ELSEVIER

Applied Ergonomics 31 (2000) 395–408

APPLIED ERGONOMICS

www.elsevier.com/locate/apergo

Making the most of ecological interface design: the role of individual differences

Gerard L. Torenvliet, Greg A. Jamieson, Kim J. Vicente*

Cognitive Engineering Laboratory, Department of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Road, Toronto, Ontario, Canada

Received 24 November 1998; accepted 26 October 1999

Abstract

As advanced control rooms for new process control plants are being designed, the question arises as to whether operators of the future need to have a particular set of cognitive characteristics to make the most of those advanced control rooms. This issue was investigated by examining the interaction between ecological interface design (EID) and individual differences in the context of a process control microworld. A number of potential predictors of performance were investigated, including: demographic data, type of interface, type of instruction, and data from two cognitive style tests. Eight linear regression analyses were conducted to determine which variables were the strongest predictors of performance. The results indicate that the strongest and most consistent predictor of performance was the interaction between a holist cognitive style score and an interface based on the principles of EID. That is, individuals who used an EID interface and who had high holist scores were the best performers. It seems that these individuals have the relational thinking ability that is required to exploit the value of the higher-order functional information provided by an EID interface. This empirical result has practical implications for operator selection. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: Cognitive style; Individual differences; Ecological interface design; Cognitive engineering; Process control; Operator selection

1. Introduction

As advanced control rooms are being designed for next-generation process control plants, the question arises as to whether operators of the future should have different characteristics than current operators. Since advanced control rooms are qualitatively different from traditional designs, it would not be surprising to find that a different set of competencies will be required to be an effective operator (Olsen and Rasmussen, 1989). If the need for a change in operator characteristics is established, then the question becomes whether such characteristics can be acquired through an appropriate training program or if they must be selected for instead. Training would be preferable since it would allow all individuals access to the job, but we cannot rule out a priori the possibility that individuals with certain characteristics may not function effectively in advanced control rooms, regardless of the training they receive. In such cases, selection criteria may have to be introduced.

The tension between training and selection has been brought to the forefront in previous research on ecological interface design (EID), a framework for designing advanced interfaces for complex sociotechnical systems (Vicente and Rasmussen, 1990, 1992; Vicente et al., 1996). EID is based on the skills, rules, knowledge taxonomy of levels of cognitive control (Rasmussen, 1983). The framework consists of three prescriptive design principles, each directed at providing the appropriate interface support for a specific level of cognitive control. First, to support skill-based behaviour, operators should be able to act directly on the interface. Furthermore, the structure of the displayed information should be isomorphic to the part–whole structure of movements. Second, to support rule-based behaviour, the interface should maintain a consistent one-to-one mapping between the work domain constraints and the perceptual cues provided in the interface. Third, to support knowledge-based behaviour, the interface should represent the work domain in the form of an abstraction hierarchy (Rasmussen, 1985), which can serve as an externalized mental model to

* Corresponding author. Tel.: (416) 978-7399; fax: (416) 978-3453URL:www.mie.utoronto.ca/labs/cel.

E-mail address: benfica@mie.utoronto.ca (K.J. Vicente).

PII: S0003-6870(00)00004-1

support problem solving. This model contains both physical and functional representations of a work domain.

Evaluations of EID conducted in our laboratory have compared the performance of two different interfaces for a process control microworld (Vicente and Rasmussen, 1990; Pawlak and Vicente, 1996). One interface was designed according to the principles of EID and thus contains both physical and functional ($P + F$) information, as identified by an abstraction hierarchy requirements analysis. Another interface was designed according to more traditional interface practices, and thus contains only physical information (P).

Several longitudinal experiments have been conducted comparing participants' performance with these two interfaces under both normal and abnormal conditions. These experiments consistently show that the $P + F$ interface based on the principles of EID leads to better performance than the more traditional P interface format (Pawlak and Vicente, 1996; Christoffersen et al., 1996, 1997, 1998; Hunter et al., 1996). At the same time, however, substantial individual differences have been observed. Some individuals using the $P + F$ interface perform much more effectively than others. These individual differences have been observed in studies where there has been no training (Pawlak and Vicente, 1996; Christoffersen et al., 1996, 1997, 1998; Janzen and Vicente, 1998), as well as in studies where participants have received some form of instruction (Hunter et al., 1996; Howie and Vicente, 1998). These results suggest that there may be one or more traits that predispose an individual to perform well with an EID interface. This issue is of significant practical importance. Since the best participants with an EID interface outperform the best participants with a more traditional interface, there are good reasons for choosing EID, but we must better understand how we can create the conditions so that operators can "make the most of EID". Can we rely on certain types of training programs, or do we also have to choose certain types of individuals for the job?

This research investigated the latter possibility by conducting a series of regression analyses for the individual differences observed in previous studies we have conducted. Our goal was to determine what factors are the best predictors of participants' performance, thereby identifying the source of the aforementioned individual differences.

1.1. Background

To set the stage for the regression analyses, we first briefly describe the four relevant experiments, all conducted in the context of the DURESS (DUal REservoir System Simulation) II testbed, a simulation of a highly simplified yet representative thermal-hydraulic process plant. DURESS II is designed to evaluate two kinds of interfaces, a conventional interface displaying only physical information (P interface) and an interface developed according to the principles of EID that displays both physical and functional information ($P + F$ interface). (For a complete description of the DURESS II testbed and the P and $P + F$ interfaces developed for it, see Pawlak and Vicente, 1996).

Each of the four experiments was designed to investigate the effects on operator adaptation of modifying one or more behaviour shaping constraints (Rasmussen et al., 1994; Vicente, 1999). Generally speaking, behaviour shaping constraints are any type of constraint that may shape how operators adapt to a work domain. Adaptation can be measured by investigating how well operators are able to control the simulation in the face of perturbations (e.g., faults of various types). The specific behaviour shaping constraints relevant to this program of research are as follows (Vicente, 1997):

- *Interface content*. Interface content can be a strong constraint on operator performance. While providing proper and relevant information is a necessary (but not sufficient) provision for functional adaptation, either neglecting to include critical information or providing irrelevant information can foster dysfunctional adaptation.
- *Interface form*. Independent of the information content of an interface is the form of information presentation. Operators may become increasingly attuned to the visual form of an interface; in other words the visual form of an interface will manipulate an operator's attention. An interface that directs an operator's attention to critical information should foster functional adaptation. Conversely, an interface that directs an operator's attention to either non-critical or irrelevant information may promote dysfunctional adaptation.
- *Type of training*. The type and amount of training operators receive influences adaptation. Operators can receive training either prior to or concurrent with operating a work domain. Training provides operators with: (1) a set of competencies tailored to a specific work situation, (2) guidance in what types of information to treat as important, and (3) experience for dealing with novel situations. Many types of training exist, and some research (e.g., Crossman and Cooke, 1962/1974) indicates that at least some types of theoretical training do not foster functional adaptation. The effect of training based on both fundamental physical principles and interface design, however, could foster functional adaptation.
- *Pre-existing competencies*. The participants in each experiment have pre-existing competencies that influence adaptation. These take on such forms as cognitive style, declarative and procedural knowledge, perceptual-motor skills, and population stereotypes.

