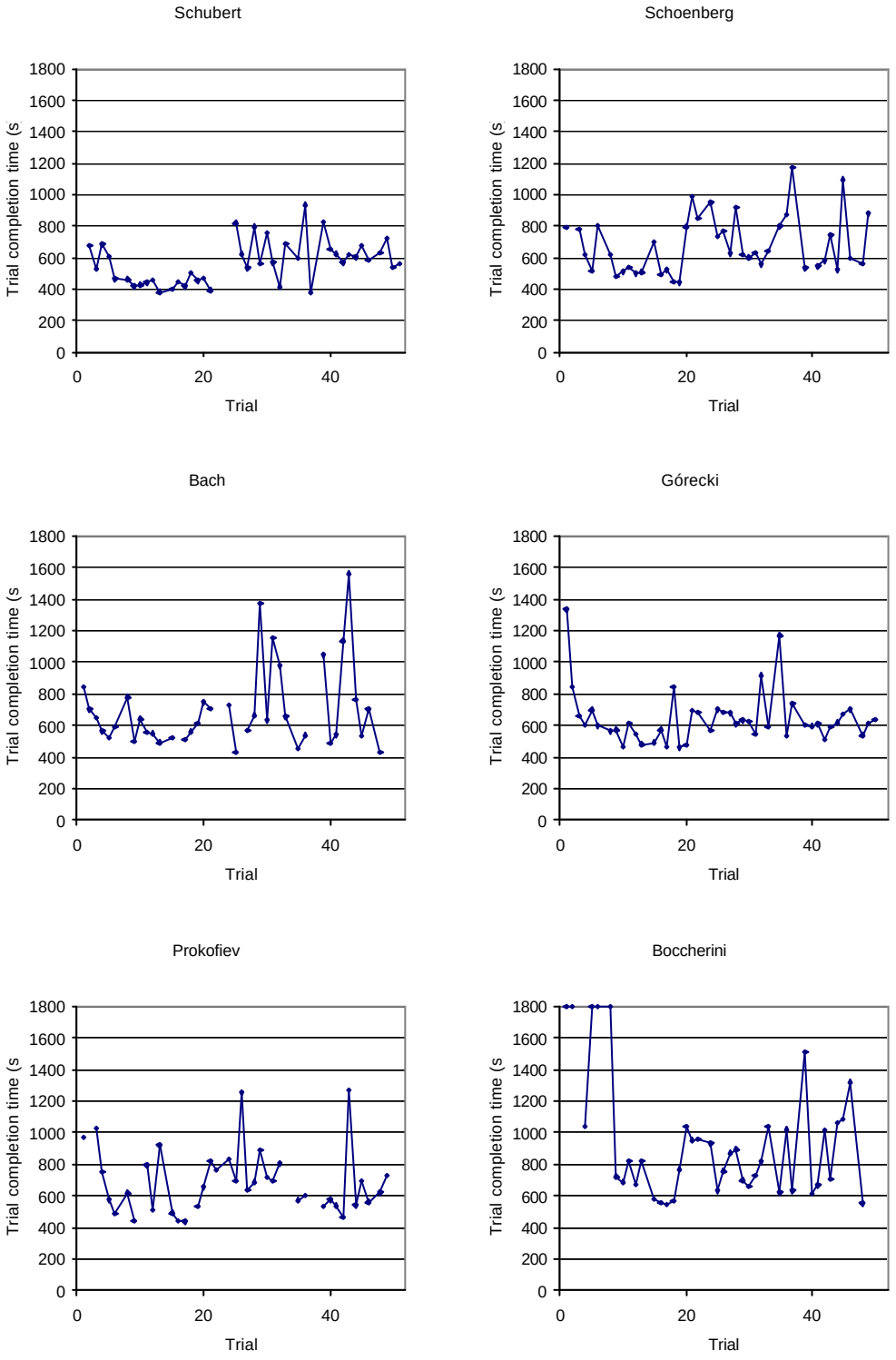


Figure 37. Individual practice curves for the Divided interface group.

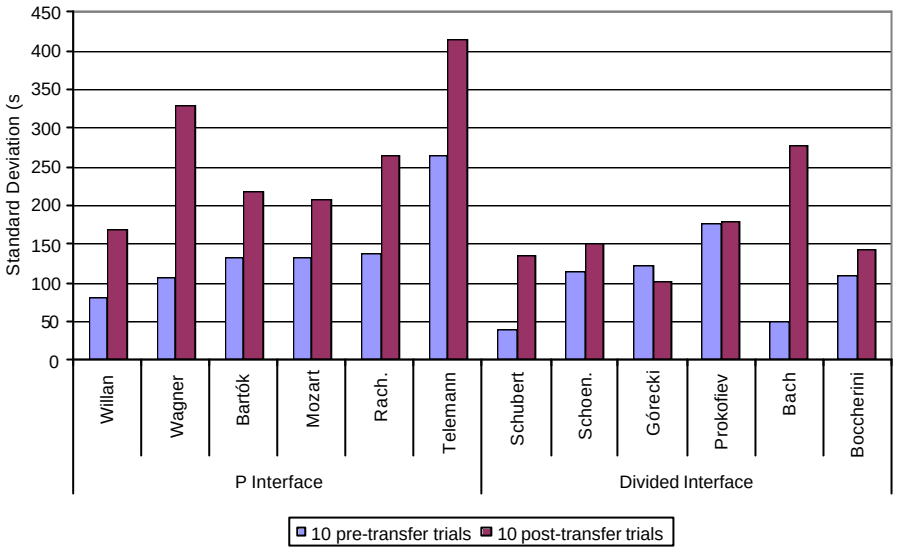


Figure 38. Comparison of variance in trial completion time for 10 trials immediately prior to and post transfer.

In order to determine if there were any dissimilarities between the two groups in the pre-transfer period, a general linear model was constructed on the trial completion time data (Table 24). Since there is no interface effect, this analysis confirms that the two groups performed similarly in the pre-transfer phase, as expected (see Figures 34 and 35). There was also a significant trial effect, which when viewed with the individual practice curves, indicates that participants' performance improved over the first 23 trials.

Table 24. GLM on trial completion time data for pre-transfer trials.

Source	DF	SS	MS	F	p
Interface	1	45687.392	45687.392	0.09	0.765
Participant(Interface)	10	4832313.446	483231.345		
Trial	18	4984005.266	276889.181	6.52	< 0.001
Interface × Trial	18	562713.518	31261.862	0.74	0.770
Participant × Trial (Interface)	152	645759.124	42465.521		

A second GLM was constructed to determine the effect of interface on performance in the post-transfer phase (Table 25). This GLM reveals that while there was no interface effect, there was an interaction between interface and trial. To understand this interaction, the performance

data for each participant in both groups was averaged for each post-transfer trial, and is plotted in Figure 39 (using a 2-period moving average to smooth the data). What can be seen here is a crossover effect: the participants who transferred to the P interface performed poorly in early trials, and then improved dramatically. The participants on the divided interface initially performed better than their counterparts, but by the end of the phase their performance was somewhat worse. It is likely that the participants on the divided interface were not as affected by the interface transfer because the interface that they transferred to still contained all of the information that they were used to, but split over a number of windows. The P group, on the other hand, had to learn to get by with less information, and this dramatically affected their performance initially. In the long run, however, the P group adapted well to their new interface, and caught up to the divided group. The divided group ended up performing slightly more poorly than the P group most likely because switching screens on the divided interface incurs a time penalty that cannot be overcome.

The results thus far have had nothing to say about the effect of tool use on trial completion time. To determine if tool use had an effect on trial completion time, correlations were calculated between: average pre-transfer trial completion time and tool use ($R_{12} < -.08$, $p = .81$), average post-transfer trial completion time and tool use ($R_{12} < -.09$, $p = .78$), average pre-transfer trial completion time and tool use variability ($R_{12} = .11$, $p = .74$), and average post-transfer trial completion time and tool use variability ($R_{12} = .49$, $p = .10$). None of these correlations are strong enough to indicate a relationship between tool use and trial completion time. It is notable, however, that the two participants who used tools the least (Willan and Schubert) were the best performers in terms of trial completion time in their groups.

Table 25. GLM on trial completion time data for post-transfer trials.

Source	DF	SS	MS	F	p
Interface	1	45184.04	45184.04	0.12	0.734
Participant(Interface)	10	3687508.44	368750.84		
Trial	24	2463364.95	102640.21	2.29	< 0.001
Interface × Trial	24	1933521.27	80563.39	1.80	0.015
Participant × Trial (Interface)	230	10317384.30	44858.19		

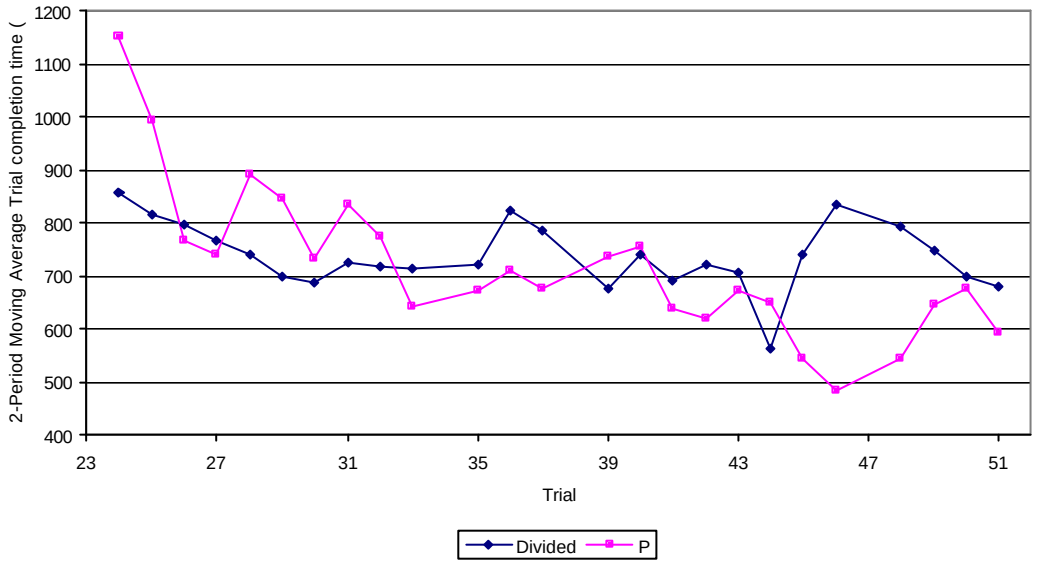


Figure 39. 2-period moving averages of the performance of all participants within each group.

Normal trial blowups. A final aspect of normal trial performance is the number of times participants violated the constraints of DURESS II and caused it to terminate abnormally, or blow up. The number of blowups for each participant in the pre- and post-transfer phases are listed in Table 26. Correlations were calculated between these data and their associated tool use and tool use variability counts, and a strong and significant positive correlation was found between post-transfer tool use and post-transfer blowups (R_{12} = 0.66, p = .02 — see Figure 40). Just as in the pilot study, the participants who used tools experienced a larger number of blowups, here in the post-transfer trials. However, in this case it should be noted that the

correlation depends on the data from one participant, Rachmaninov, who used many tools and experienced 3 post-transfer blowups. Recall that Rachmaninov persisted in marking the active portions of the feedwater streams — a tool that was most likely an aid to learning — for the entire experiment. It is likely that this tool use either became a rote action with little motivation or that Rachmaninov was attempting to produce 'good' data. As a result, this data point can be considered to be an outlier. Without this datum, the strength and significance of this correlation drops — $R_{11} = 0.26$, $p = .44$. Because it is based on one potentially outlying data point, this is not a strong finding.

Summary. The strongest finding from these analyses of normal trial performance is that there was no systematic correlation between tool use and performance. Tool use and performance in terms of trial completion time and number of blowups are not correlated. Hence, tool use does not seem to have an effect on performance under normal conditions. In addition, these data confirm that the interface transfer perturbed the participants' performance by negatively impacting the performance of both groups.

Table 26. Number of normal trial blowups.

	P Interface						*Divided Interface*					
	Willan	**Wagner**	**Mozart**	**Bartók**	**Rachmaninov**	**Telemann**	**Schubert**	**Schoenberg**	**Bach**	**Górecki**	**Prokofiev**	**Boccherini**
Pre-transfer	2	4	5	3	2	5	1	3	4	1	5	4
Post-transfer	0	0	0	0	3	0	1	1	2	0	2	0
	Pre 3.5 / Post 0.5						Pre 3 / Post 2					

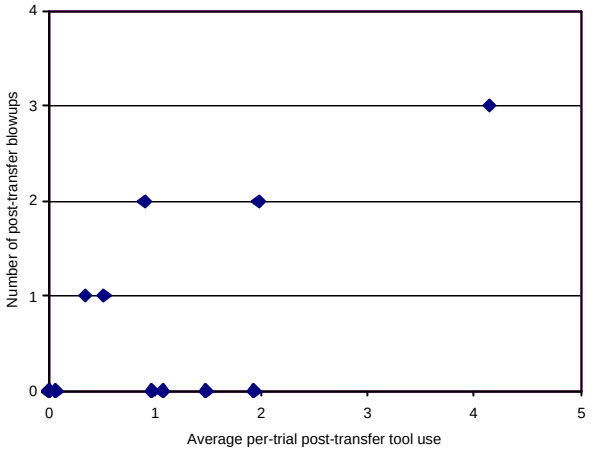


Figure 40. Plot of average per trial post-transfer tool use vs. number of post-transfer blowups.

Results: Fault Trials

Analyses of participants' fault performance for were conducted for the measures of fault detection time, number of faults detected, fault diagnosis time, number of faults diagnosed, accuracy of diagnosis, and compensation time. Since these analyses revealed no systematic differences between groups and no correlation between tool use and performance on fault trials, they are not presented here, but can be found in Appendix D.

