Making the Most of Ecological Interface Design: The Role of Cognitive Style

Gerard L. Torenvliet, Greg A. Jamieson, & Kim J. Vicente
Cognitive Engineering Laboratory
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, Ontario, Canada

Abstract

A composite database was created by compiling data obtained from 45 participants in four experiments with the DURESS II microworld. This database consisted of measures of performance, demographic data, a summary of the experimental manipulations conducted in these studies, 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 ecological interface design (EID). Thus, 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 in an EID interface. This empirical result has important implications for future research on EID, and more importantly, for 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 will need different characteristics from current operators. Since advanced control rooms are qualitatively different from more traditional designs, it would not be surprising to find that a different set of competencies are required to be an effective operator in an advanced control room. If the need for a change in operator characteristics is established, then the immediate question becomes whether such characteristics can be acquired through an appropriate training program or whether they must be selected for. 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. Several studies have shown the advantages of an interface based on the principles of ecological interface design (EID). However, at the same time, substantial individual differences have been observed as well. Some individuals using an EID interface perform much more effectively than others. These individual differences have been observed in studies where there has been no training [3,6], as well as in studies where participants have received some form of instruction [6,7]. These results suggest that there may be one or more traits that predispose an individual to perform well with an EID interface. This is an issue 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. Given this choice, however, 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 investigates the latter possibility by describing a series of regression analyses designed to analyse the individual differences observed in previous studies we have conducted. Our goal is 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 will first briefly describe the details of the data from the four studies that comprise the database for these analyses. All of these studies have been conducted in the context of the DURESS II testbed, a simulation of a highly simplified yet representative thermal hydraulic plant. DURESS II can be controlled using either a conventional interface

displaying only physical information (P interface) or 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, see [14].)

Using DURESS II, each experiment was designed to investigate the effects on operator adaptation of modifying one or more behaviour shaping constraints [12]. Generally speaking, behaviour shaping constraints are any type of constraint that may shape how operators adapt to a work domain. The specific behaviour shaping constraints relevant to this program of research are:

- *Interface Content.* Interface content is a strong constraint on operator performance [3]. 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, or 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 that operators receive influences adaptation. Operators can receive training either prior to or concurrent with operating a system. 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., [4]) 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 subjects in each experiment have pre-existing competencies that influence adaptation. These take on such forms as cognitive style, declarative knowledge, perceptual and motor skills, and population stereotypes. 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. These experiments are described in the text below.

1. *Experiment I* [3]: 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 subjects operated either the P or the P+F interface of the DURESS II microworld quasi-daily over a period of six months (217 trials).
2. *Experiment II* [7]: This experiment investigated the interaction between interface design and model-based training on adaptation. Twenty-four subjects participated in a 2 x 2 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).
3. *Experiment III* [6]: This experiment investigated the impact of interface form on adaptation. The P+F interface of DURESS II was compared to a divided P+F interface with one level of information for each level of the abstraction hierarchy. 12 subjects, divided into two groups (integrated vs. divided), participated in this experiment for one month (67 trials) each.
4. *Experiment IV* [6]: This experiment investigated the

Experiment	*Behaviour Shaping Constraints*				Number of Subjects
	Interface Content	Interface Form	Type of Training	Pre-existing Competencies	
I	P vs. P+F	P vs. P+F	None	Demographic Data + Cognitive Style	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		12*
IV	P+F	P+F	None vs. Dplayer vs. Dplayer + SE		18*

*Experiments III and IV made use of a shared control group of six subjects. In total, 24 subjects participated in these experiments.

Table 1. Integrated summary of the three-year research program, showing the behaviour shaping constraints investigated in each experiment. (Adapted from [15].)

effect of a second type of training, self instruction via performance reviews and/or self-explanation, on operator performance. 18 subjects were divided 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. 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 the Dplayer program, a 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-subjects 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. Furthermore, these individual differences also have methodological implications for research. The high variability between subjects results in statistical analyses with low power, thereby making it difficult to separate the effects of deliberate experimental manipulations from the effects of individual differences. If we knew the source of these individual differences, then we could make sure that different experimental groups were evenly matched to reduce the variability across groups, and we could select subjects within a restricted range of the relevant trait(s) to reduce the variability within groups. Both of these steps could result in an increase in statistical power, and thus improve our chances of designing more sensitive experiments. Therefore, there are several reasons why it is important to uncover the root cause of the individual differences in performance observed in previous studies.

Preliminary analyses [5] indicate that the holist/serialist cognitive style distinction [10] is relevant to understanding adaptation in the DURESS II microworld. From previous experiments, a comprehensive database exists containing relevant information for almost all subjects to support a more comprehensive analysis of individual differences. It is hoped that this analysis of individual differences will help to identify how the holist/serialist cognitive style distinction affects adaptation in the DURESS II microworld. This understanding should help in solving the problems identified above.

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, however, only those which were relevant are reported.

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 one of the motivations for this research was to aid in the selection and groupwise matching of subjects to reduce the effects of individual differences, specific performance measures were selected from these that are relevant to this aim.

- *Asymptotic Trial Completion Time (TCT).* Asymptotic TCT is measured as the average amount of time required for a subject to bring the system 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 subject violated any of the system's constraints, causing the system to 'blow up.'

2.2: Performance measures for fault performance

Faults were administered to subjects at random intervals over the duration of each experiment. Since there were 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. In order to realise the benefits of both types of analyses while at the same time compensating for their deficiencies, two types of individual differences analyses on fault data were performed. The first type is an aggregate analysis in which the various measures of fault performance (described below) were aggregated over all faults for each subject. 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 experience. The second type of analysis considers each fault of each subject 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 subjects verbalised that there was a problem with the system behaviour, 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 subject'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 subjects' diagnoses, ranging from a score of 0 for an irrelevant utterance to 3 for a statement of the location and root cause of the fault [11]. 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 time elapsed between the occurrence of a fault and the subject'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 subjects, this data is not well suited to aggregation. Rather, regressions were performed on the interaction between diagnosis time and experience.