Table 1
Integrated summary of the three-year research program, showing the behaviour shaping constraints investigated in each experiment (adapted from Vicente, 1997)

Experiment	Behaviour shaping constraints				Number of participants
	Interface content	Interface form	Type of training	Pre-existing competencies	
I	P vs. $P + F$	P vs. $P + F$	None		6
II	P vs. $P + F$	P vs. $P + F$	None vs. AH		24
III	$P + F$	P vs. $P + F$ vs. Divided $P + F$	None	Demographic data + Cognitive style	12[a]
IV	$P + F$	$P + F$	None vs. dplayer vs. Dplayer + SE		18[a]

[a]Experiments III and IV made use of a shared control group of six participants. In total, there were 24 participants in these two experiments.

Although this set of behaviour shaping constraints cannot be controlled in the same fashion as those listed above, an understanding of their effects is important for both experimental design and analysis of results.

Table 1 summarises the manipulations of these behaviour shaping constraints across a series of four experiments as follows.

- *Experiment I* (Christoffersen et al., 1996, 1997, 1998): This experiment was designed to assess the impact of interface content and form on long-term adaptation. It involved a longitudinal investigation in which six participants operated either the P or the $P + F$ interface for the DURESS II microworld quasi-daily over a period of six months (217 trials per participant).
- *Experiment II* (Hunter et al., 1996): This experiment investigated the interaction between interface design and model-based training on adaptation. Twenty-four participants took part in a 2×2 between-participants experiment, with two levels of interface design (P vs. $P + F$) and two levels of training (none vs. abstraction hierarchy [AH] training), over a period of one month (67 trials per participant).
- *Experiment III* (Janzen and Vicente, 1998): This experiment investigated the impact of interface form on adaptation. Two groups (integrated vs. divided) participated for one month (67 trials per participant). The integrated group used the standard $P + F$ interface which presented information about four levels of the abstraction hierarchy all in one integrated view. In contrast, the divided group used an interface that presented information about each level of abstraction on a separate window, for a total of four windows that could only be viewed one at a time.
- *Experiment IV* (Howie and Vicente, 1998): This experiment investigated the effect of a second type of training, self instruction, via performance reviews and/or self-explanation, on operator performance. Eighteen participants were divided equally into three groups, each of which performed identical tasks on the $P + F$ interface while engaging in different levels of performance reviews and/or self-explanation (67 trials per participant). The first group did not review their performance or engage in any self-explanation of control actions. The second group periodically reviewed their performance using Dplayer, a computer program that plays back trials in real-time from data contained in the simulator log files. The third group also periodically reviewed their performance using Dplayer, but its members were also instructed and encouraged to engage in self-explanation of control actions while reviewing their trials.

1.2. Motivation

Previous analyses of the data from these experiments have revealed large between-participants differences in performance that cannot be accounted for solely by experimental manipulations. These differences raise the practical question of whether we can make the most of EID by training or by selection (among other factors).

Preliminary analyses (Howie, 1996) indicated that the holist/serialist cognitive style distinction (Pask and Scott, 1972) was relevant to understanding adaptation in the DURESS II microworld. From previous experiments, a comprehensive database existed containing relevant information for almost all participants to support a more comprehensive analysis of individual differences. It was hoped that this analysis of individual differences would help to identify how the holist/serialist cognitive style distinction affects adaptation in the DURESS II microworld. This understanding should help in addressing the training vs. selection question.

2. Method

The first step in our analysis was to compile and summarise the data from the previous investigations in terms of performance measures of interest (dependent variables) and potentially relevant predictors (independent variables). A large number of performance measures and predictors were included in our database, but only those that were statistically significant are reported here.

2.1. Performance measures for normal trials

In the course of this experimental programme, many measures of operator performance on both normal and fault trials in the DURESS II microworld were adopted. Since the motivation for this research was to investigate individual differences, specific performance measures were selected that are relevant to this aim.

- *Asymptotic trial completion time (TCT)*. Asymptotic TCT is measured as the average amount of time required for a participant to bring the simulation from a shutdown to a steady state for a late block of trials. Since all experiments had at least 67 trials, this late block includes the ten normal (i.e., non-fault) trials prior to and including trial 67.
- *Asymptotic trial completion time variance (CTV)*. Asymptotic CTV is defined as the variance in asymptotic TCT.
- *Number of incomplete normal trials (INT)*. Normal trials are considered incomplete if a participant violated any of the simulation's constraints, causing it to "blow up".

2.2. Performance measures for fault performance

Faults were administered to participants at random intervals over the duration of each experiment. Since there was a relatively small number of faults at irregular intervals, fault data are not well suited to blocking or aggregation as these techniques may result in a loss of understanding of the effects of experience on fault performance. On the other hand, raw (i.e., non-aggregated) fault data may also be noisy and erratic, and may themselves not be suited to making any useful generalisations or conclusions. To realise the benefits of both types of analysis while simultaneously compensating for their deficiencies, two types of individual differences analysis were performed on fault data. The first type is an aggregate analysis in which the various measures of fault performance (described below) were aggregated over all faults for each participant. This analysis could help in isolating the individual differences that affect overall performance on faults. This aggregate analysis has the drawback that it masks the effect of interaction between individual differences and practice. The second type of analysis considers each fault trial for each participant in order to help in understanding the effects of experience and the interaction between experience and individual differences.

The four measures of fault performance described below were used in an individual differences analysis.

- *Fault detections (NFD)*. If participants verbalised that there was a problem with the behaviour of DURESS II, they were said to have detected a fault. A regression was performed on the total number of faults detected over an entire experiment.
- *Fault detection time (TDETECT)*. Fault detection time is the time elapsed between the occurrence of a fault and the participant's verbal detection. A regression was performed on the interaction between fault detection time and experience.
- *Diagnosis accuracy (DA)*. Fault diagnosis scores were assigned to participants' diagnoses, ranging from a score of 0 for an irrelevant utterance to 3 for a correct statement of the precise location and root cause of the fault (Pawlak and Vicente, 1996). A fault was considered to be detected if the diagnosis score is greater than or equal to 1. A regression was performed on the interaction between diagnosis accuracy and experience.
- *Diagnosis time (TDIAG)*. Diagnosis time is the elapsed time between the occurrence of a fault and the participant's verbalisation of a correct root cause diagnosis (score of 3). Since not all faults were diagnosed at this level and since different faults were diagnosed at this level by different participants, this measure is not well suited to aggregation. Thus, regressions were performed on the interaction between diagnosis time and experience for this measure.