Results: Control Recipes

The control recipes completed by participants periodically over the course of the experiment were analysed using a battery of methods adopted from Christoffersen, et al. (1998). These analyses included a tallying of the number of statements in participants' control recipes under three broad categories: length and chunking, differentiation, and knowledge organisation. Measures of length and chunking included counts of the number of words and steps in each recipe as well as the number of words per step. Measures of differentiation included counts of the number of references to the steady state heater ratios, the number of references to a decoupled feedwater stream configuration, and the number of asymmetrical statements which

revealed participants' understanding of the different characteristics of the two reservoirs. Measures of knowledge organisation included counts of the number of explicit references to the perceptual features of the display, statements specifying precise quantities for component settings, statements of declarative knowledge, and statements justifying or explaining recipe steps. Three more measures were developed to address the issues raised during the pilot investigation analyses of tool use: number of references to tool use, number of pictures drawn, and number of statements of warning about the work-domain constraints.

While these analyses were expected to provide some insight into the effect of tool use and operator modifications on system understanding, this was not realised. Accordingly, these analyses are not reproduced here. The only result of note was that Wagner and Rachmaninov referred to their tool uses of deriving flows and deriving heater settings, respectively. That they perceived these to be important instructions to pass on to a novice user indicates the value that they placed in these modifications. (These analyses are reproduced in Appendix E.)

That few significant patterns emerged from these analyses is not surprising. Christoffersen et al. (1998) only found significant results after five months of experience with DURESS, a much longer period of time than the one-month span of this experiment.

Unfortunately, this calls into question the results on control recipes found during the pilot study. In that study, it was found that the participants who tended to use tools also tended to write longer control recipes that included both warnings about the work-domain constraints and pictures. In view of these new data, it is likely that the results of the pilot study are best attributed to some factor not measured, and are not the result of tool use.

Results — Abstraction Hierarchy Analyses

Since it is possible that tool use might be correlated with different strategies for operating

DURESS II, a number of measures developed by Yu, Chow, Jamieson, Khayat, Lau, Torenvliet, Vicente, and Carter (1997) were applied to participants' data to investigate this possibility. These measures Rasmussen's (1986) abstraction hierarchy as developed for DURESS II (Bisantz & Vicente, 1994; Vicente & Rasmussen, 1990) to define four multiple, complementary frames of reference in which changes in operator behaviour or work domain state can be plotted as trajectories over time. Using these trajectories, variability in operator behaviour and work domain state can be calculated across trials at the levels of:

- Functional purpose / system (outputs to the environment)
- Abstract function / subsystem (mass and energy topologies)
- Generalized function / component (liquid flow and heat transfer rates)
- Physical function / component (component settings)

Accordingly, measures of output variability, mass and energy variability, liquid flow and heat transfer variability, and component setting variability can be calculated for each participant across trials. Previous research (Yu et al., 1997) has shown that these measures may point to differences in strategies between participants.

These measures were applied to the data from this experiment. The normal trials of each participant were divided into four blocks – trials 1 – 11 (10 normal trials), trials 12-23 (10 normal trials), trials 24-34 (10 normal trials), and trials 35-52 (15 normal trials). Blocks 1 and 2 contained pre-transfer trials, and blocks 3 and 4 contained post-transfer trials.

These measures were not able to reveal any differences between interface groups in the post-transfer period, as well as no consistent affect of the interface transfer on performance within groups (these data can be found in Appendix F). To see if any of these variance measures were correlated with either total tool use or tool use variability, average tool use and tool use

variability were calculated for each participant for each block of trials. Correlations between the tool use measures and abstraction hierarchy variance were calculated for each block. Three strong correlations were found:

- For block 3, output variance was strongly and positively correlated with variance in tool use (R_{12} = .90, p < .0001).
- For block 3, mass and energy variance was strongly and positively correlated with both average tool use (R_{12} = .63, p = .03) and tool use variance (R_{12} = .75, p = .005).

Unfortunately, these correlations are isolated to one block of trials and do not point to any patterns across blocks. Further, these three correlations were gleaned from a pool of 32 correlations. In order for the overall rate of type I error to be .05, individual correlations need to achieve a p-value of .05/32, or .002. Viewed in this way, only one of the correlations was significant, and one significant correlation across 32 does not point to any meaningful trend.

That these measures have revealed no differences between groups or correlation to tool use does not necessarily imply that these effects do not exist. Yu et al. (1997) only found consistent differences between participants after approximately 160 trials. It is possible that the current experiment did not give participants enough practice for any differences to become stable.

DISCUSSION

From the outset, this exploratory research has been guided by four questions. This section will revisit these questions, and will summarise the answers and directions that were found. The questions are repeated below:

- Why do operators modify their interfaces?
- How do operator modifications develop over time?
- Do operator modifications affect task performance?
- Do operator modifications affect understanding?

The contributions of this research in answering these questions are discussed below.

Why do operators modify their interfaces?

The first question guiding this research speaks to the underlying purpose of tool use. Kirsh (1995) has already spoken of modifications with respect to their cognitive and perceptual ends, but the effort here was to answer this question in the context of the operation of process control microworld. Fortunately, this research was able to answer this question quite thoroughly.

The results from the pilot study already indicated a difference between the logs of the participants on the P and P+F interfaces: participants on the P interface tended to write logs at a lower-level, while participants on the P+F interface wrote logs at a higher-level. Unfortunately, because this study did not generate a great deal of other tool uses, it was not clear if this interface difference would hold for other types of modifications as well. Further, it was not clear if these modifications were efforts to 'finish the design' or if they were simply reflections of the information contained on both interfaces.

The second experiment was able to address this question more fully. Participants engaged in in a wide variety of tool uses, and this provided ample data for investigating the motivation behind tool use. The patterns in participants' tool use led to the development of a taxonomy of

four different motivations for tool use:

- Tool use as an aid to learning. In early experimental trials, participants were observed using tools to help them in learning and exploring DURESS II. Some participants used their logs to discuss various work-domain characteristics and configuration strategies, others used a straightedge and a magnifier to help in attuning themselves to the perceptual features of the interface, and still others went through the exercise of writing the names of the various components on-screen. While there were many idiosyncrasies in these tool uses, they each had one characteristic in common: they had only a limited use, and so were discarded as soon as they were no longer perceived to be useful. As participants became more attuned to DURESS II and the visual and informational content of the various interfaces, this category of tool use faded.

  In the context of the literature review, this finding is notable. All of the research reviewed there discussed operator modifications in the context of expert behaviour. This category of tool use, on the other hand, shows that tool use can also have an important function in the development of expert behaviour. The tools used in this developmental stage may not be as sophisticated as some of the other tool uses observed, but they nevertheless served the important cognitive function of facilitating attunement to the work-domain and interface.

- Tool use to aid in information integration. A second important function of tools was to help participants on the divided P+F interface compensate for the division of information across the various levels of their display. Faced with a new interface that structured information more poorly than the P+F interface they were used to, a number of participants worked to finish (or perhaps, restore) the design back to the format they were accustomed to. It is important to note that the hallmark of this type of tool use was integration, and this integration

pointed at precisely the deficiency of the divided interface when compared to the P+F interface.

- Tool use to aid in deriving information about the system state. Participants who transferred to the P interface engaged in the same type of activity as their divided interface counterparts, except that their efforts to 'restore the design' were uniquely tailored to the new demands placed on them by the P interface. Rachmaninov's and Mozart's efforts to learn the steady state heater ratios and Wagner's efforts to derive flow values for each of the valves addressed exactly the types of task support that were lost in the interface transfer. The hallmark of this type of activity was the derivation of system parameters. Again, this mirrors precisely the support that was lost in the transfer from the P+F to the P interface.
- Tool use to extend the functionality of the interfaces. In addition to tool uses aimed at restoring information that participants were accustomed to, a number of tool uses were also observed that extended the functionality of the various interfaces. The setting memories developed were intelligent solutions to the problem of adjusting the components to keep all goal variables within their limits, the reservoir volume alarms were inventive ways to reduce the amount of monitoring necessary to determine if the volume of a reservoir was changing, and the use of the stopwatch was a simple but effective way to simplify the task of keeping time. All of these tool uses were simple, but highly effective.

Moving away from this specific experimental context, this research indicates three purposes for tool use in complex systems. First of all, tool use can be an important aid to system learning. Second, tool use can help operators to restore beneficial functio ns that they had become used to. Third, tools can be used to extend the functionality of the interface in order to make work more effective or more efficient.

In retrospect, these categories were already identified in the literature review. Cook and Woods (1996) dealt with the restorative function of tool use, as the anaesthesiologists worked to modify their new interface so that it would display forms they were accustomed to in their old interface. Henderson and Kyng (1991), Kirsh (Kirsh, 1995), Hutchins (Hutchins, 1995), Vicente and Burns (Vicente & Burns, 1995), and Vicente et al. (Vicente et al., 1997) have all dealt with the use of tools to extend the functionality of some device. Finally, Hammond et al. (Hammond et al., 1995) implicitly recognize the value of tools to learning by speaking of a 'process of stabilization' in which actor environment are tailored by one another. However, this is the first time that these different purposes of operator modifications have been observed in the laboratory and structured with respect to one another.

How do operator modifications develop over time?

This research was also able to provide a strong answer to this question. Operator modifications have been shown to follow a fairly predictable trajectory that starts with a period of exploration that features high tool use variability and eventually settles into stable patterns that have low tool use variability. It is notable that this pattern was independent of the amount of tool use engaged in by participants: whether participants used many or few tools, they generally exhibited exploration and stability at different times in the experiment.

The results on the amount of tool use between participants were more difficult to interpret. This research indicated no significant difference in the amount of tool use engaged in under the three interfaces studied. Even if possibilities for tool use exist, different individuals will explore these possibilities in different ways. In the end, the amount of tool use engaged in is probably only partially determined by the number of possibilities for tool use that exist. A number of results pointed to the likelihood that tool use is connected to a participants' predisposition to self-

explain his or her actions (Howie & Vicente, 1998). These two phenomena, in turn, might be connected to some underlying individual difference.

Do operator modifications affect task performance?

The pattern of null results presented in connection with performance on normal and fault trials is important. Some might predict that tool use would distract operators, and so negatively impact their performance. Since neither a negative nor a positive correlation could be found between tool use and performance, this research provides some preliminary evidence to disagree with this assertion. While the lack of any effect was likely partially due to a lack of experimental power, these results indicate that any correlations that exist are likely not strong.

Further studies will have to be performed to determine if Kirsh (1995) is correct in stating that tool use actually improves performance under diverse task situations. The possibility of this being true is still open; a more sensitive experiment is needed.

Do operator modifications affect understanding?

Unfortunately, this research was not able to this question. The fact that the performance of tool using participants in the pilot study did not degrade dramatically when their tools were taken away implies that tools are not a cognitive crutch, but more sensitive measures are needed to determine if and how tools affect understanding.

CONCLUSIONS

Contributions

This first laboratory investigation of operator modifications has been fruitful. In the context of a series of exploratory investigations designed to investigate a number of questions about tool use in a preliminary way, a number of important results have been found. The most important contribution of this research has been to identify a taxonomy that considers the purpose and motivation behind operator modifications. Operators use tools to attune themselves to the properties of a system (i.e., learning), to extend the functionality of the interface that they are given, and — in the context of a transfer to a new interface — to restore the useful functionality of the old interface. In terms of performance, operator modifications and performance have been shown to not have a strong correlation. Finally, compelling evidence has been presented to make good on Kirsh's (1995) claims that operator modifications achieve their ends with intelligence, elegance, and simplicity.

Unfortunately, this research was not able to address the impact of operator modifications on understanding, except in a very preliminary way. More sensitive measures of understanding are needed. While it is clear that tool use does not degrade understanding, little has been said about how tools might improve it.

Nonetheless, this research is a strong first step toward understanding operator modifications. The taxonomy proposed above is a framework for future experimentation to examine each of the different types of operator modifications. Experimentation in more complex situations with more sensitive measures is needed to determine the extent to which modifications are correlated with performance and understanding.

This research has also made two methodological contributions. First of all, a novel measure has been proposed to account for the variability in tool use. Second, this research has

successfully applied the results of Torenvliet et al. (in press) to the selection and matching of experimental participants. In this study, the impact of individual differences on performance were balanced between groups, indicating that the holist-serialist cognitive style distinction is a useful way to balance experimental groups in the context of the DURESS II microworld.

Limitations and Future research

In spite of the useful results that were obtained, this research has also suffered from a number of limitations. The most notable issue is bias. In a control room operators have to look around for the tools needed to modify their interfaces. It seems likely that task demands suggest a need, and then operators supply the tools. In this investigation, the tools were supplied leaving operators to look for a need. This may have led to an over-application of tools, especially in the case of Rachmaninov, who continued to mark the active feedwater stream perhaps for no other reason other than because he thought he had to use tools. The idea in this research was to remove this bias by providing a large variety of tools, many of which seemed spurious (e.g., paperclips), but this method has drawbacks. Unfortunately, in the absence of a theory dictating which tools are useful under different settings, solving this issue in the laboratory is a difficult problem.

This research also only considered the effects of an interface transfer in one direction: from the information-rich P+F interface to the qualitatively poorer P and divided P+F interfaces. To help in understanding which of the tool uses were related to the order of the interface transfer as opposed to the qualities of the interfaces themselves, a similar experiment should be conducted to explore the effects on tool use of transferring from the P or divided P+F interfaces to the P+F interface.

In view of the large individual differences in tool use that were observed, future experiments

into tool use might be able to employ a psychometric test to help in selecting participants that can be expected to use tools similarly. The anthropology community has produced some research that could help in this area (C. Burns, personal communication, August 9, 1999); this research should be explored to inform future experimentation.

In addition, this research addressed the specific case of hard-wired displays. Even though DURESS II is implemented on a computer workstation, it is essentially a hard-wired interface that requires no interface navigation. The divided P+F interface does require navigation, but even that is a rather simple case of four screens. It remains to be understood how operators go about resolving the intrinsic degrees of freedom in a display while at the same time making extrinsic modifications. Understanding this interaction is integral to scaling this research concept up to the designs of contemporary and future control rooms.

A similar issue relates to the nature of the tools employed in making modifications. This research investigated the case of tangible, 'hard' tools. It is unclear if these results would scale to 'soft' tools that can be implemented in computer applications. Many interfaces are designed to include 'soft' tailorability (cf. Henderson & Kyng, 1991), and it is possible that these tools are used differently and have a different impact on performance than 'hard' tools.