Previous analyses have also used fault compensation time (i.e., the time from the occurrence of a fault until successful trial completion) as a measure of fault performance. However, the methodological differences between Experiments I and II–IV make the use of this performance measure difficult in the context of a cross-experiments analysis. While trials in Experiments II–IV terminated at the end of the start-up phase (simulation at steady state for five consecutive minutes), the majority of Experiment I trials included a start-up phase as well as tuning and shut down phases. Faults could occur either in the start-up or the tuning phase, and a trial was considered to terminate successfully after the new tuning goals had been met and the simulation was brought to a shut down state. Since trial termination is measured over different periods between Experiments I and II–IV, compensation time does not have a consistent meaning across these experiments. Accordingly, it was not used as a performance measure in this investigation.

2.3. Predictors — cognitive style

Cognitive style refers to "psychological dimensions that represent consistencies in an individual's manner of acquiring and processing information" (Ausburn and Ausburn (1978, p. 338) quoted in Jonassen and Grabowski (1993)). Jonassen and Grabowski (1993) indicate that cognitive style is a stable trait, an inherent quality of an individual and not a learned property. Thus, it cannot be taught through training but can instead serve as a basis for operator selection.

Among the many tests of cognitive style that are available, we considered two that appeared to be relevant to our goals (Howie, 1996). The Study Process Questionnaire (SPQ) attempts to categorise learning styles by a merging of learning strategy and motive into an overall learning class (Biggs, 1987). This test is administered using a questionnaire in which participants grade their own study habits and learning motives. Using the SPQ, students are classified as deep, surface, or achieving learners. *Deep* learners tend to study participants and tasks to learn them intimately, even if the level of knowledge they hope to gain shows no promise of immediate benefit. *Surface* learners, on the other hand, tend to learn only as much as is perceived to be needed to demonstrate knowledge, or pass a test. *Achieving* learners tend to be goal-driven, and combine both deep and surface strategies to gain enough knowledge to demonstrate excellence at a task.

The Spy Ring History Test (SRT) assigns participants to one of three cognitive style categories: serialist, holist, or versatile (Pask and Scott, 1972). Participants are classified on their ability to learn and reproduce several "spy ring" communication networks that show developments over a number of years. They are asked to learn the configuration of these "spy rings" one year at a time, and are asked questions that evaluate their ability both to reproduce directly and integrate the "spy rings" from various years. Those who perform best at direct reproduction of the networks are classified as *serialists*, while those who excel at higher levels of information integration are classified as *holists*. *Versatile* learners are able to adopt either a serialist or a holist style to suit the situation. The information provided by the test is richer than just a discrete categorisation, however. The test also results in percentage scores for holist and serialist questions, the highest score of which indicates a participant's cognitive style (if the two scores were within 10% of each other, participants are classed as versatile, (Howie, 1996)). For example, if a participant had a 70% holist score and a 90% serialist score, they would be classified as a serialist, whereas if they had a 70% holist score and a 75% serialist score, they would be classified as versatile. The holist and serialist scores also allow us to compare, for instance, the holist scores of two participants. For example, if one participant had a 90% holist score and another had a 50% holist score, then the former would be considered to be the 'stronger' holist.

The SPQ and SRT are unusual tests of individual differences that have not enjoyed the same wide use as other tests. Nevertheless, we chose them because, out of all of the tests described by Jonassen and Grabowski (1993), these two seemed to best fit the factors that appeared to be responsible for the individual differences we observed in our experiments (Howie, 1996). The present individual differences analyses used both tests as predictors of performance with DURESS II. Data existed for 45 of 60 participants on both the SPQ and SRT, in the following categories:

- discrete holist/serialist classification, as derived from the SRT,
- quantitative holist/serialist overall score, as derived from the SRT,
- quantitative holist score, as derived from the SRT,
- quantitative serialist score, as derived from the SRT,
- discrete learning style classification, as derived from the SPQ.

2.4. Predictors — control variables

There are several other variables not related to cognitive style that could potentially account for some, or perhaps even more, of the variance in participants' performance on the DURESS II simulation. These variables, which encompass both demographic data and experimental manipulations, served as control variables in our search for underlying sources of individual differences.

- *Demographic data*. This category of predictors includes *gender* (male/female), *education level* (undergraduate/master's/Ph.D./post-doctorate), and *education relevance* (determined by the number of physics and thermodynamics courses taken).
- *Experimental manipulations*. This category of predictors includes the type of *interface* used by participants ($P/P+F$/Divided $P+F$), the type of *training* administered (none/AH/Dplayer review/Dplayer review + self-explanation), and the *goal tolerances* for their trials (goal tolerances were set to 2.0°C in Experiment I and to 1.5°C in all other experiments). This category also includes *fault order*. Participants were exposed to routine and non-routine faults, where non-routine faults were actually two interacting routine faults. Since we have data for each of the faults in a non-routine fault sequence, faults were coded based on their position in this sequence. The first fault of a non-routine fault sequence as well as routine faults were coded as "first". The second faults of non-routine fault sequences were coded as "second".

If our speculations about cognitive style are incorrect, then these control variables will account for more of the variance in the data than the SPQ and SRT test scores.

2.5. Data preparation

Seven of the predictors represent categorical data. These data were prepared for use in a linear regression model by assigning $n-1$ indicator variables for the n categories of each qualitative predictor (Neter et al., 1990). For instance, the predictor gender has two qualitative levels (male, female), and so necessitates the creation of one indicator variable. We arbitrarily defined this variable to take on a value of 1 for males and 0 for females. Similar indicator variables were created for the holist/serialist and learning style classifications, education level, interface type, training type, and goal tolerances.

We also suspected that there might be a significant interaction effect between cognitive style and interface. To test for this possibility, ten interaction terms covering all possible permutations of cognitive style (scores and class) and interface were included in the database by multiplying the individual terms.

2.6. Model selection procedure

The performance measures introduced above were regressed on the aforementioned predictors using a stepwise regression procedure. The overall best equations were chosen using an iterative procedure based on criteria derived from Neter et al. (1990). The selected performance measure and all predictors were first processed using the stepwise regression function of the SAS statistical software package. Residual plots and analyses of outliers and influential observations were used to identify problematic observations that could be considered as belonging to a different population than the one under analysis. Since not all outlying cases are necessarily influential, observations were deleted from the model only if the model DFFITS and one or more DFBETAS were greater than 1, unless values close to 1 for both DFFITS and DFBETAS were obtained.[1] If influential observations were identified and deleted, the stepwise procedure and analyses of outliers and influential observations were performed on the reduced data set. All models were also checked for multicollinearity, but as no model had a variance inflation factor greater than 10, remedial action was not necessary. Once the overall best model was selected, residual analyses and tests for normality were performed to ensure that all model assumptions were met. If this was the case, a given model was counted as suitable. For two models, the data were not normally distributed (see below). In both of these cases, appropriate data transformations (Law and Kelton, 1991) were performed. A significance criterion of $\alpha = 0.05$ was adopted.

[1] DFFITS is the standardized difference between a fitted *value* for a given observation when all observations are used in a regression and the predicted value for the given observation when it is omitted from the regression model. DFBETAS is defined as the standardized difference between the estimated regression *coefficient* for a given observation when all observations are used and the regression coefficient obtained when the given observation is omitted.

3. Results

3.1. Regression on trial completion time

A stepwise linear regression on asymptotic TCT included four predictors (Table 2):

- The interaction between the $P + F$ interface and holist SRT score was a *negative* predictor (F (1,42) $= 11.2$, partial $r^2 = 0.21$, $p = 0.002$).
- The indicator variable for the P interface was a *negative* predictor (F (2,41) $= 6.2$, partial $r^2 = 0.10$, $p = 0.02$).
- The interaction between the P interface and a serialist cognitive style was a *positive* predictor (F (3,40) $= 3.4$, partial $r^2 = 0.05$, $p = 0.07$).
- The number of physics courses taken was a *positive* predictor (F (4,39) $= 4.0$, partial $r^2 = 0.06$, $p = 0.05$).

This model is highly significant (F (4,39) $= 7.2$, $p < 0.001$) and accounts for 43% of the variance in the data. It indicates that interactions between cognitive style and interface are significant predictors of TCT. Specifically, the interaction between the $P + F$ interface and a participant's holist score on the SRT (PF × HOL) reduces TCT, while the interaction between the P interface and a serialist cognitive style (P × CSS) increases TCT. Note that the interaction between the $P + F$ interface and a participant's quantitative holist SRT score, as opposed to their categorical cognitive style designation, is the strongest of these predictors. This is notable for two reasons. First, it indicates that quantitative SRT scores are more predictive of performance than a qualitative cognitive style classification. Second, this holist score is independent of serialist ability. Participants who have

Table 2
Prediction equation for trial completion time

Regression on trial completion time (TCT)

$$TCT = 560.6 - 191.5\ PF \times HOL - 101.4\ INTP + 91.3\ P \times CSS + 6.5\ PHYSICS$$

where PF × HOL = interaction between $P + F$ interface and holist SRT score
INTP = 1 for P interface, 0 otherwise
$P \times$ CSS = interaction between P interface and serialist cognitive style
PHYSICS = number of physics courses taken

	Variable			
	PF × HOL	INTP	$P \times$ CSS	PHYSICS
Normalized beta weight	−0.58	−0.48	0.23	0.15
Partial[a] r^2	0.21	0.10	0.05	0.06
F	11.2	6.2	3.4	4.0
p	0.002	0.02	0.07	0.05

[a] Model adjusted $r^2 = 0.37$.

high holist scores may also have high serialist scores, or may even be serialists by classification. According to the prediction equation, this will not detrimentally affect their performance on the $P + F$ interface. The same is not true for the interaction of a serialist cognitive style and the P interface. "Stronger" serialists will experience the same increase in TCT as "weaker" serialists.

The significance of the P interface term is more subtle. By itself, this term indicates that the P interface induces an improvement in TCT. However, a true understanding of this result can only be achieved by interpreting this term in the context of the overall equation. First, the general benefit of the P interface is nearly cancelled for serialists using that interface, as indicated by the $P \times CSS$ interaction. Second, since the mean holist score was 53%, most participants using the $P + F$ interface would experience a benefit (by the $PF \times HOL$ term) equal to that of the P interface. Rather than pointing to performance improvements using the P interface, the inclusion of the P interface term is better regarded as a placeholder for the divided interface, which induced no performance improvement.

The term for the number of physics courses taken is most likely included because of the influence of a participant who claimed to have taken 21 physics courses, compared to an average for all participants of 3.03. The observation for this participant had a DFFITS of -0.96 and physics DFBETA of 0.94, indicating that it is a problematic observation, but not one that is influential enough to remove from the model.

3.2. Regression on completion time variance

A stepwise linear regression on CTV included two predictors (Table 3):

Table 3
Prediction equation for completion time variance

Regression on completion time variance (CTV)

$CTV = 85.1 + 100.3\ P \times CSS - 41.7\ PF \times HOL$
where $P \times CSS$ = interaction between P interface and serialist cognitive style
$PF \times HOL$ = interaction between $P + F$ interface and holist SRT score

	Variable	
	$P \times CSS$	$PF \times HOL$
Normalized beta weight	0.47	-0.24
Partial[a] r^2	0.29	0.05
F	16.6	3.2
p	< 0.001	0.08

[a]Model adjusted $r^2 = 0.31$.

- The interaction between the P interface and a serialist cognitive style was a *positive* predictor (F (1,42) = 16.6, partial $r^2 = 0.29$, $p < 0.001$).
- The interaction between the $P + F$ interface and holist SRT was a *negative* predictor (F (2,41) = 3.2, partial $r^2 = 0.05$, $p = 0.08$).

This model is highly significant (F (2, 41) = 10.4, $p < 0.001$), and accounts for 34% of the variance in the data. It indicates that interactions between cognitive style and interface are also significant predictors of CTV. An interaction between a serialist cognitive style and the P interface ($P \times CSS$) increased completion time variance, while an interaction between holist score and the $P + F$ interface ($PF \times HOL$) decreased completion time variance. Note that the latter interaction is not nearly as strong as the former.

Just as with TCT, serialists have a marked disadvantage when using the P interface, but not when using the $P + F$ interface, and participants with high holist SRT scores do particularly well with the $P + F$ interface.

3.3. Regression on incomplete normal trials

A stepwise linear regression on INT included five predictors (Table 4):

- Holist SRT score was a *negative* predictor (F (1,42) = 11.0, partial $r^2 = 0.21$, $p = 0.002$).
- Wide goal tolerances (i.e., 2.0°C) was a *positive* predictor (F (2,41) = 6.8, partial $r^2 = 0.11$, $p = 0.01$).
- Gender was a significant predictor, with males having fewer incomplete normal trials (F (3,40) = 4.4, partial $r^2 = 0.07$, $p = 0.04$).
- Abstraction hierarchy training was a *positive* predictor (F (4,39) = 6.1, partial $r^2 = 0.08$, $p = 0.02$).
- Dplayer review training was a *positive* predictor (F (5,38) = 6.3, partial $r^2 = 0.08$, $p = 0.02$).

This model is highly significant (F (5,38) = 9.1, $p < 0.001$) and accounts for 55% of the variance in the data. The main predictor of incomplete normal trials is a participant's holist SRT score, although there is no interaction with interface in this case. The term for goal tolerances was meant to identify the effect of the wider goal tolerances of Experiment I, but it is hard to understand why wider goal tolerances would promote a greater number of INTs. Although it is possible that wider goal tolerances promote riskier behaviour on the part of participants, it is more likely that the inclusion of GT2 in this model is the result of some unaccounted for experimental differences between Experiments I and Experiments II–IV.