Finally, this investigation has treated tool use as an individual phenomenon, while in most situations in complex systems it is a social phenomenon. The way that teams develop and settle on a set of tools is still an open research question. Admittedly, the ways that teams interact in complex sociotechnical systems is itself a relatively open issue (although there is a literature on this topic, more is needed, cf., Vicente, 1999). New experimental methods are needed to investigate this phenomenon.

REFERENCES

Becker, G. (1991). Analysis of human behaviour during NPP incidents: A case study, Balancing automation and human action in nuclear power plants (pp. 517-526). Vienna, Austria: IAEA.

Bisantz, A. M., & Vicente, K. J. (1994). Making the abstraction hierarchy concrete. International Journal of Human-Computer Studies, 40, 83-117.

Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1996). A longitudinal study of the effects of ecological interface design on skill acquisition. Human Factors, 38, 523-541.

Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1997). A longitudinal study of the effects of ecological interface design on fault management performance. International Journal of Cognitive Ergonomics, 1, 1-24.

Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1998). A longitudinal study of the effects of ecological interface design on deep knowledge. International Journal of Human-Computer Studies, 48, 729-762.

Cook, R. I., & Woods, D. D. (1996). Adapting to new technology in the operating room. Human Factors, 38, 593-613.

Flach, J., Hancock, P., Caird, J., & Vicente, K. J. (Eds.) (1995). Global perspectives on the ecology of human-machine systems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Gibson, J. J. (1979/1986). The ecological approach to visual perception. Mahwah, NJ: Lawrence Erlbaum Associates.

Giraudo, M. D., & Pailhous, J. (in press). Dynamic instability of visuo-spatial images. Journal of Experimental Psychology: Human Performance and Perception.

Hammond, K. J., Converse, T. M., & Grass, J. W. (1995). The stabilization of environments. Artificial Intelligence, 72, 305-327.

Hancock, P., Flach, J., Caird, J., & Vicente, K. J. (Eds.) (1995). Local applications of the ecological approach to human-machine systems. Hillsdale, NJ: Lawrence Erlbaum Associates.

Henderson, A., & Kyng, M. (1991). There's no place like home: Continuing design in use. In J. Greenbaum & M. Kyng (Eds.), Design at work: Cooperative design of computer systems (pp. 219-240). Hillsdale, NJ: Lawrence Erlbaum Associates.

Howie, D., & Vicente, K. J. (1998). Making the most of ecological interface design: The role of self explanation. International Journal of Human-Computer Studies, 49, 651-674.

Howie, D. E., Janzen, M. E., & Vicente, K. J. (1996). Research on factors influencing cognitive behaviour (III) (CEL 96-06). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Hunter, C. N., Janzen, M. E., & Vicente, K. J. (1995). Research on factors influencing human cognitive behaviour (II) (CEL 95-08). Toronto: University of Toronto, Cognitive Engineering Laboratory.

Hutchins, E. (1995). Cognition in the wild. Cambridge, MA: MIT Press.

Irmer, C., & Reason, J. T. (1991). Early learning in simulated forest fire-fighting. Paper presented at the Simulations, Evaluations, and Models: Proceedings of the Fourth MOHAWC Workshop, Roskilde, Denmark.

Janzen, M. E., & Vicente, K. J. (1998). Attention allocation within the abstraction hierarchy. International Journal of Human-Computer Studies, 48, 521-545.

Kirk, R. E. (1995). Experimental design: Procedures for the behavioral sciences. (Third ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Kirsh, D. (1995). The intelligent use of space. Artificial Intelligence, 73, 31-68.

Lave, J. (1988). Cognition in practice. New York: Cambridge University Press.

Law, A. M., & Kelton, W. D. (1991). Simulation modelling & analysis. New York: McGraw-Hill.

Malone, T. B., Kirkpatrick, M., Mallory, K., Eike, D., Johnson, J. H., & Walker, R. W. (1980). Human factors evaluation of control room design and operator performance at Three Mile Island-2 (NUREG/CR-1270). Washington, DC: USNRC.

Mumaw, R. J., Woods, D. D., & Eastman, M. C. (1992). Interim report on techniques and principles for computer-based display of data (STC Report 92-ISJ3-CHICR-R1). Pittsburgh, PA: Westinghouse Science and Technology Center.

Pask, G., & Scott, B. C. (1972). Learning styles and individual competence. International Journal of Man-Machine Studies, 4, 217-253.

Pawlak, W. S., & Vicente, K. J. (1996). Inducing effective operator control through ecological interface design. International Journal of Human-Computer Studies, 44, 653-688.

Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Transactions on Systems, Man, and Cybernetics, SMC-13, 257-266.

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision making and system management. IEEE Transactions on Systems, Man, and Cybernetics, 15, 234-243.

Rasmussen, J. (1986). Information processing and human-machine interaction: An approach to cognitive engineering. Amsterdam: North-Holland.

Seminara, J. L., Gonzalez, W. R., & Parsons, S. O. (1977). Human factors review of nuclear power plant control room design (EPRI NP-309). Palo Alto, CA: EPRI.

Suchman, L. (1987). Plans and situated actions. Cambridge: Cambridge University Press.

Torenvliet, G. L., Jamieson, G. A., & Vicente, K. J. (in press). Making the most of ecological interface design: Is operator selection necessary? Applied Ergonomics.

van Foerster, H. (1984). Observing systems. Seaside, CA: Intersystems Publications.

Vicente, K. J. (1991). Supporting knowledge-based behaviour through ecological interface design. Unpublished Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Vicente, K. J. (1997a). Heeding the legacy of Meister, Brunswik, and Gibson: Toward a broader view of human factors research. Human Factors, 39, 323-328.

Vicente, K. J. (1997b). Operator adaptation in process control: A three-year research program. Control Engineering Practice, 5, 407-416.

Vicente, K. J. (1999). Cognitive work analysis: Toward safe, productive, and healthy computer-based work. Mahwah, NJ: Lawrence Erlbaum Associates.

Vicente, K. J., & Burns, C. M. (1995). A field study of operator cognitive monitoring at Pickering nuclear generating station - B (CEL 95-04). Toronto: University of Toronto.

Vicente, K. J., Christoffersen, K., & Pereklita, A. (1995). Supporting operator problem solving through ecological interface design. IEEE Transactions on Systems, Man, and Cybernetics, SMC-25, 529-545.

Vicente, K. J., Mumaw, R. J., & Roth, E. M. (1997). Cognitive functioning of control room operators: Final phase (CEL 97-01). Toronto: University of Toronto.

Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems II: Mediating "direct-perception" in complex work domains. Ecological Psychology, 2, 207-249.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. IEEE Transactions on Systems, Man, and Cybernetics, 22, 589-606.

Wickens, C., & Carswell, C. M. (1995). The proximity compatibility principle: Its psychological foundation and relevance to display design. Human Factors, 37, 473-494.

Yu, X., Jamieson, G. A., Khayat, R., Lau, E., Torenvliet, G. L., Vicente, K. J., & Carter, M. W. (1997). Research on the characteristics of long-term adaptation (CEL 97-04). Toronto: University of Toronto.

APPENDIX A: TRIAL SCHEDULE — PILOT STUDY

Day	Trial	D1	D2	Event	Description
0				SRT Test Day	Introduction to research Consent form (SRT) SRT
1				Introduction	Demographic questionnaire Technical description of DURESS II List of variable names / pretest Schedule / Data Card
2				Introduction, cont'd	Interface description (P or P+F) *Intro to Recipe Test* Recipe test 1
3	1	8	2	Trials begin	
	2	7	4		
	3	3	9		
4	4	3	16	Comparison RF1	
	5	3	16	Routine fault RF1	VA1 blockage at 4 min
5	6	2	5		
	7	8	7		
	8	4	3	Comparison RF2	Recipe Test 2
6	9	4	3	Routine fault RF2	HF1 to 50% at 1 min
	10	5	10		
7	11	7	11		
	12	6	8	Comparison NRF1	
	13	2	6		
8	14	6	8	Non-routine fault NRF1	RL1 –4 at 4 min; EH +.478e6 R1 at 6 min + PH
	15	5	12		
	16	2	14		
9	17	3	12	Comparison RF3	
	18	5	2		
	19	3	12	Routine fault RF3	VA2 sticks at setting of 10 at 2 min
	20	5	10		
10	21	2	5		
	22	4	11	Comparison RF4	
	23	1	18		
	24	6	12		
11	25	4	11	Routine fault RF4	RL2 –4 at 2 min
	26	5	13		
	27	4	8		
	28	8	5		
12	29	9	3		
	30	4	11		
	31	3	8		
	32	4	6		
	33	1	10		Recipe Test 3
13	34	4	12		
	35	2	9		
	36	2	7		
	37	1	6		
	38	5	5		

Day	Trial	D1	D2	Event	Description
14	39	3	4		
	40	3	11		
	41	4	9	Comparison RF5	
	42	5	2		
	43	3	4		
15	44	4	8		
	45	6	6		
	46	4	9	Routine fault RF5	VB2 blockage at 5 min
	47	8	4		
16	48	6	2		
	49	2	8		
	50	1	18		
	51	6	9		(comparison NRF2)
	52	5	11		
17	53	6	9	Non-routine fault NRF2	RL1 –2 into R2 at 3 min
	54	3	4		
	55	5	6		
	56	5	11		
	57	8	6		
18	58	3	16		
	59	4	11		
	60	3	15		
	61	4	3		(comparison RF6)
	62	4	3	Routine fault RF6	HF1 to 50% at 1 min (same as trial 9)
19	63	4	8		(comparison NRF3)
	64	4	8	Non-routine fault NRF3	HF1 to 150% at 3 min; input to 20°C at 5 min
	65	6	11		
	66	6	7		(comparison RF7)
	67	6	7	Routine fault RF7	*RL2 –7 at 4 min* Recipe Test 4
20	Tr01-59	4	11		
	Tr02-58	3	16		
	Tr03-61	4	3		(comparison Tr-RF1)
	Tr04-56	5	11		
	Tr05-52	5	11		
21	Tr06-64	4	8	Non-routine Fault Tr-NRF1	HF1 to 150% at 3 min; input to 20°C at 5 min
	Tr07-54	3	4		
	Tr08-63	4	8		(comparison Tr-NRF1)
	Tr09-66	6	7		(comparison Tr-RF2)
	Tr10-53	6	9	Non-routine Fault Tr-NRF2	RL1 –2 into R2 at 3 min
22	Tr11-55	5	6		
	Tr12-48	6	2		
	Tr13-51	6	9		(comparison Tr-NRF2)
	Tr14-62	4	3	Routine Fault Tr-RF1	HF1 to 50% at 1 min
	Tr15-67	6	7	Routine Fault Tr-RF2	RL2 –7 at 4 min
23	Tr16-57	8	6		
	Tr17-50	1	18		
	Tr18-49	2	8		
	Tr19-60	3	15		
	Tr20-65	6	11		*debriefing + Recipe Test 4*

APPENDIX B: TRIAL SCHEDULE — EXPERIMENT

Day	Trial	Type	Config No.	D1	D2	Description
1	--		–	--	--	Intro Session 1
2	--		–	--	--	Intro Session 2 + Recipe Test 1
3	1		1	8	2	
	2		2	7	4	
	3		3	3	9	
4	4		4	3	16	
	5		6	2	5	
	6		7	8	7	
5	7	Routine Fault	5	3	16	VA1 blockage at 4 min
	8		8	4	3	
	9		10	5	10	
6	10		11	7	11	
	11		12	6	8	
	12		13	2	6	Recipe Test 2
7	13		15	5	12	
	14	Routine Fault	9	4	3	HF1 to 50% at 1 min
	15		16	2	14	
8	16		17	3	12	
	17		18	5	2	
	18		20	5	10	
	19		21	2	5	
9	20		24	6	12	
	21		22	4	11	
	22		23	1	18	
	23	Non-Routine Fault	64	4	8	HF1 to 150% at 3 min; input to 20oC at 5 min
10	24		26	5	13	Interface Changeover
	25		27	4	8	
	26		28	8	5	
	27		29	9	3	
11	28		30	4	11	
	29		31	3	8	
	30		32	4	6	
	31		33	1	10	
12	32		34	4	12	
	33		35	2	9	
	34	Routine Fault	19	3	12	VA2 sticks at setting of 10 at 2 min + Recipe Test 3
13	35		36	2	7	
	36		37	1	6	
	37		38	5	5	
	38	Routine Fault	9	4	3	HF1 to 50% at 1 min (same as trial 13)
	39		39	3	4	

Day	Trial	Type	Config No.	D1	D2	Description
	40		40	3	11	
	41		41	4	9	
14	42		42	5	2	
	43		43	3	4	
	44		44	4	8	
	45		45	6	6	
	46		47	8	4	
15	47	Non-Routine Fault	53	6	9	RL1 -2 into R2 at 3 min
	48		48	6	2	
	49		49	2	18	
	50		50	1	9	
16	51		51	6	11	
	52	Routine Fault	67	6	7	RL2 -7 at 4 min + Recipe Test 4

APPENDIX C: TOOL USE PROFILES

Tool use profiles for each participant appear on the next pages.

Table 27. Tool use profile for Bartók (P interface).

Bartók (P)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																					
NB:Derivation																																																				
NB: Learning	•							•							•	•																																				
NB:Events																																																				
NB:Times																																																				
NB:Reason																																																				
Stickers:FWS																																																			
SWatch:Timing																																																			
Mag:Details																																																			
Post-it:Heater																																																			
Post/Gr:ResAlarm																																																			
Gr:CompNames																																																			
Gr:Flows																																																			
Edge:Lineup																																																			
Post/Gr:Setting																																																			
Flag:OutputT																																																			
Control Panel																																																				

Table 28. Tool use profile for Mozart (P interface).

Mozart (P)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																						
NB:Derivation																										•	•	•	•	•	•				•																			
NB: Learning	•				•		•	•																																													
NB:Events																																																						
NB:Times																																																						
NB:Reason																															•	•		•	•	•	•																
Stickers:FWS																																																					
SWatch:Timing		•				•		•			•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Mag:Details						•		•								•								•																												
Post-it:Heater																																																					
Post/Gr:ResAlarm																•	•	•	•	•	•	•	•	•																												
Gr:CompNames																																																					
Gr:Flows																																																					
Edge:Lineup		•						•				•	•																																								
Post/Gr:Setting																																																					
Flag:OutputT																																																					
Control Panel																																																					

Table 29. Tool use profile for Rachmaninov (P interface).