The inclusion of the gender term in the model is caused by one influential data point. One of the four female participants had 9 INTs, 2.9 standard deviations greater

Table 4
Prediction equation for incomplete normal trials

Regression on incomplete normal trials (INT)

$$\mathrm{INT} = 6.10 - 3.04\ \mathrm{HOLIST} + 2.87\ \mathrm{GT2} - 1.88\ \mathrm{GENMALE} + 1.66\ \mathrm{TRAH} + 1.66\ \mathrm{TRREV}$$

where HOLIST = holist SRT score
GT2 = 1 if 2°C goal tolerances, 0 if 1.5° goal tolerances
GENMALE = 1 if male, 0 if female
TRAH = 1 if AH training, 0 otherwise
TRREV = 1 if Dplayer review training, 0 otherwise

	Variable				
	HOLIST	GT2	GENMALE	TRAH	TRREV
Normalized beta weight	− 0.39	0.45	− 0.29	0.35	0.31
Partial[a] r^2	0.21	0.11	0.07	0.08	0.08
F	11.0	6.8	4.4	6.1	6.3
p	0.002	0.01	0.04	0.02	0.02

[a]Model adjusted $r^2 = 0.49$.

than the mean for all participants. Although the DFFITS for this observation was 1.10, none of its DFBETAS were large enough to justify dropping this observation from the model. Thus, it is not possible to make any generalisations about the effects of gender on performance from this result.

This model also indicates that participants in the Dplayer review training group had a relatively large number of INTs. This confirms the results of Howie and Vicente (1998), who reported that participants in the self-explanation group had fewer incomplete trials than participants in the no training and Dplayer review training groups. The inclusion of the AH training term in the model can be accounted for in a different manner. Hunter et al. (1996) attributed the relatively poor performance of the AH training group to an initially low level of ability when compared to the control group. In other words, it is quite possible that it was not the AH training that is being indicated in this model, but rather a group whose overall INT performance was poor to begin with.

3.4. Regression on number of fault detections

A stepwise linear regression on NFD included three predictors (Table 5):

- Undergraduate education level was a *negative* predictor (F (1,43) = 13.6, partial $r^2 = 0.24$, $p = 0.001$).
- Holist SRT score was a *positive* predictor (F (2,42) = 8.7, partial $r^2 = 0.13$, $p = 0.005$).
- Dplayer review training was a *positive* predictor (F (3,41) = 4.2, partial $r^2 = 0.06$, $p = 0.05$).

This model is highly significant (F (3,41) = 10.2, $p < 0.001$) and accounts for 43% of the variance in the data. The main predictor is education level, with the undergraduates being generally able to detect fewer faults than other participants. The second predictor is holist SRT score. Participants with higher holist SRT scores were able to detect a greater number of faults. The third predictor of NFD is the type of training given to participants. Participants in the Dplayer review and self-explanation group were able to detect a greater number of faults than participants with all other types of training This result confirms that obtained previously by Howie and Vicente (1998), who attributed it to a deeper knowledge often gained by participants who engage in self-explanation (see also Chi et al., 1994)

Table 5
Prediction equation for number of fault detections

Regression on number of fault detections (NFD)

$$\mathrm{NFD} = 5.83 - 4.52\ \mathrm{EDUG} + 4.27\ \mathrm{HOLIST} + 2.14\ \mathrm{TRREVSE}$$

where EDUG = 1 if undergraduate, 0 otherwise
HOLIST = holist SRT score
TRREVSE = 1 if Dplayer review + self-explanation training, 0 otherwise

	Variable		
	EDUG	HOLIST	TRREVSE
Normalized beta weight	− 0.51	0.34	0.24
Partial[a] r^2	0.24	0.13	0.06
F	13.6	8.7	4.2
p	0.001	0.005	0.05

[a]Model adjusted $r^2 = 0.39$.

Notably absent from this model is a term for interface. In absolute terms, participants using the $P + F$ interface were able to detect more faults than either the P or the divided interfaces. Nonetheless, the partial correlation between interface and NFD was not strong enough to necessitate including an interface term in the model.

3.5. Regressions on diagnosis accuracy

Two regressions were performed on the dependent variable diagnosis accuracy (DA). The first analysis included all observations, and the resulting model indicates which individual differences affect overall ability to detect and diagnose faults. The second analysis includes only those observations for which DA $\geqslant 1$. Thus, the resulting model indicates which individual differences affected accuracy of diagnosis *given that a fault had been detected*.

When interpreting the results that follow, it must be appreciated that diagnosis accuracy lies on an ordinal (as opposed to an interval) scale. Strictly speaking, we can make no assumptions about the relative distance between, say, diagnosis scores of 0 and 1 versus the distance between scores of 2 and 3. Accordingly, r^2 and normalized beta weights have little meaning, and thus are not reported. F-tests and significance values, both for individual parameters and for the overall model, are still meaningful and hence are reported.

A stepwise linear regression on DA for all observations included six predictors, and was highly significant (F (6,439) = 24.6, $p < 0.001$) (Table 6):

- The interaction between the $P + F$ interface and holist SRT score was a *positive* predictor (F (1,444) = 32.9, $p < 0.001$).
- Deep learning style was a *positive* predictor (F (2,443) = 31.4, $p < 0.001$).
- Dplayer review training was a *negative* predictor (F (3,442) = 31.2, $p < 0.001$).
- Fault sequence was a *negative* predictor (F (4,441) = 28.3, $p < 0.001$).
- The interaction between the P interface and a serialist cognitive style was a *negative* predictor (F (5,440) = 26.4, $p < 0.001$).
- The interaction between the P interface and a holist cognitive style was a *negative* predictor (F (6,439) = 25.6, $p < 0.001$).

Just as in the regressions for TCT and CTV, interactions between cognitive style and interface are significant predictors of a participant's diagnosis score. An interaction between the $P + F$ interface and a participant's holist SRT score was the most significant predictor and results in an increase in diagnosis score. Interactions between the P interface and a holist or a serialist cognitive style decrease diagnosis scores. This model also indicates that participants with a deep learning style were better able to detect faults and to diagnose them at a higher level. The fault sequence term indicates that participants had a more difficult time diagnosing the second fault of a non-routine fault sequence than either routine faults or the first fault of a non-routine fault. As the second faults of the non-routine fault sequences were designed to be difficult to diagnose and compensate for, this result is not surprising. The negative impact of Dplayer review training on diagnosis score confirms the results of Howie and Vicente (1998), who found that the Dplayer review group had the largest proportion of diagnosis scores of 0.

A second regression on DA was performed for all observations that had DA $\geqslant 1$. This equation included three predictors and is also highly significant (F (3, 326) = 16.3, $p < 0.001$) (Table 7):

- The interaction between the $P + F$ interface and a holist SRT score was a *positive* predictor (F (1,328) = 25.7, $p < 0.001$).

Table 6
Prediction equation for diagnosis accuracy for all observations

Regression on diagnosis accuracy (DA) for all observations

$$\mathrm{DA} = 1.87 + 0.76\ \mathrm{PF} \times \mathrm{HOL} + 0.60\ \mathrm{LSDEEP} - 0.90\ \mathrm{TRREV} - 0.46\ \mathrm{SEQUENCE} - 0.86\ P \times \mathrm{CSS} - 0.89\ P \times \mathrm{CSH}$$

where PF × HOL = interaction between $P + F$ interface and holist SRT score
LSDEEP = 1 if deep learning style, 0 otherwise
TRREV = 1 if Dplayer review training, 0 otherwise
SEQUENCE = 1 if normal fault or first fault of non-routine fault sequence
$P \times$ CSS = interaction between P interface and serialist cognitive style
$P \times$ CSH = interaction between $P + F$ interface and holist cognitive style

	Variable					
	PF × HOL	LSDEEP	TRREV	SEQUENCE	$P \times$ CSS	$P \times$ CSH
F	32.9	31.4	31.2	28.3	26.4	25.6
p	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001	< 0.001

Table 7
Prediction equation for diagnosis accuracy for observations with DA $\geqslant 1$

Regression on diagnosis accuracy (DA) for observations with DA $\geqslant 1$			
DA = 1.442 + 0.83 PF × HOL + 0.48 LSDEEP + 0.45 PHYSICS where PF × HOL = interaction between $P + F$ interface and holist SRT score LSDEEP = 1 if deep learning style, 0 otherwise PHYSICS = number of physics courses taken			
	Variable		
	PF × HOL	LSDEEP	PHYSICS
F	25.7	25.8	20.6
p	< 0.001	< 0.001	< 0.001

Table 8
Prediction equation for fault detection time

Regression on fault detection time (TDETECT)		
TDETECT = exp (4.21 − 1.01 PF × HOL − 0.058 FAULT) where PF × HOL = interaction between $P + F$ interface and holist SRT score FAULT = fault number		
	Variable	
	PF × HOL	FAULT
Normalized beta weight	− 0.27	− 0.16
Partial[a] r^2	0.07	0.03
F	25.5	10.1
p	< 0.001	0.002

[a] Model adjusted $r^2 = 0.09$.

- Deep learning style was a *positive* predictor (F (2,327) = 25.8, $p < 0.001$).
- The number of physics courses taken was a *positive* predictor (F (3,326) = 20.6, $p < 0.001$).

Again, the interaction between the $P + F$ interface and a participant's holist SRT score is the most significant predictor of performance. In this case, given that they had detected the fault, participants with a high holist SRT score using the $P + F$ interface were more likely to diagnose faults at a higher level. Given that a fault has been detected, a deep learning style also positively affects fault diagnosis. Note that these two predictors correspond to the results for the regression on DA with all observations.

Participants who took a greater number of physics courses were also more likely to diagnose faults at a deeper level. Although it is tempting to draw conclusions from this result about the effects of prior knowledge on performance, we are cautious of reading too much into this result. It may be caused by a poorly worded question on the demographic test administered to all participants that left open to interpretation whether a half- or a full-year course constituted one course for the purposes of the test. Due to these difficulties in interpretation, we cannot be sure that these numbers accurately reflect the relative amount of physics knowledge possessed by a participant.

3.6. Regression on fault detection time

Residual analyses of all of the regressions performed up to this stage confirmed that the data for each model were normally distributed. The data for fault detection and diagnosis time were not normally distributed, but rather are best modelled by a lognormal distribution (Law and Kelton, 1991). Accordingly, these data were transformed to their natural logarithms for the analyses that follow. Normal probability plots of the regression residuals confirmed that this transformation was valid for both analyses.

A regression on the natural logarithm of fault detection time included two predictors (Table 8):

- The interaction between the $P + F$ interface and holist SRT score was a *negative* predictor (F (1,337) = 25.5, partial $r^2 = 0.07$, $p < 0.001$).
- Fault number was a *negative* predictor (F (2,338) = 10.1, partial $r^2 = 0.03$, $p = 0.002$).

This model accounts for only 10% of the variance in the data. Nevertheless, it is highly significant (F (2, 338) = 18.1, $p < 0.001$). Although the variance accounted for is quite low, the qualitative results reinforce the general trend that can be seen in previous models. Once again, the main predictor of fault detection time is an interaction between cognitive style and interface. In this case, the interaction between a participant's holist SRT score and the $P + F$ interface is associated with reduced fault detection time. The significant fault variable indicates that fault detection times improved with experience. These two trends can be seen in Fig. 1, which depicts predicted fault detection time as a function of holist SRT score and fault number (or, experience).

3.7. Regression on fault diagnosis time

A regression on the natural logarithm of fault diagnosis time included three predictors:

- The $P + F$ interface was a *negative* predictor (F (1,159) = 14.7, partial $r^2 = 0.09$, $p < 0.001$).
- The number of physics courses taken was a *negative* predictor (F (2,158) = 7.0, partial $r^2 = 0.04$, $p < 0.01$).
- The trial number at which the fault occurred was a *negative* predictor (F (3,157) = 4.9, partial $r^2 = 0.03$, $p = 0.03$).

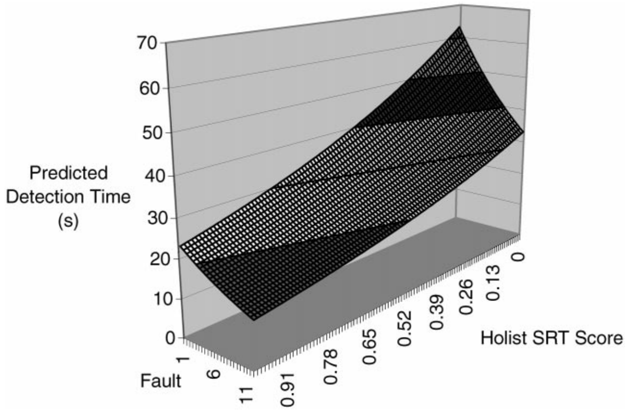

Fig. 1. Predicted fault detection time as a function of holist SRT score and fault number.

The model is highly significant (F (3, 157) = 9.3, $p < 0.001$) and accounts for 15.0% of the variance in the data. The best predictor of fault diagnosis time is interface. Participants using the $P + F$ interface have significantly shorter fault detection times than participants using either the P or the divided interface. This result confirms previous findings (Christoffersen et al., 1997).

Finally, the trial number at which the fault occurred is also a predictor of fault diagnosis time. This result does not indicate an experience effect so much as a difference in fault trial performance between Experiment I and all others. In Experiment I, faults were administered starting at trial 63 while in Experiments II–IV all faults were completed by trial 64. Faults in Experiment I may have been more difficult to diagnose, explaining the longer diagnosis time for later trials predicted by the above equation.

Education relevance is also a predictor of fault diagnosis time. Fault diagnosis time (Table 9) is predicted to decrease as a function of the number of physics courses taken. Again, we are cautious of reading too much into this result as participants seemed to have been inconsistent in their interpretation of what constituted one physics course (see above).

Table 9
Prediction equation for fault diagnosis time

Regression on fault diagnosis time (TDIAG)

TDIAG = exp (5.635 − 0.80 INTPF
− 0.068 PHYSICS − 0.0047 TRIAL
where INTPF = 1 for $P + F$ interface, 0 otherwise
PHYSICS = number of physics courses taken
TRIAL = trial number at which fault occurred

	Variable		
	INTPF	PHYSICS	TRIAL
Normalized beta weight	− 0.28	− 0.22	− 0.17
Partial[a] r^2	0.09	0.04	0.03
F	14.7	7.0	4.9
p	< 0.001	< 0.01	0.03

[a] Model adjusted $r^2 = 0.13$.