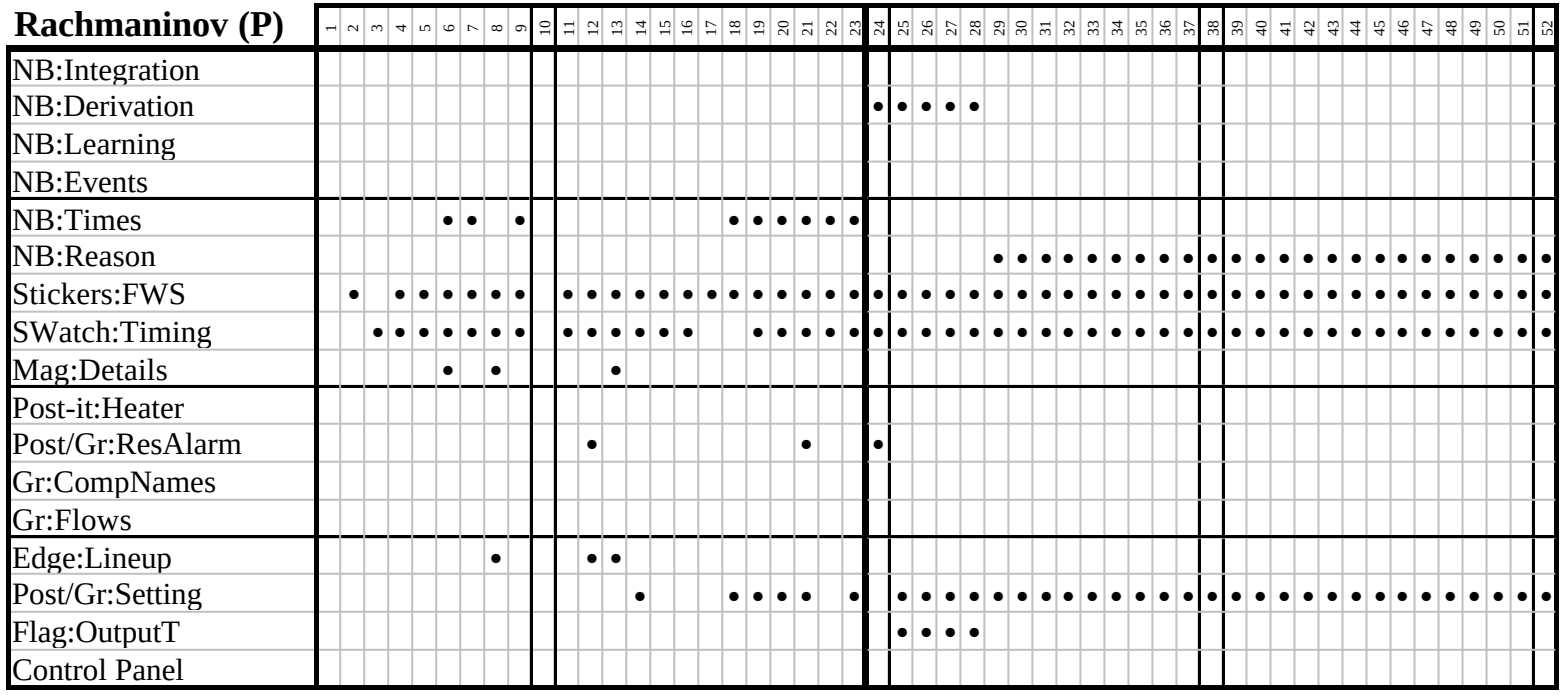


Table 30. Tool use profile for Telemann (P interface).

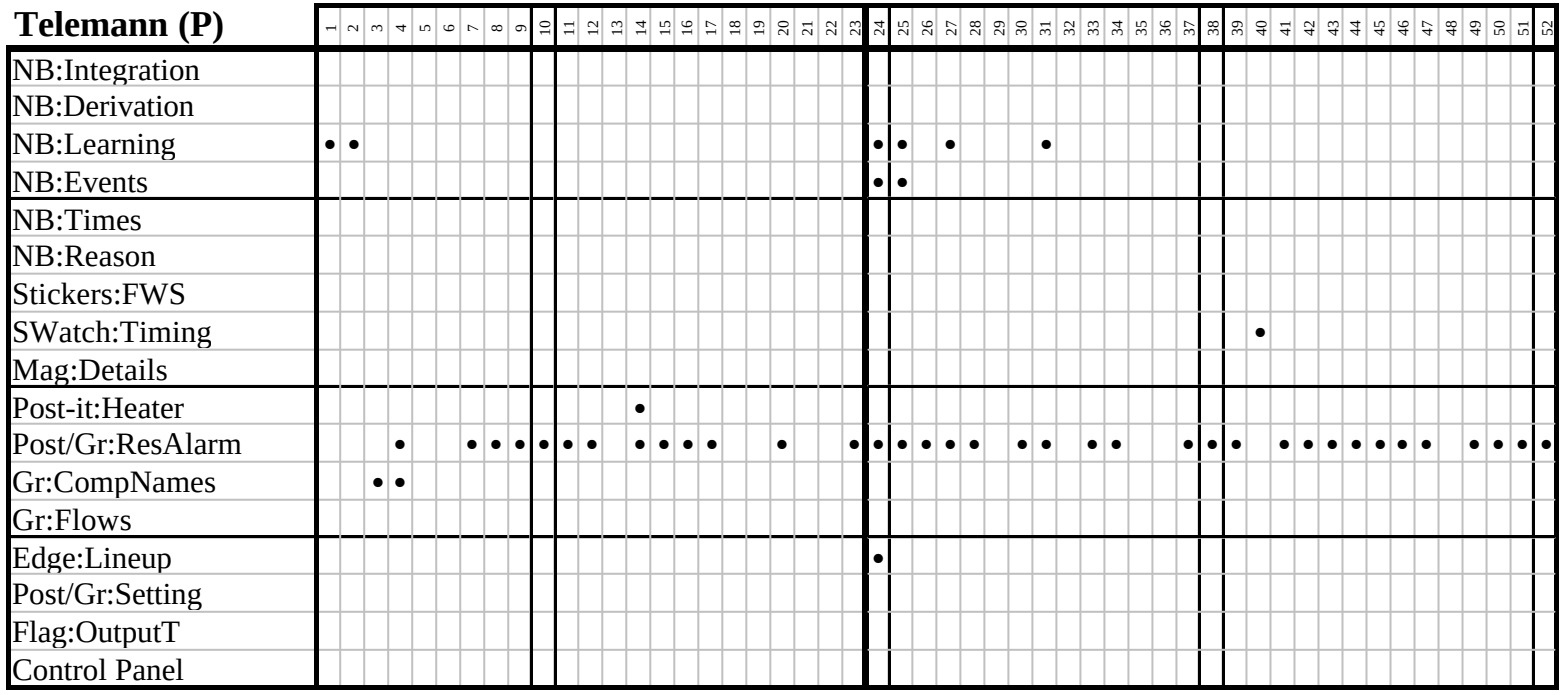


Table 31. Tool use profile for Wagner (P interface).

Wagner (P)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																				
NB:Derivation																																																				
NB:Learning																																																				
NB:Events																																																				
NB:Times																																																				
NB:Reason																																																				
Stickers:FWS																																																				
SWatch:Timing									•																																										
Mag:Details																																																				
Post-it:Heater																																																	•	•	•		
Post/Gr:ResAlarm																																																				
Gr:CompNames		•																																																	
Gr:Flows																												•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•
Edge:Lineup												•																																							
Post/Gr:Setting																																																			
Flag:OutputT																																																			
Control Panel																																																			

Table 32. Tool use profile for Willan (P interface).

Willan (P)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																					
NB:Derivation																																																				
NB:Learning																																																				
NB:Events																																																				
NB:Times																																																				
NB:Reason																																																				
Stickers:FWS																																																				
SWatch:Timing																																																				
Mag:Details																																																				
Post-it:Heater																																																				
Post/Gr:ResAlarm																																																				
Gr:CompNames																																																				
Gr:Flows																																																				
Edge:Lineup																																																				
Post/Gr:Setting																																																				
Flag:OutputT																																																				
Control Panel																																																				

Table 33. Tool use profile for Bach (divided interface).

Bach (Div)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																					
NB:Derivation																																																				
NB:Learning			•		•	•	•	•			•	•									•						•																										
NB:Events																																																				
NB:Times																																																				
NB:Reason																																																				
Stickers:FWS																																																				
SWatch:Timing																																																				
Mag:Details																																																				
Post-it:Heater																																																				
Post/Gr:ResAlarm																																																				
Gr:CompNames																																																				
Gr:Flows																																																				
Edge:Lineup																																																				
Post/Gr:Setting							•	•																																											
Flag:OutputT																																																				
Control Panel																												•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•	•

Table 34. Tool use profile for Boccherini (divided interface).

Boccherini (Div)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																					
NB:Derivation																																																				
NB:Learning				•		•									•							•				•																									
NB:Events																																																				
NB:Times																							•																													
NB:Reason																																																				
Stickers:FWS																																																			
SWatch:Timing																																																			
Mag:Details																																																			
Post-it:Heater																																																			
Post/Gr:ResAlarm																																																			
Gr:CompNames																																																			
Gr:Flows																																																			
Edge:Lineup																																																			
Post/Gr:Setting																																																			
Flag:OutputT																									•																										
Control Panel																																																			

Table 35. Tool use profile for Górecki (divided interface).

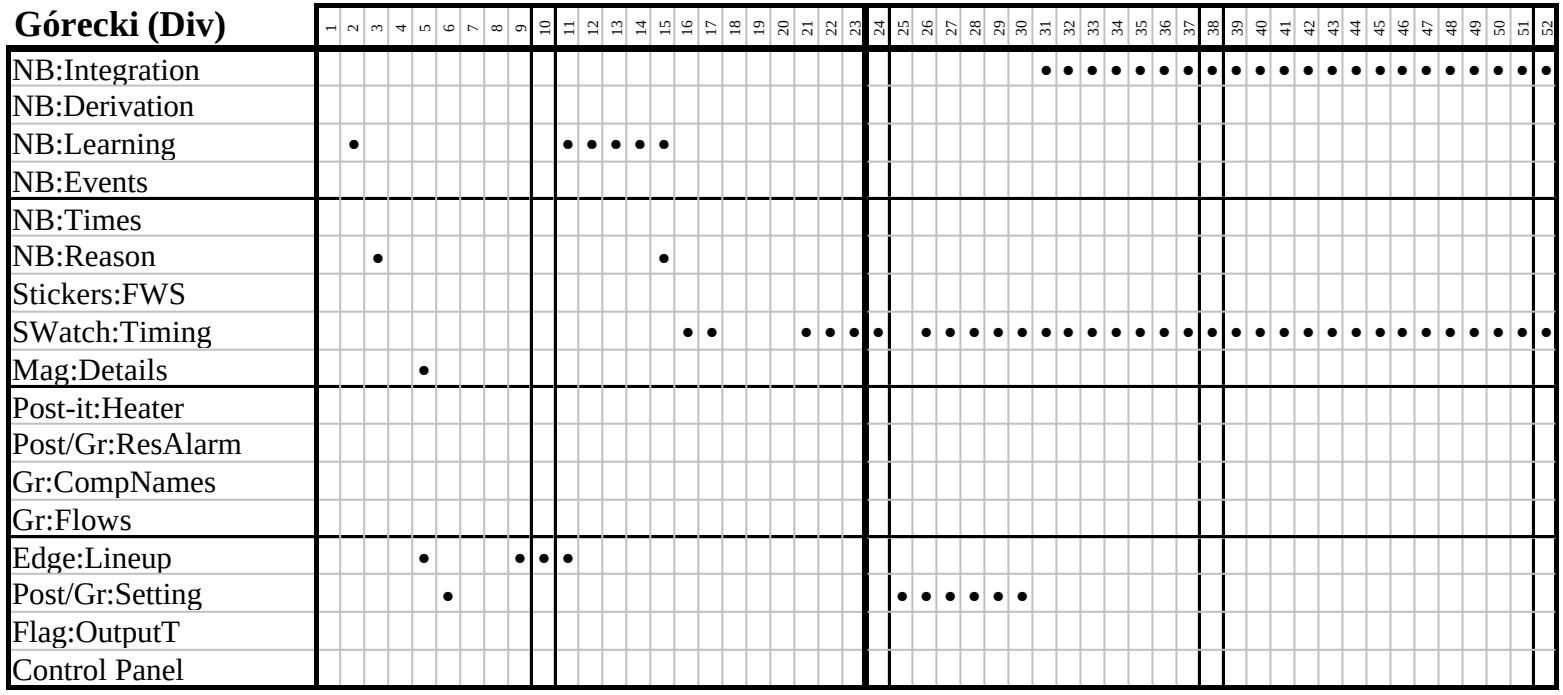


Table 36. Tool use profile for Prokofiev (divided interface).

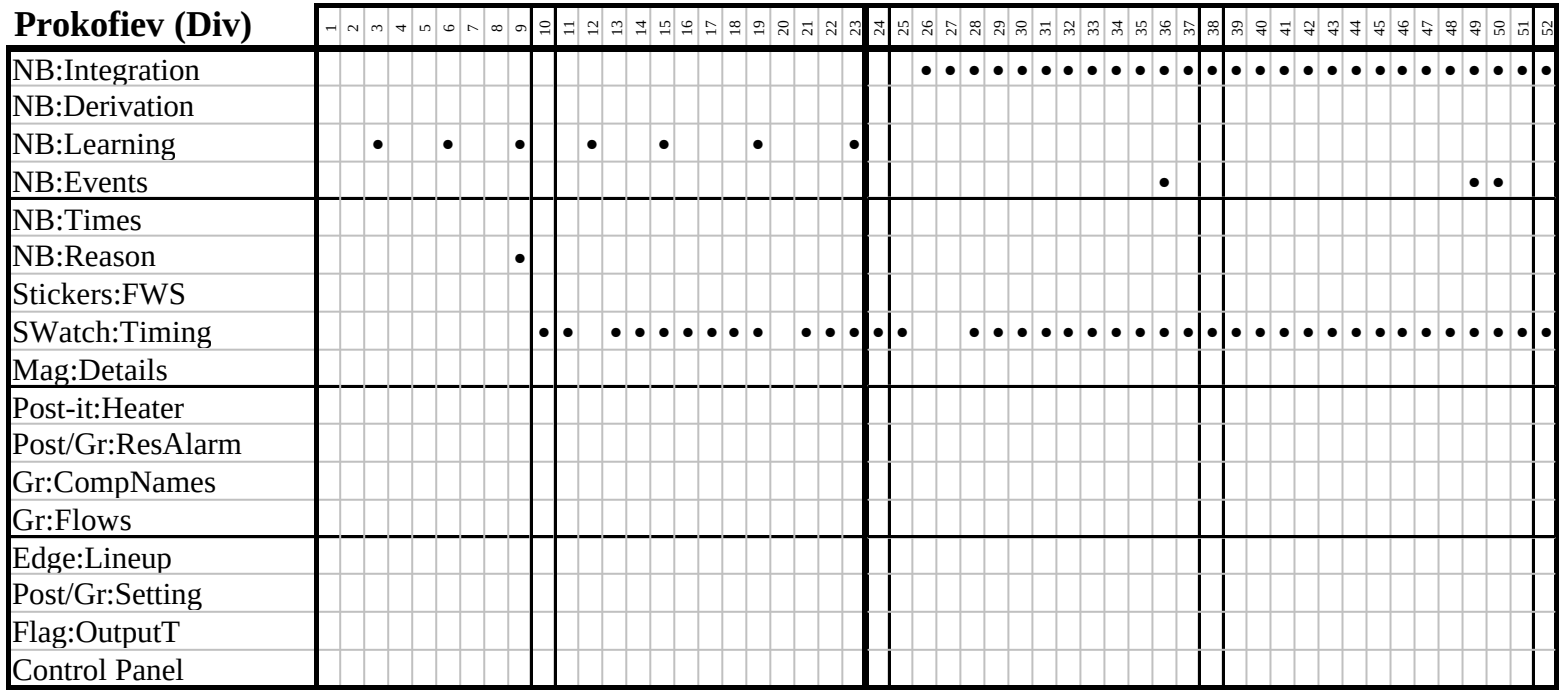


Table 37. Tool use profile for Schoenberg (divided interface).

Schoenberg (Div)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																				
NB:Derivation																																																				
NB:Learning																																																				
NB:Events																																																				
NB:Times																																																				
NB:Reason																																																				
Stickers:FWS																																																				
SWatch:Timing															•	•	•	•	•	•	•	•	•																									•				
Mag:Details																																																				
Post-it:Heater																																																				
Post/Gr:ResAlarm																																																				
Gr:CompNames																																																				
Gr:Flows																																																				
Edge:Lineup																																																				
Post/Gr:Setting																																																				
Flag:OutputT																																																				
Control Panel																																						•	•			•	•	•		•		•	•	•

Table 38. Tool use profile for Schubert (divided interface).

Schubert (Div)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52
NB:Integration																																																					
NB:Derivation																																																				
NB:Learning																																																				
NB:Events							•							•								•	•	•																												
NB:Times								•	•	•	•	•	•	•	•	•	•	•	•	•	•		•	•	•	•	•	•	•	•	•	•	•	•	•	•																
NB:Reason																																																			
Stickers:FWS																																																			
SWatch:Timing																																																			
Mag:Details																																																			
Post-it:Heater																																																			
Post/Gr:ResAlarm																																																			
Gr:CompNames																																																			
Gr:Flows																																																			
Edge:Lineup																																																			
Post/Gr:Setting																																																			
Flag:OutputT																																																			
Control Panel																																																			

APPENDIX D: FAULT ANALYSES

Fault data are presented below for three measures: detection, diagnosis, and compensation. At the beginning of the experiment, all participants were told that in the event that they encountered a fault, their task was to verbalise their detection and diagnosis, and then to compensate for the fault and reach steady state. Accordingly, the data for detection and diagnosis comes from participants' verbal protocols. Participants' level of diagnosis was scored based on their level of accuracy, and were categorized into four levels (Pawlak & Vicente, 1996):

0 - the participant says nothing relevant to the fault or nothing at all

1 - the participant states that the system is not behaving as expected and describes the symptoms of the fault at a general level (i.e., "The level of reservoir 1 is dropping.")

2 - the participant describes fault symptoms at a more functional level, but still fails to localize the fault (i.e., "I'm losing flow into reservoir one somehow.")

3 - the participant correctly localizes the root cause of the fault (i.e., "VA1 is blocked.")

Using this scheme, higher diagnosis scores indicate a higher understanding of the functioning of the system.

Compensation time is the time taken to successfully complete a fault trial. Compensation times were counted regardless of whether or not a fault was detected.

Results are presented below: Table 39 presents fault detection times, Table 40 presents the number of faults detected, Table 41 presents the average highest detection score reached, Table 42 presents the average detection times, and Table 43 presents the number of faults correctly diagnosed (i.e., given a diagnosis score of 3). Each table presents results for both the pre- and post-transfer phases. Table 44 presents averages of each of these measures across interfaces to help in comparing the performance of the two groups. Finally, Table 45 presents the average

compensation times for all participants in the pre- and post-transfer phases, as well as the within-interface averages. Note that in these analyses the non-routine fault that occurred in trial 23 was considered as two separate faults (fault 1 was the increase in heater output at 3 minutes and fault 2 was the increase in temperature of the input water at 5 minutes), while the non-routine fault in trial 47 (reservoir 1 leaking into reservoir 2) was treated as one fault.

The purpose of the subsequent analyses is two-fold. First of all, it is again important to establish if there was any difference between the two groups in the pre-transfer phase. Second, given that the ability of the groups on fault tasks was roughly equivalent, the next task is to establish any connection that exists between tool use and performance on fault trials. Such a connection might indicate the effect of tool use on understanding.

Table 39. Average detection times in pre- and post-transfer trials (s).

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer	41	14	66	95	7	20	79	54	82	134	50	4
Post-transfer	102	82	24	43	66	63	22	68	184	77	147	30
Averages	Pre 40.5 / Post 63.3						Pre 67.2 / Post 88.0					

Table 40. Number of faults detected in pre- and post-transfer trials.

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer (4 faults)	4	1	4	4	3	4	4	3	4	2	4	2
Post-transfer (4 faults)	4	3	2	3	3	4	3	4	4	3	4	4
Averages	Pre 3.3 / Post 3.2						Pre 3.2 / Post 3.7					

Table 41. Average highest diagnosis score reached in pre- and post-transfer trials.

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer	2.25	0.75	2.25	2.25	1.25	2.75	2.50	2.00	2.50	1.00	2.50	1.25
Post-transfer	2.00	1.50	1.00	1.50	0.75	2.25	1.50	2.25	2.50	0.50	2.25	1.75
Averages	Pre 2.0 / Post 1.5						Pre 1.9 / Post 1.8					

Table 42. Average diagnosis time in pre- and post-transfer trials (s).

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer	54	14	145	197	77	19	277	82	120	379	129	11
Post-transfer	125	n/a	16	104	n/a	82	n/a	237	340	n/a	295	229
Averages	Pre 84.3 / Post 82.5						Pre 166.3 / Post 275.3					

Table 43. Number of faults diagnosed (i.e. diagnosis score = 3) in pre- and post-transfer trials.

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer (4 faults)	1	1	3	3	1	3	3	2	3	1	2	1
Post-transfer (4 faults)	2	0	1	1	0	2	1	1	2	0	1	1
Averages	Pre 2 / Post 1						Pre 2 / Post 1					

Table 44. Comparison of fault measures across interfaces. The best group on each measure for each phase is in boldface.

	Average Detection Time		*Number of Faults Detected*		*Highest Diagnosis Score*		*Average Diagnosis Time*		*Number of Faults Diagnosed*	
	P	**Div**	**P**	**Div**	**P**	**Div**	**P**	**Div**	**P**	**Div**
Pre-transfer (4 faults)	**40.5**	67.2	**3.3**	3.2	**2.0**	1.9	**84.3**	166.3	2	2
Post-transfer (4 faults)	**63.3**	88.0	3.2	**3.7**	1.5	**1.8**	**82.5**	275.3	1	1

Table 45. Average compensation times for pre- and post- transfer trials (s).

	P Interface						*Divided Interface*					
	Bach	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Pre-transfer	975	916	665	1049	849	822	746	800	820	1421	805	592
Post-transfer	1186	1166	710	1230	672	802	812	1104	963	692	702	1137
Averages	Pre 879 / Post 961						Pre 864 / Post 901					

Beginning with the results on detection time and number of faults detected, both groups of participants seem to be fairly closely matched in the pre-transfer period. Boccherini (P), Telemann (divided) and Willan (divided) detected the least number of faults, and Boccherini (P) and Telemann (divided) had the worst detection times. Still, there is does not seem to be a large difference between the groups. A GLM [8] on fault detection time (Table 46) and an independent samples *t*-tests on the number of detections (t_{10} = -.26, p = .80) confirmed this result. Note, however, that there is a significant interaction between interface and trial, indicating a difference in performance on the non-routine fault in trial 23 (see Figure 41). This interaction does not imply any large difference between the groups.

The results on average highest diagnosis score in the pre-transfer period are similar. Boccherini (P), Telemann (divided), and Willan (divided) all had poor scores in this period, but there seems to be no large difference between the groups. An independent samples *t*-test confirmed this result (t_{10} = .10, p = .92).

[8] Since fault detection and diagnosis times are generally well modelled by a lognormal distribution (Law & Kelton, 1991), the data for this GLM and those that follow in this section were transformed to their logarithms to make a better fit with the normal distribution.

Table 46. GLM on pre-transfer detection time.

Source	DF	SS	MS	F	p
Interface	1	.022	.022	.06	.817
Participant(Interface)	10	3.979	0.398		
Trial	2	9.118	1.559	12.66	< .001
Interface × Trial	2	1.017	.509	4.13	.041
Participant × Trial (Interface)	12	1.601	.123		

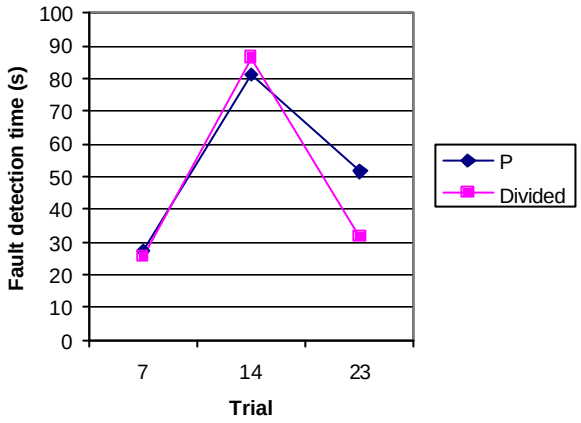


Figure 41. Plot of the interaction between trial and interface group for the pre-transfer group.

Moving on to diagnosis times and the number of faults diagnosed, a slightly different result is seen. The P interface group had lower diagnosis times overall, but were not as consistent in terms of number of faults diagnosed as the divided group. Still, a GLM on diagnosis time (Table 47) and an independent samples *t*-test on the number of faults diagnosed ($t_{10} < .01$, $p = 1.00$) confirms that the differences between the groups are not reliable.

Finally, the pre-transfer compensation times were evenly distributed in the two groups, with the exception of Prokofiev (P) and Telemann (divided) who had unusually high compensation times. Again, an independent samples *t*-test confirms that there is no difference between the groups on this measure ($t_{10} = -.12$, $p = .91$).

Table 47. GLM on pre-transfer fault diagnosis time.

Source	DF	SS	MS	F	p
Interface	1	.121	.121	0.31	0.588
Participant(Interface)	10	3.874	.387		
Trial	2	.474	.237	1.66	.249
Interface × Trial	2	.383	.192	1.35	.313
Participant × Trial (Interface)	8	1.138	.142		

For post-transfer performance, although the P group had faster detection times than the divided group in the post-transfer phase, the difference in times is consistent with the small difference already noted in pre-transfer performance. Since this is the case, the difference could be a result of differences between participants. At any rate, a GLM (Table 48) confirmed that this difference was not significant. Also, even though the divided group detected more post-transfer faults than the P group, this difference was not significant ($t_{10} = 1.34$, $p = .21$).

Further, the P group on average diagnosed the same number of faults as the divided group ($t_{10} = 0.00$, $p = 1.00$) and had only a slightly lower average diagnosis score than the divided group ($t_{10} = .77$, $p = .46$). A GLM (Table 49) revealed that the P group's diagnosis times were almost significantly lower than those of the divided group. Finally, even though there was a slight difference between the compensation times for the P and divided groups in favour of the divided group, this difference was not significant ($t_{10} = -.45$, $p = .66$).

In sum, there were no systematic differences between the groups during pre-transfer fault performance, and their performance remained remarkably consistent into the post-transfer phase.

Table 48. GLM on post-transfer fault detection times.

Source	DF	SS	MS	F	p
Interface	1	> .001	> .001	0.00	.966
Participant(Interface)	10	2.477	.248		
Trial	3	2.144	.714	4.30	.016
Interface × Trial	3	.570	.190	1.14	.354
Participant × Trial (Interface)	22	3.659	.166		

Table 49. GLM on post-transfer fault diagnosis times.