4. Discussion

We have presented eight regression analyses of a composite data set from four experiments, with 15 different predictor variables for each analysis. With such a complex data set and so many potentially significant predictors, by sheer chance alone we would expect to find a number of idiosyncratic and unrelated significant effects. We would do well to not put too much weight on such findings. On the other hand, it would be surprising if there were any recurring significant effects in which the same variable is a strong predictor in several different analyses. Such findings would be worthy of our attention. As shown by the summary in Table 10, our results have generated two such recurring patterns.

The strongest finding in these individual differences analyses is the repeated role played by the interaction between the $P + F$ interface and the holist score of the SRT. This interaction, summarised generically in Fig. 2, was a statistically significant predictor in five of the eight regression analyses. In four of these five cases, it was the strongest predictor in terms of variance accounted for. In all five cases, the interaction between $P + F$ and holist score had a beneficial impact on performance. These findings serve to reinforce the advantage of the $P + F$ interface that we have observed individually in each of the experiments comprising our composite data set. More importantly, however, these results provide new and convincing evidence that individuals who have a high holist score and who use the $P + F$ interface are the top performers overall. For some reason, individuals with a lower holist score do not seem to be able to take full advantage of the benefits that the $P + F$ interface has to offer. As we mentioned, these top performers need not be classified as holists (i.e., their serialist score can be even higher). What seems to matter is that they have a high holist score.

The second important finding was the role played by the interaction between the P interface and a serialist cognitive style designation on the SRT. Although this finding was not as strong as the first, it is notable as well. This interaction was a statistically significant predictor in three of the eight regression analyses presented. In one of these three cases, it was the strongest predictor in terms of variance accounted for. In all three cases, the interaction between the P interface and a serialist designation had a negative impact on performance. Thus, it seems

Table 10
Summary of significant regression terms, by dependent variable

Dependent variable	Predictor	Direction	r^2	P
Asymptotic trial	PF × HOL	−	0.21	< 0.01
Completion time	P interface	−	0.10	0.02
	P × CSS	+	0.05	0.07
	No. physics courses taken	+	0.06	0.05
Asymptotic completion	P × CSS	+	0.29	< 0.01
Time variance	PF × HOL	−	0.05	0.08
Incomplete normal trials	Holist SRT score	−	0.21	< 0.01
	Goal tolerances	+	0.11	0.01
	Male gender	−	0.07	0.04
	AH training	+	0.08	0.02
	Dplayer training	+	0.08	0.02
Number of fault detections	Undergraduate education	−	0.24	< 0.01
	Holist SRT score	+	0.13	< 0.01
	Dplayer training	+	0.06	0.05
Diagnosis accuracy for all faults[a]	PF × HOL	+		< 0.01
	Deep learning style	+		< 0.01
	Dplayer training	−		< 0.01
	Fault sequence	−		< 0.01
	P × CSS	−		< 0.01
	P × HOL	−		< 0.01
Diagnosis accuracy for observations with DA $\geqslant$ 1[a]	PF × HOL	+		< 0.01
	Deep learning style	+		< 0.01
	No. physics courses taken	+		< 0.01
Fault detection time	PF × HOL	−	0.07	< 0.01
	Fault number	−	0.03	< 0.01
Fault diagnosis time	$P + F$ interface	−	0.09	< 0.01
	No. physics courses taken	−	0.04	< 0.01
	Trial number	−	0.03	0.03

[a]Since diagnosis accuracy was measured on an ordinal scale, the r^2 statistics have little meaning and are therefore not reported for these dependent variables.

that individuals who are categorised as serialists (i.e., whose serialist score on the SRT exceeds their holist score by more than 10 points) and who use the P interface do particularly poorly. Note that these individuals can still have a high holist score, as long as their serialist score was even higher.

As we will discuss below, the recurring significant, beneficial interaction between the $P + F$ interface and holist score has important practical implications. Is there any way to explain why this result was obtained? Recall that one of the features that distinguishes the $P + F$ from the P interface is that it presents functional information as well as physical information. Interestingly, this functional information is primarily *relational* in nature. That is, functional information shows how the individual physical variables are actually related to each other by the higher-order, goal-relevant constraints identified by an abstraction hierarchy analysis (Vicente and Rasmussen, 1990). It seems to follow that to benefit from an interface with such information, participants must be proficient in systems thinking so that they can think relationally and get a "big picture" understanding of what is going on in the process. This seems to be a relatively straightforward implication that follows from the characteristics of the $P + F$ interface. The important applied question is whether such systems thinking can be taught or whether it is a stable trait of an individual.

As mentioned earlier, cognitive style is a stable trait that cannot be trained. In other words, we cannot create serialists, holists, or versatile individuals. The work of Pask and Scott (1972) tells us that some individuals (i.e., holists) have a natural tendency to engage in relational thinking, whereas other individuals (i.e., serialists) do not. Thus, training a serialist to make the most effective use of

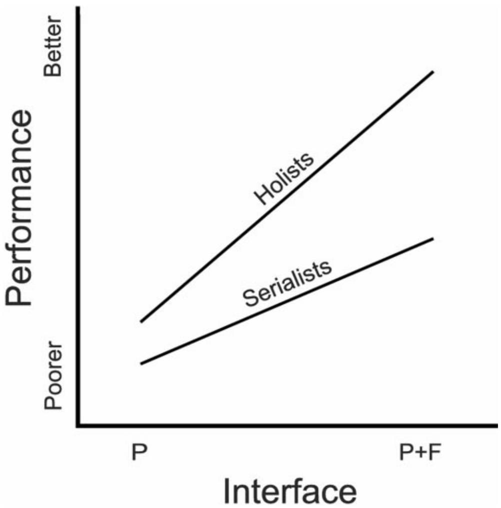

Fig. 2. Generic summary of the primary findings consistently observed across the various regression analyses.

an EID interface, while it cannot be completely ruled out, seems to be an unlikely possibility. Furthermore, creating an alternative interface format that would enhance the performance of a serialist to the same level as that observed with an EID interface also cannot be completely ruled out, but also seems to be an unlikely possibility. Participants using an interface based on EID achieve a level of performance that exceeds that observed with a more traditional interface format. Even serialists tended to perform better on the $P + F$ interface than on the P (see Fig. 2). Therefore, the set of empirical findings presented above, while certainly not definitive, suggests the following interpretation. A strong holist ability is required to take full advantage of the functional information in the $P + F$ interface. The stronger the holist score, the better an individual performs with the $P + F$ interface. Moreover, the level of performance achieved by this cognitive style $\times$ interface combination exceeds that obtained by any other combination that has been explored so far.

5. Conclusions

The individual differences analyses described in this paper were motivated by an applied concern. Specifically, we were interested in knowing whether we can "make the most" of an EID interface by training alone, or whether we have to resort to selection criteria. Note that making different interfaces available to different operators as a function of cognitive style does not appear to be a viable option because participants using an interface based on EID achieve a level of performance that exceeds that observed with more traditional interface formats. There is no trade-off in this respect.