Source	DF	SS	MS	F	p
Interface	1	.423	.423	4.53	.077
Participant(Interface)	6	.560	.093		
Trial	1	.065	.065	.72	.552
Interface × Trial	1	.048	.048	.53	.599
Participant × Trial (Interface)	1	.090	.090		

Given this null result, the next step is to determine if tool use was responsible for any individual differences in performance that might have masked the expected between interface differences. To answer this question, a number of correlations were performed between the measures of tool use and the fault measures. For both the pre- and post-transfer periods, correlations were performed between the tool use measures of average tool use and tool use variability and all of the fault measures discussed above. Only one correlation turned out to be strong enough to be nearly significant: post-transfer tool use with post-transfer detection time (R_{12} = .56, p = .06). A scatterplot of these data are presented in Figure 42. This plot indicates that the correlation might be unduly biased by the data from Rachmaninov, who had both a large number of tool uses and a high average detection time. If Rachmaninov's data are removed, the correlation loses its strength and significance (R_{11} = .002, p = .99). (Even though Rachmaninov had a high average detection time, it should be noted that he was the best overall performer in terms of the number of faults detected and diagnosed.)

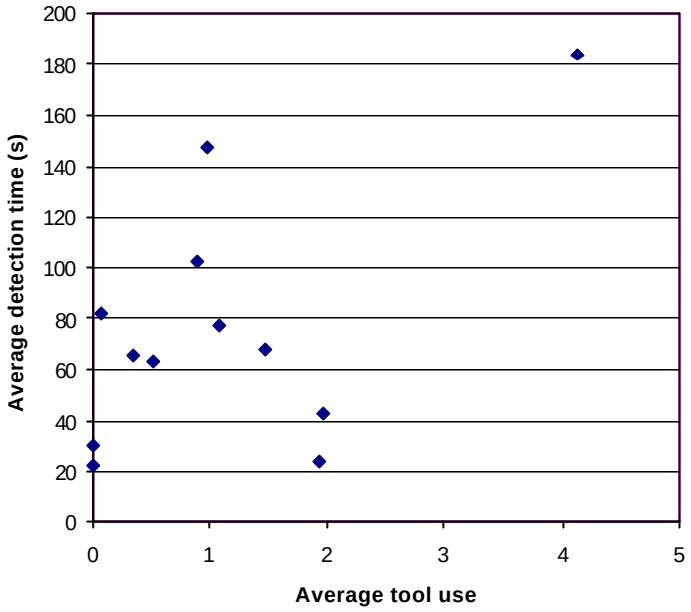


Figure 42. Scatterplot of the correlation between average detection time and average tool use in the post-transfer phase.

Summary. These analyses were unable to uncover any connection between tool use and performance on fault trials. This result leaves open the question as to whether or not tool use reflects an increased understanding of the system.

APPENDIX E: CONTROL RECIPE ANALYSES [9]

Measures of Length and Chunking

a. Length of control recipes.

		Recipe completed at trial no.				
Interface	**Subject**	**0**	**12**	**33**	**52**	**Mean**
Divided	Bach	215	175	267	138	199
	Boccherini	208	160	416	277	265
	Górecki	259	247	258	247	253
	Prokofiev	383	387	395	278	361
	Schoenberg	48	99		97	81
	Schubert	102	146	143	125	129
P	Bartók	99	141		166	135
	Mozart		283	300	187	257
	Rachmaninov	186	225	308	254	243
	Telemann	209	220	204	209	211
	Wagner	96	128	133	100	114
	Willan	40	104	63	101	77
Mean for P Interface		203	202	296	194	224
Mean for Divided Interface		126	184	202	170	170
Mean overall		168	193	249	182	198

b. Number of distinct steps in each recipe.

		Recipe completed at trial no.				
Interface	**Subject**	**0**	**12**	**33**	**52**	**Mean**
Divided	Bach	9	10	9	11	10
	Boccherini	7	8	14		10
	Górecki	15	14	9	12	13
	Prokofiev	14	7	11		11
	Schoenberg	5	7		5	6
	Schubert	6	9	13	7	9
P	Bartók	7		5	8	7
	Mozart		13	7	8	9
	Rachmaninov	8	6	9	11	9
	Telemann	13	11	6	8	10
	Wagner	6	6	5	5	6
	Willan	5	10	9	9	8
Mean for P Interface		9	9	11	9	10
Mean for Divided Interface		8	9	7	8	8
Mean overall		9	9	9	8	9

[9] The data for the control recipes for Schoenberg at trial 33 and Bartók at trial 0 are missing data.

c. Average number of words per chunk.

Interface	Subject	Recipe completed at trial no. 0	12	33	52	Mean
Divided	Bach	24	18	30	13	21
	Boccherini	30	20	30		26
	Górecki	17	18	29	21	21
	Prokofiev	27	55	36		40
	Schoenberg	10	14		19	14
	Schubert	17	16	11	18	16
P	Bartók	14		0	21	12
	Mozart		22	43	23	29
	Rachmaninov	23	38	34	23	30
	Telemann	16	20	34	26	24
	Wagner	16	21	27	20	21
	Willan	8	10	7	11	9
Mean for P Interface		21	23	27	18	22
Mean for Divided Interface		15	22	24	21	21
Mean overall		18	23	25	19	22

Measures of Differentiation

a. Number of references to steady state heater setting as a ratio of reservoir demand, for each reservoir.

Interface	Subject	Recipe completed at trial no. 0	12	33	52
Divided	Bach				
	Boccherini				
	Górecki				
	Prokofiev				
	Schoenberg				
	Schubert				
P	Bartók				
	Mozart			2	2
	Rachmaninov			2	2
	Telemann				
	Wagner				
	Willan				

b. Explicit references to decoupled feedwater stream strategy.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini				
	Górecki		1*		1*
	Prokofiev			1	1*
	Schoenberg				
	Schubert		1	1	1*
P	Bartók		1*		1*
	Mozart		1*	1*	1*
	Rachmaninov		1*	1*	1*
	Telemann			1	
	Wagner				
	Willan		1*	1*	1*

c. Number of asymmetrical statements.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini				
	Górecki				
	Prokofiev		1		1
	Schoenberg				
	Schubert		1		
P	Bartók				
	Mozart			1	1
	Rachmaninov			1	1
	Telemann				
	Wagner				
	Willan				1

Measures of Knowledge Organisation

a. Number of explicit references to perceptual features of the display.

		Recipe completed at tria l no.			
Interface	Subject	0	12	33	52
Divided	Bach	2			
	Boccherini	2	4	5	2
	Górecki	1	2	5	3
	Prokofiev	5	7	2	4
	Schoenberg				2
	Schubert		2	2	2
P	Bartók				
	Mozart	2	3		3
	Rachmaninov	1	2		1
	Telemann	6			3
	Wagner	1	2		
	Willan				

b. Number of statements specifying quantitative values for component settings.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach	—	1	3	2
	Boccherini	—	1	4	1
	Górecki	—	1	1	2
	Prokofiev	—	3	2	4
	Schoenberg	—			
	Schubert	—			
P	Bartók	—	1		1
	Mozart	—	3	3	3
	Rachmaninov	—	2	2	1
	Telemann	—			
	Wagner	—	2	2	2
	Willan	—			2

c. Number of declarative knowledge statements

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach	4			
	Boccherini	3			
	Górecki				
	Prokofiev	1	1		
	Schoenberg				
	Schubert				
P	Bartók				
	Mozart		1		
	Rachmaninov	2		1	
	Telemann				
	Wagner				
	Willan				

d. Number of statements justifying or explaining recipe steps.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini				
	Górecki	1			
	Prokofiev				
	Schoenberg				
	Schubert				
P	Bartók				
	Mozart				
	Rachmaninov		1		
	Telemann				
	Wagner				
	Willan				

e. Number of statements specifying the goal to be achieved without listing the precise actions that need to be performed or the relationships that need to be considered.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini			1	1
	Górecki		1	1	1
	Prokofiev	1	2	2	2
	Schoenberg		1		1
	Schubert				
P	Bartók				
	Mozart				
	Rachmaninov				
	Telemann	1	3	1	1
	Wagner				
	Willan				

Measures Related to Tool Use

a. Number of explicit references to tool use.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini				
	Górecki				
	Prokofiev				
	Schoenberg				
	Schubert				
P	Bartók				
	Mozart		3		
	Rachmaninov			4	4
	Telemann				
	Wagner			1	2
	Willan				

b. Number of pictures and figures.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach				
	Boccherini		1	1	
	Górecki				
	Prokofiev		1	2	
	Schoenberg				3
	Schubert				
P	Bartók				
	Mozart		3	2	1
	Rachmaninov		2	2	2
	Telemann	3	1	1	1
	Wagner				
	Willan				

c. Number of statements warning about system constraints.

		Recipe completed at trial no.			
Interface	**Subject**	**0**	**12**	**33**	**52**
Divided	Bach	2	1		
	Boccherini				
	Górecki				
	Prokofiev				
	Schoenberg				1
	Schubert				
P	Bartók	1			
	Mozart		4	2	3
	Rachmaninov	2	1	1	
	Telemann		2	2	1
	Wagner				
	Willan		3		

APPENDIX F: ABSTRACTION HIERARCHY ANALYSES

The following pages contain the data and graphs results from the application of abstraction hierarchy analyses (Yu et al., 1997) to the experiment log files.

		Divided Interface						P Interface					
	Block*	**Bach**	**Boccherini**	**Górecki**	**Prokofiev**	**Schoenberg**	**Schubert**	**Bartók**	**Mozart**	**Rachmaninov**	**Telemann**	**Wagner**	**Willan**
Outputs	**1**	0.18	0.79	0.42	0.35	0.18	0.27	1.78	0.46	0.24	0.46	1.04	0.12
	2	0.06	0.30	0.12	0.29	0.16	1.69	0.13	0.29	0.17	0.79	0.19	0.06
	3	0.09	0.33	0.10	0.17	0.21	0.16	0.09	0.52	0.40	0.65	0.10	0.06
	4	0.08	0.12	0.09	0.17	23.25	1.27	0.15	0.27	0.16	0.20	0.12	0.05
Mass & Energy Variance	**1**	0.53	1.97	0.72	0.86	0.52	0.80	3.72	1.52	0.93	1.87	0.29	0.49
	2	1.27	1.53	0.12	0.72	0.17	0.86	0.39	4.51	1.82	2.19	0.29	0.56
	3	0.75	1.61	0.10	0.48	0.34	1.78	0.50	2.28	3.08	2.70	0.27	0.72
	4	1.20	1.88	0.21	0.94	0.26	2.67	0.86	2.18	3.55	2.13	0.22	0.73
Flows	**1**	0.25	0.35	0.45	0.19	0.25	0.29	0.55	0.35	0.28	0.41	0.29	0.18
	2	0.25	0.45	0.56	0.55	0.32	0.20	0.41	0.45	0.35	0.40	0.29	0.34
	3	0.17	0.30	0.41	0.29	0.20	0.16	0.34	0.30	0.27	0.31	0.27	0.17
	4	0.19	0.31	0.49	0.22	0.23	0.17	0.36	0.31	0.19	0.32	0.22	0.20
Actions	**1**	0.33	0.66	0.82	0.26	0.39	0.45	0.91	0.64	0.55	0.76	0.42	0.29
	2	0.30	1.00	0.97	1.07	0.43	0.34	0.86	0.43	0.57	0.57	0.39	0.44
	3	0.21	0.67	0.64	0.39	0.25	0.26	0.60	0.35	0.46	0.50	0.36	0.19
	4	0.23	0.66	0.92	0.32	0.54	0.26	0.86	0.37	0.38	0.52	0.59	0.27

*Blocks 1 and 2 are from the pre-transfer phase, and blocks 3 and 4 are from the post transfer phase. Each block contains 10 normal trials, except for block 4 which contains 15 normal trials.

Note: In the graphs that follow, cohorts between the P and divided interface groups are matched by position (i.e, graphs for Bach always appear in the upper left hand corner of the divided interface graphs, and those for Bach's cohort, Willan, appear in the upper left hand corner of the P interface graphs).

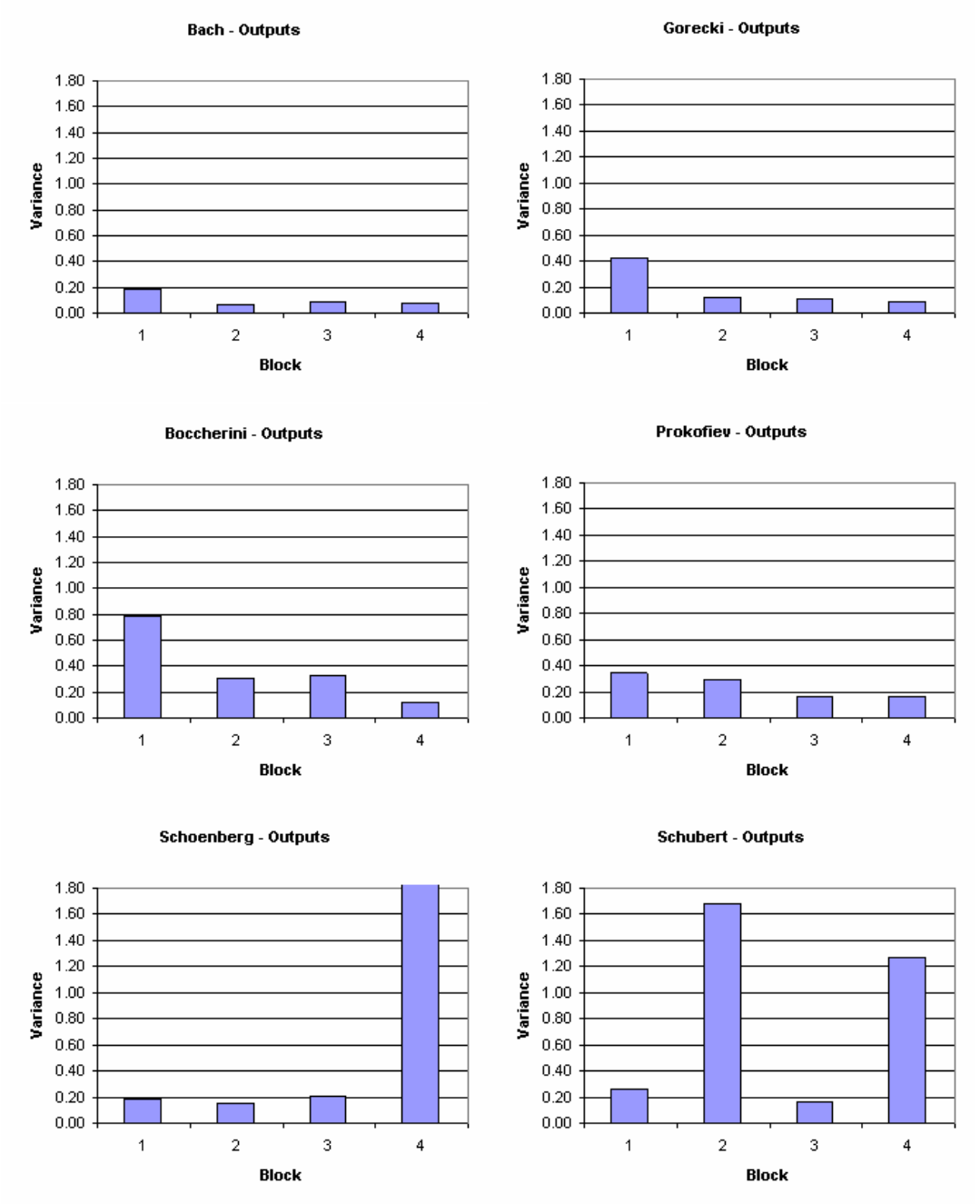


Figure 43. Output variance, divided interface group

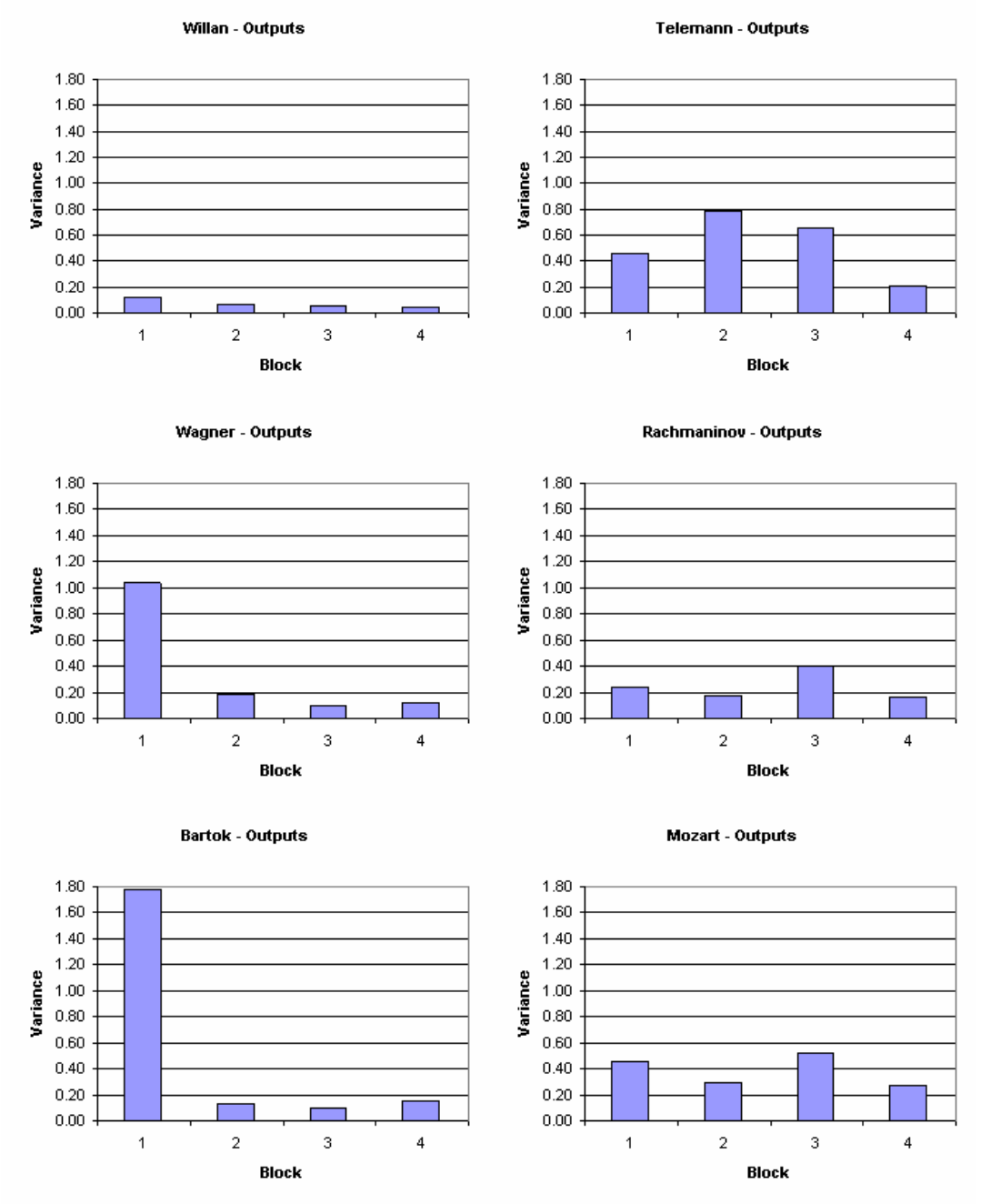


Figure 44. Output variance, P interface group

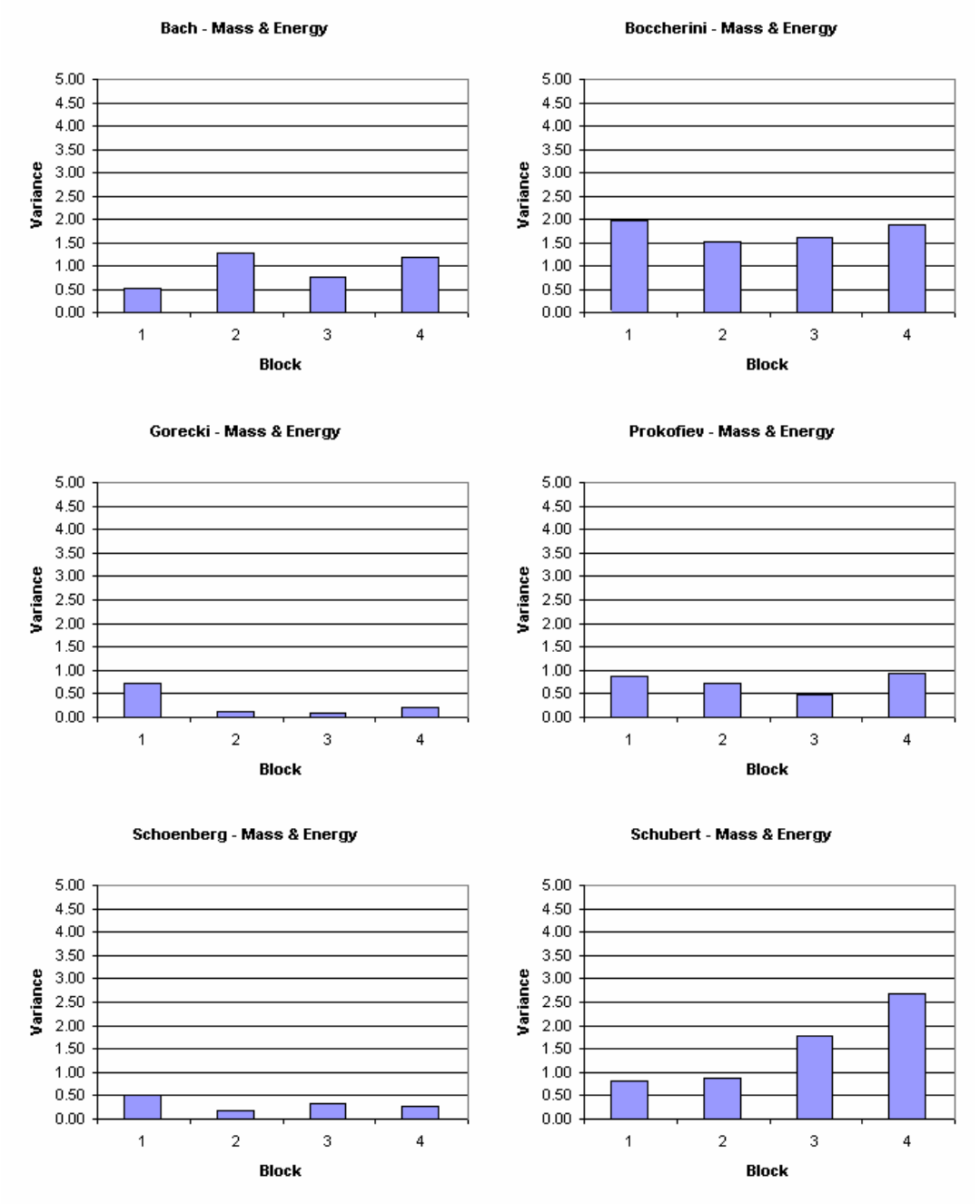


Figure 45. Mass and energy variance, divided interface group

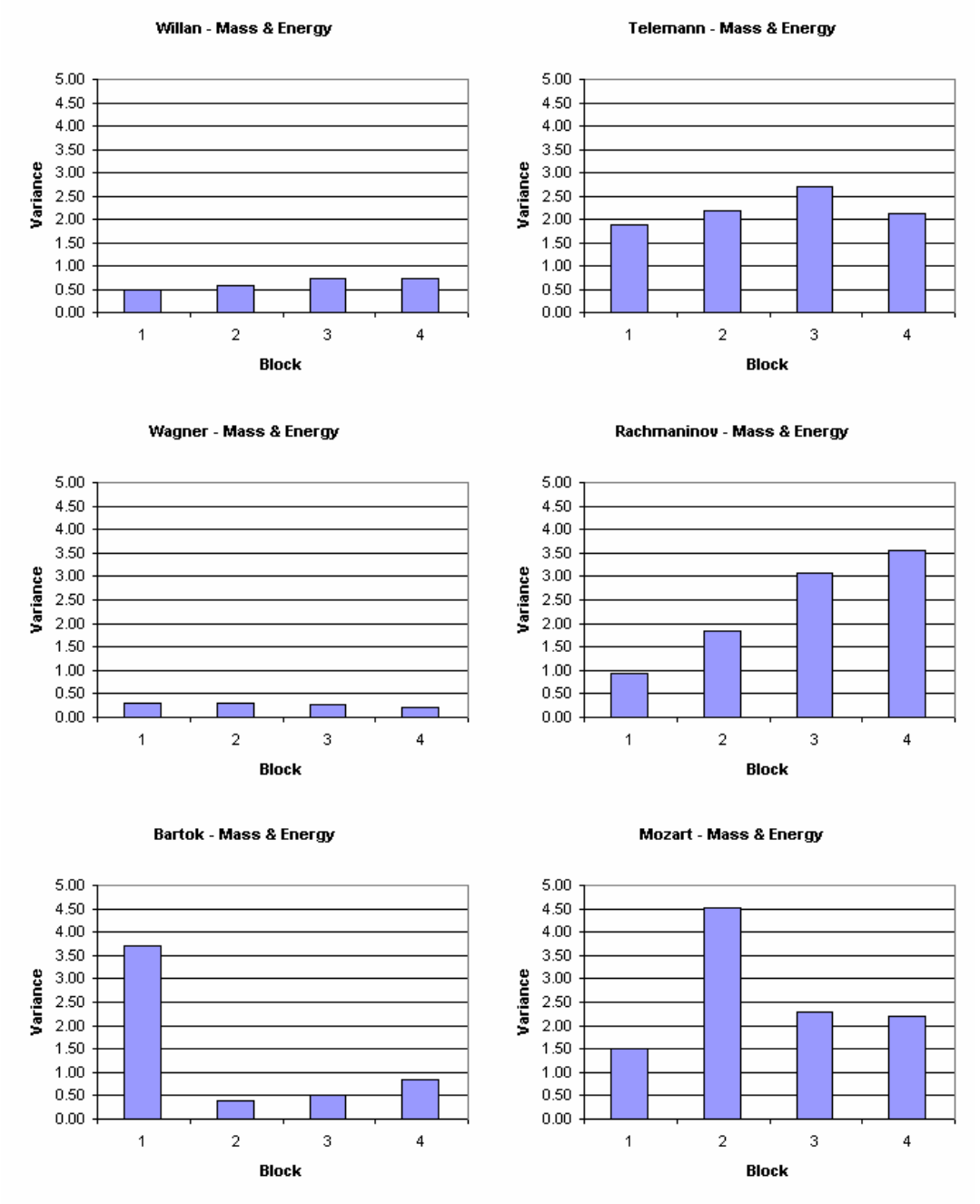


Figure 46. Mass and energy variance, P interface group

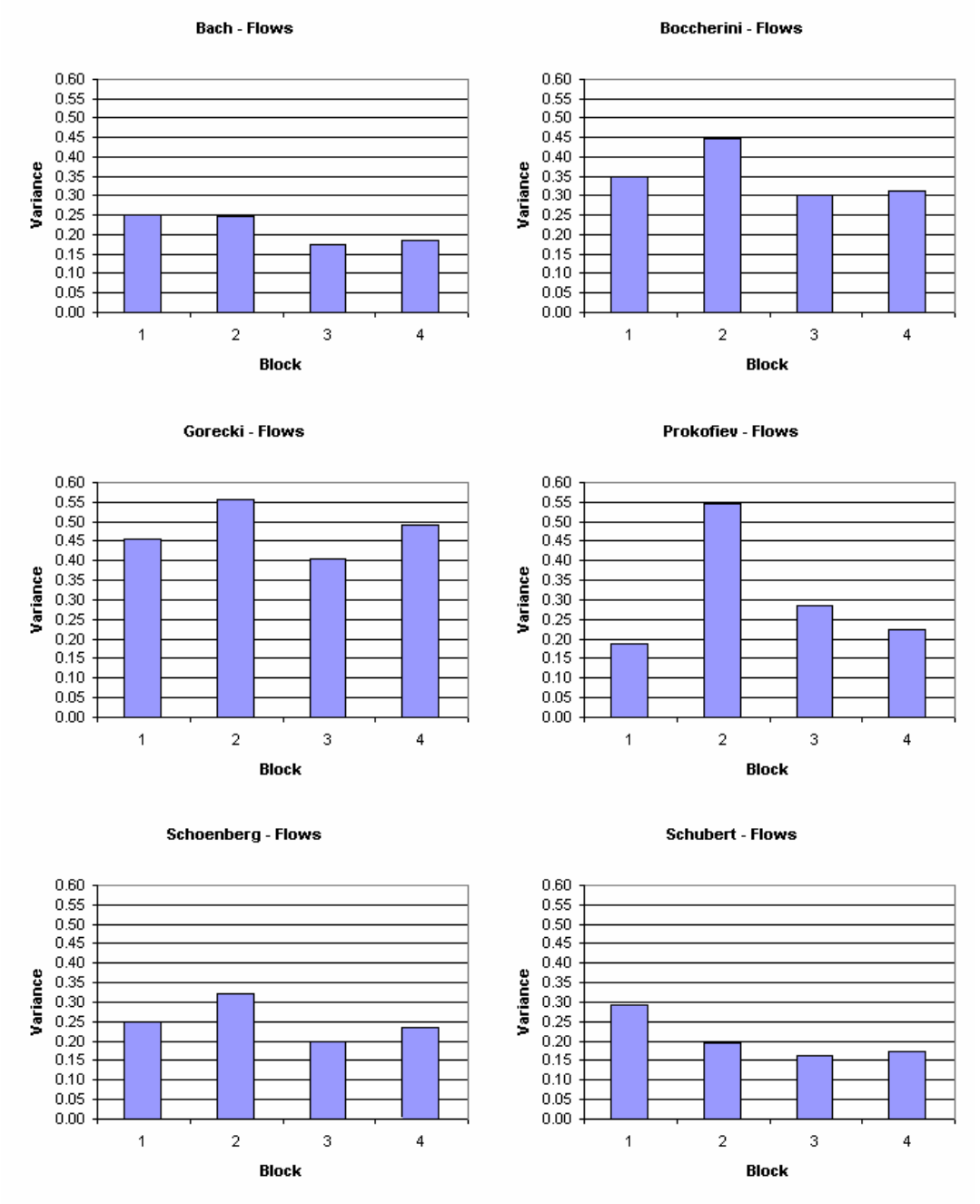


Figure 47. Flows variance, divided interface group

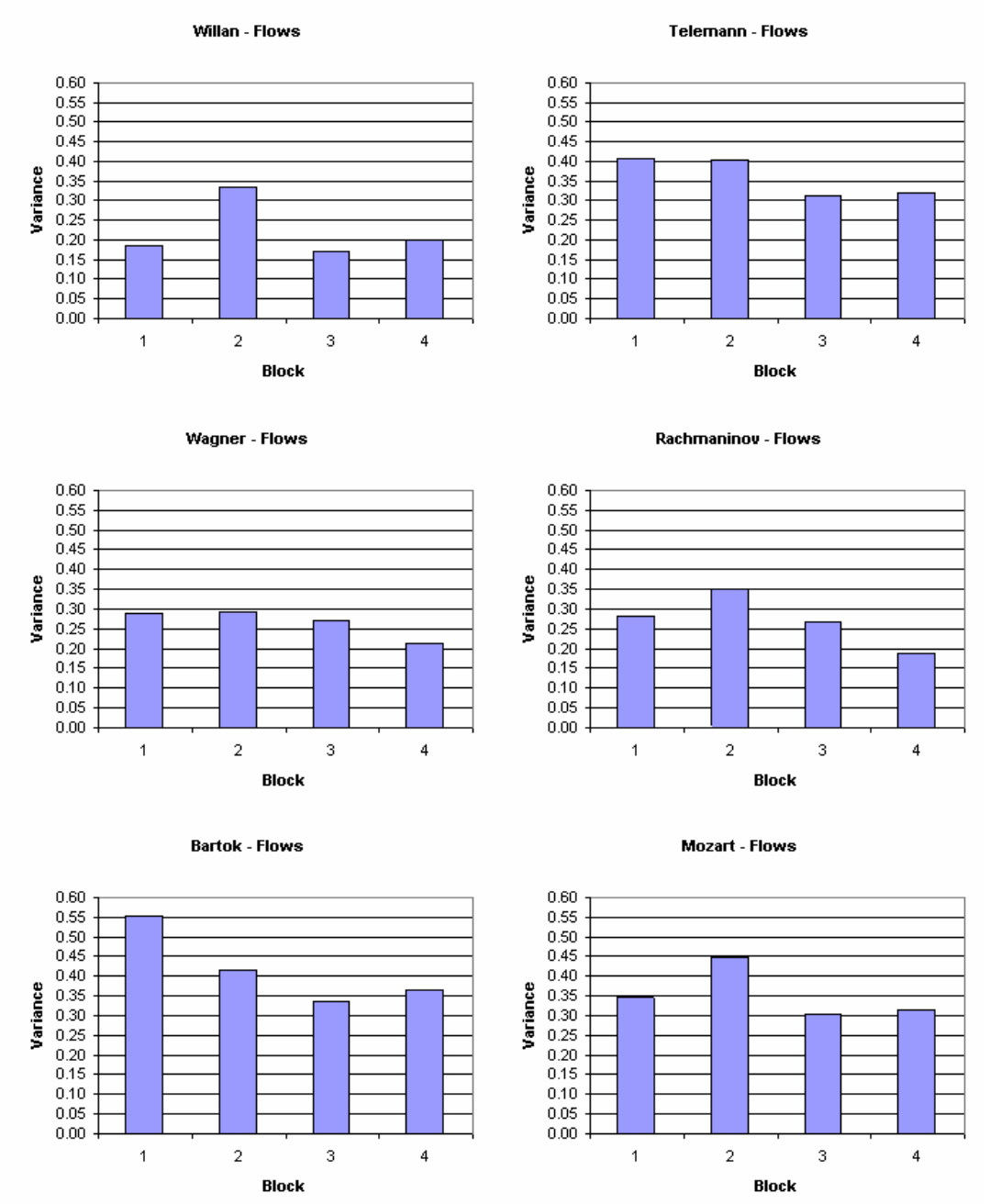


Figure 48. Flows variance, P interface group

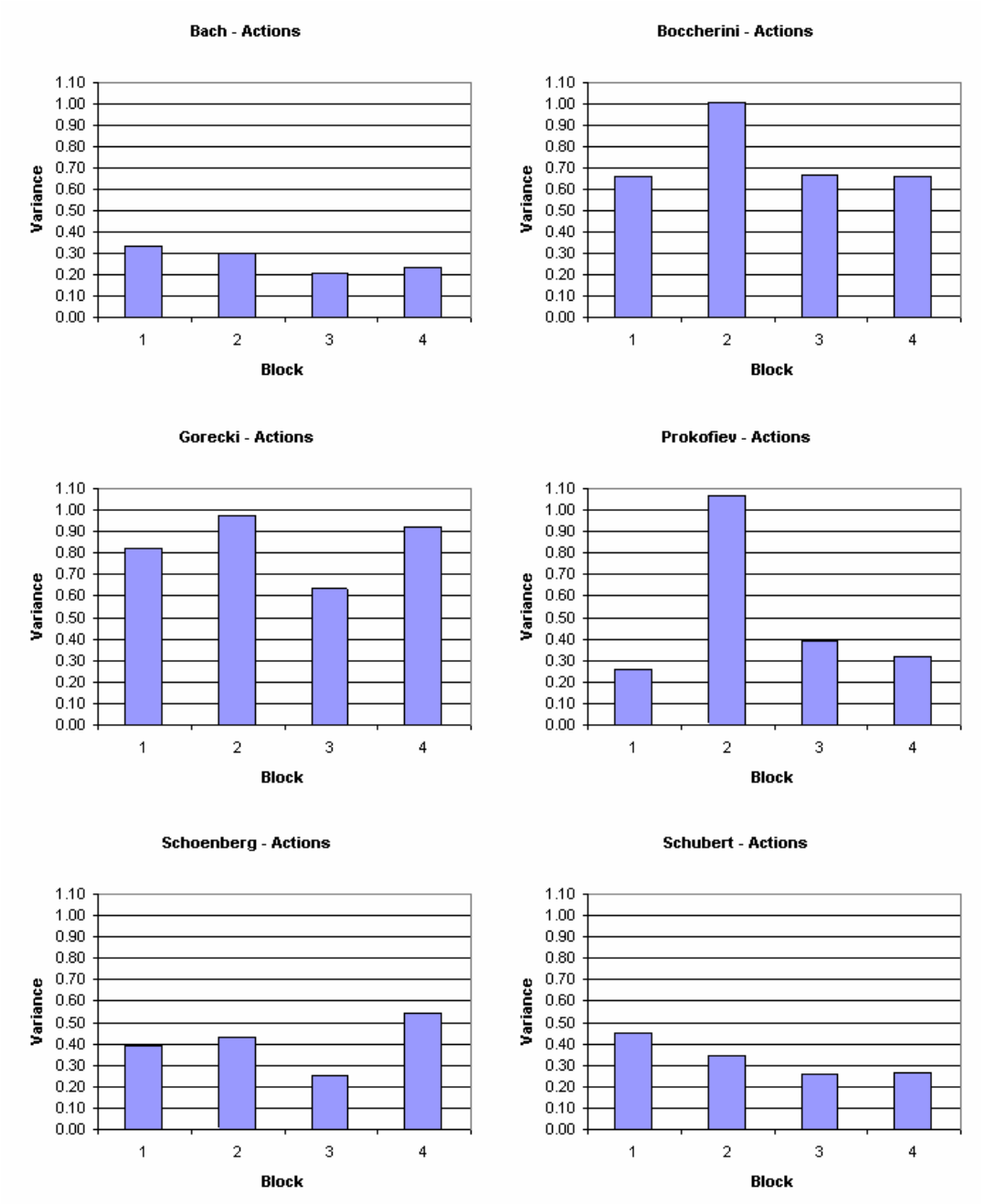


Figure 49. Action variance, divided interface group