Our results show that the holist/serialist cognitive style is a statistically significant predictor of the performance variability we had observed between participants. As with most studies of individual differences, the proportion of variance accounted for is not large (Stanton and Ashleigh, 1996), in our case ranging from approximately 5% to 21%. Nevertheless, these analyses provide, for the first time, an empirical link between the holist/serialist cognitive style and performance with an interface based on the principles of EID. The top performers were those who were given an EID interface *and* who had high holist scores. It seems that these individuals had the relational thinking skills that are required to interpret the higher-order functional information presented in an EID interface. If generalisable, this result has important implications for the selection of operators. To make the most of an EID interface, operators should be selected on the basis of their holist tendencies. This conclusion does not mean that training is not important. In fact, in previous studies (Hunter et al., 1996; Howie and Vicente, 1998), we showed that two different types of instruction can also lead to improved performance. There is no reason why training and selection cannot be used in tandem. Thus, a more accurate interpretation of these results is that training alone is not enough. There is a substantial proportion of the variance in performance that can only be attributed to cognitive style. Therefore, to make the most of EID, it seems that we should make sure that operators have strong holist tendencies, in addition to receiving suitable training.

5.1. Limitations

Despite these important contributions, these results have several important limitations. First, the analyses were conducted a posteriori on data from experiments that were originally designed for different purposes. It would be useful to conduct an experiment with several groups of participants that were clearly distinguished categorically according to cognitive style criteria (e.g., high/low serialists vs. high/low holists), thereby allowing us to test a priori hypotheses. Second, the sample size of these analyses is low compared to the norm in studies of individual differences (although the complexity of the task and the duration of the studies are much greater and longer, respectively, than the norm). Thus, the reliability of the results could be improved. Third, this research was performed in the context of a simulated microworld. While this microworld was designed to be representative (Brunswik, 1956) of complex domains (see Vicente, 1991), it is far from the level of complexity found in industry-scale plants. Finally, while the participants in these studies had extensive experience, they were not professional operators. Thus, future research should see if the results obtained here generalise to professional operators working with an industry-scale process plant.

Acknowledgements

This research project was sponsored by a contract from the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, contract monitor) as well as research and equipment grants from the Natural Sciences and Engineering Research Council of Canada. We would like to thank Dr. Tanabe his contributions and help, and the two reviewers for their constructive suggestions.

References

Ausburn, L.J., Ausburn, F.B., 1978. Cognitive styles: some information and implications for instructional design. Educ. Commu. Techn. J. 26, 337–354.

Biggs, J.B., 1987. Student Approaches to Learning. Australian Council for Educational Research, Hawthorn, Australia.

Brunswik, E., 1956. Perception and the Representative Design of Psychological Experiments 2nd edition. University of California Press, Berkeley, CA.

Chi, M.T.H., DeLeeuw, N., Chiu, M.H., LaVancher, C., 1994. Eliciting self-explanations improves understanding. Cognitive Sci. 18, 439–477.

Christoffersen, K., Hunter, C.N., Vicente, K.J., 1996. A longitudinal study of the effects of ecological interface design on skill acquisition. Hum. 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. Int. J. Cognitive Ergon. 1, 1–24.

Christoffersen, K., Hunter, C.N., Vicente, K.J., 1998. A longitudinal study of the impact of ecological interface design on deep knowledge. Int. J. Hum-Comput. Stud. 48, 729–762.

Crossman, E.R.F.W., Cooke, J.E., 1962/1974. Manual control of slow response systems. In: Edwards, E., Lees, F.P. (Eds.), The Human Operator and Process Control, Taylor and Francis, London, pp. 51–66.

Howie, D.E., 1996. Shaping expertise through ecological interface design: metacognition and individual differences. (CEL 96-01). University of Toronto, Cognitive Engineering Laboratory, Toronto.

Howie, D.E., Vicente, K.J., 1998. Making the most of ecological interface design: The role of self-explanation. Int. J. Hum-Comput. Stud. 49, 651–674.

Hunter, C.N., Vicente, K.J., Tanabe, F., 1996. Can 'theoretical' training improve fault management performance? In Proceedings of the 1996 American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Control and Human-Machine Interface Technologies, ANS, La Grange Park, IL, pp. 683–690.

Janzen, M.E., Vicente, K.J., 1998. Attention allocation within the abstraction hierarchy. Int. J. Hum-Comput. Stud. 48, 521–545.

Jonassen, D.H., Grabowski, B.L., 1993. Handbook of Individual Differences, Learning, and Instruction. Erlbaum, Mahwah, NJ.

Law, A.M., Kelton, W.D., 1991. Simulation Modelling and Analysis. McGraw-Hill, New York.

Neter, J., Wasserman, W., Kutner, M.H., 1990. Applied Linear Statistical Models. Richard D. Irwin, Toronto.

Olsen, S.E., Rasmussen, J., 1989. The reflective expert and the prenovice: notes on skill-, rule-, and knowledge-based performance in the setting of instruction and training. In: Bainbridge, L., Quintanilla, S.A.R. (Eds.), Developing Skills with Information Technology. Wiley, Chichester, UK, pp. 9–33.

Pask, G., Scott, B.C., 1972. Learning styles and individual competence. Int. J. Man-Mach. Stud. 4, 217–253.

Pawlak, W.S., Vicente, K.J., 1996. Inducing effective operator control through ecological interface design. Int. J. Hum. Comput. Stud. 44, 653–688.

Rasmussen, J., 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. IEEE Trans. Systems, Man, Cybernet. SMC-13, 257–266.

Rasmussen, J., 1985. The role of hierarchical knowledge representation in decision making and system management. IEEE Trans. Systems, Man, Cybernet. SMC-15, 234–243.

Rasmussen, J., Pejtersen, A.M., Goodstein, L.P., 1994. Cognitive Systems Engineering. Wiley, New York.

Stanton, N., Ashleigh, M., 1996. Selecting personnel in the nuclear power industry. In: Stanton, N. (Ed.), Human factors in nuclear safety. Taylor and Francis, London.

Vicente, K.J., 1991. Supporting knowledge-based behavior through ecological interface design. Unpublished Doctoral Dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

Vicente, K.J., 1997. Operator adaptation in process control: a three-year research program. Control Eng. Practice 5 (3), 407–416.

Vicente, K.J., 1999. Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work. Erlbaum, Mahwah, NJ.

Vicente, K.J., Christoffersen, K., Hunter, C.N., 1996. Response to Maddox critique. Hum. Factors 38, 546–549.

Vicente, K.J., Rasmussen, J., 1990. The ecology of human — machine systems II: Mediating 'direct perception' in complex work domains. Ecol. Psychol. 2, 207–250.

Vicente, K.J., Rasmussen, J., 1992. Ecological interface design: theoretical foundations. IEEE Trans. Systems, Man, Cybernet. SMC-22, 589–606.