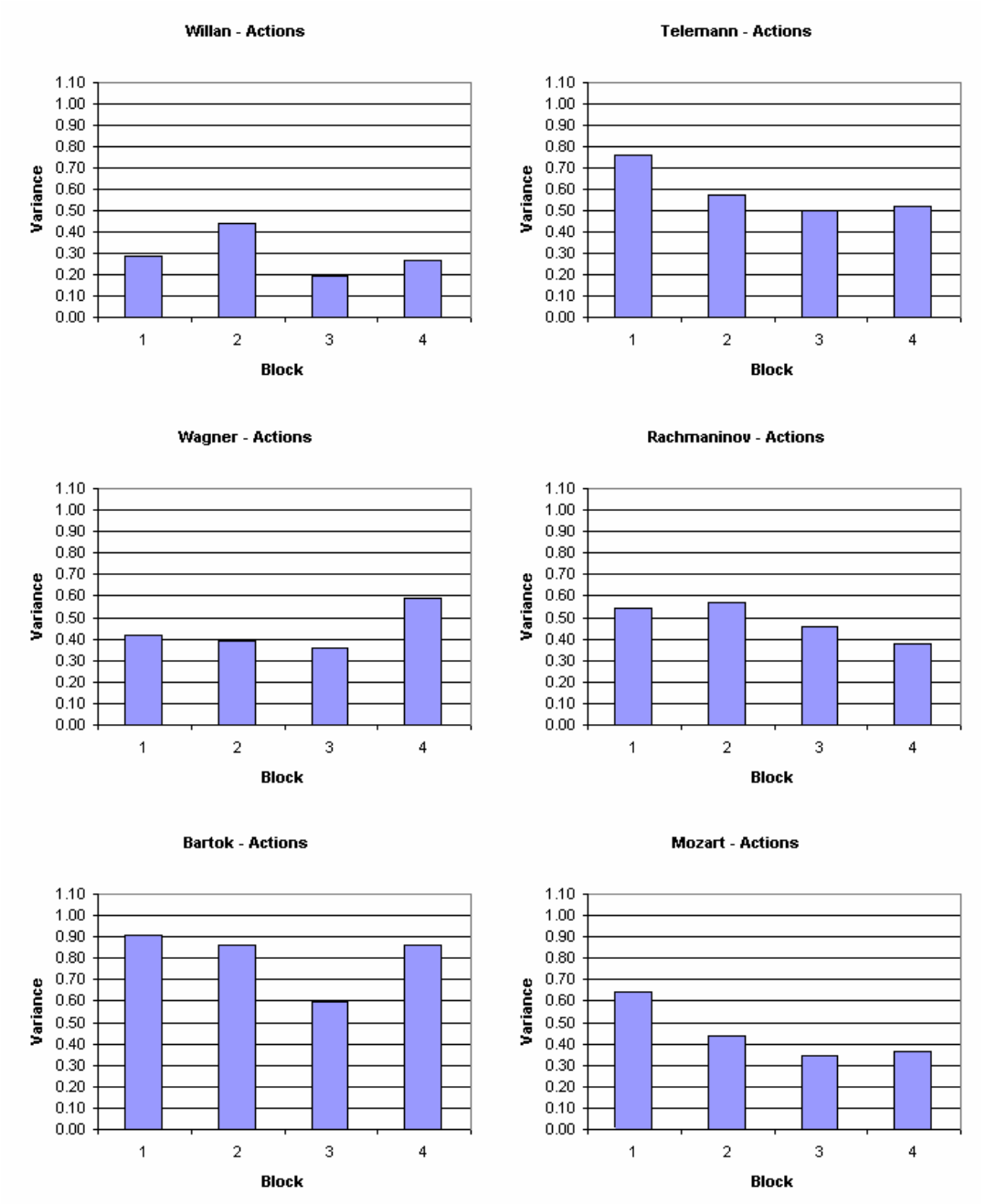


Figure 50. Action variance, P interface group

APPENDIX G: EXPERIMENTAL PROTOCOLS

The following pages contain all of the protocols and materials given to participants over the course of the experiments detailed in this thesis. These are:

- Introduction to the Experiment (Session 0)
- Introduction to the Experiment (Session 1)
- Consent To Take Part in Research
- Demographic Questionnaire
- Technical Description of DURESS II
- Experimental Procedure: P Interface
- Experimental Procedure: P+F Interface
- Experimental Procedure: Divided Interface
- Verbal Protocol Instructions
- Control Recipe Instructions & Form
- Spy Ring History Test