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 (plant 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 successfully terminate after the new tuning goals had been met and the plant was brought to a shut down state. Since trial termination is measured over different periods between experiment I and II-IV, compensation time does not have a consistent meaning across these experiments. Accordingly, fault compensation time was not used as a performance measure in this investigation.

2.3: Predictors — cognitive style

Two tests of cognitive style have been proposed separately by Biggs [1] and Pask & Scott [10]. Biggs proposed the Study Process Questionnaire (SPQ) which attempts to categorise learning styles by a merging of learning strategy and motive into an overall learning class. This test is administered using a questionnaire in which subjects 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 subjects 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 needed to demonstrate knowledge, or pass a test. *Achieving* learners tend to be goal driven, and combine both deep and surface strategies in order to gain enough knowledge to demonstrate excellence at a task.

Pask & Scott's Spy Ring History Test (SRT) assigns subjects to one of three cognitive style categories: serialist, holist, or versatile. Subjects are classified on their ability to learn and reproduce several "spy ring" communication networks that show developments over a number of years. Subjects are asked to learn the configuration of these "spy rings" one year at a time, and are asked questions which rank 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 subjects' cognitive style (if the two scores were within 10% of each other, subjects are classed as versatile [5]). An average overall score can also be calculated. These different scores allow us to compare, for instance, the holist scores of two subjects to determine which is the 'stronger' holist.

The present individual differences analyses used both tests as predictors of performance with DURESS II. Data exists for 45 of 60 subjects 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

As we now have SPQ and SRT data for a large number of subjects, these analyses should have a greater power than the preliminary analyses we conducted [5].

2.4: Predictors — control variables

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

1. **Demographic data:**
- *Gender* (male / female)
- *Education level* (undergraduate / master's / Ph.D. / post-doctorate)
- *Education relevance* (number of physics and thermodynamics courses taken)

2. **Experimental manipulations:**
- *Interface*. The type of interface subjects used.
- *Training.* The type of training (none, AH, Dplayer review, or Dplayer review + self-explanation) that subjects received.
- *Goal Tolerances.* Tolerances around the goals were set to 2.0° C in Experiment I and to 1.5° C in all other experiments.
- *Fault Order*. Subjects were exposed to two types of faults, routine and non-routine, where non-routine faults were actually two interacting faults. For the purpose of this analysis, faults were coded based on their position in a fault sequence. Routine faults and the first fault of a non-routine fault were coded as first in a sequence, while the second faults of non-routine faults were coded as second in a sequence.

2.5: Data preparation

Seven of the predictors represent qualitative 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 [9]. 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.

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.* [9]. 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 which 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 [9]. 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 maximum variance inflation factor of greater than 10, no remedial action was necessary [9]. 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. In two cases (see below), this was not the case. In both of these cases, appropriate data transformations [8] solved the problem.

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=0.017).
- 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=0.072).
- The number of physics courses taken was a *positive* predictor (F (4,39)=4.0, partial r^2=0.06, p=0.054).

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 subject's holist score on the SRT (PFXHOL) reduces TCT, while the interaction between the P interface and a serialist

Regression on Trial Completion Time (TCT)

TCT = 560.6 – 191.5 PFXHOL –101.4 INTP + 91.3 PXCSS + 6.5 PHYSICS

where
- PFXHOL = interaction between P+F interface and holist SRT score
- INTP = 1 for P interface, 0 otherwise
- PXCSS = interaction between P interface and serialist cognitive style
- PHYSICS = number of physics courses taken

	Variable			
	PFXHOL	*INTP*	*PXCSS*	*PHY-SICS*
Normalized Beta Weight	-0.58	-0.48	0.23	0.15
*Partial 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

*model adjusted r^2 = 0.37

Table 2. Prediction equation for Trial Completion Time.

cognitive style (PXCSS) increases TCT. Note that the interaction between the P+F interface and a subject'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. Subjects who have 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 PXCSS interaction. Second, since the mean holist score was 53%, most subjects using the P+F interface would experience a benefit (by the PFXHOL 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 subject who claimed to have taken 21 physics courses, compared to an average for all subjects of 3.03. The observation for this subject had a DFFITS of –0.9660 and physics DFBETA of 0.9380, 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):

- 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 score (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 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 (PXCSS) increased completion time variance, while an interaction between holist score and the P+F interface (PFXHOL) 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 subjects with high holist SRT scores do particularly well with the P+F interface.

Regression on Completion Time Variance (CTV)

CTV = 85.1 + 100.3 PXCSS – 41.7 PFXHOL

where
- PFXHOL = interaction between P+F interface and holist SRT score
- PXCSS = interaction between P interface and serialist cognitive style

	Variable	
	PFXHOL	*PXCSS*
Normalized Beta Weight	0.47	-0.24
*Partial r^2**	0.29	0.05
F	16.6	3.2
p	< 0.001	0.08

*model adjusted r^2 = 0.31

Table 3. Prediction equation for Completion Time Variance.

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).
- Male gender was a *negative* predictor (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 subject'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 operators, it is more likely that the inclusion of GT2 in this model is the result of some unaccounted for experimental differences between experiment 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 subjects had 9 INTs, which is 2.9 standard deviations greater than the mean for all subjects. 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 subjects in the Dplayer review training group had a relatively large number of INTs. This confirms the results of Howie *et al.*, who reported that "overall, participants in the [self-explanation] group had fewer incomplete trials than subjects in the [no training] and [Dplayer review training] groups" [6, p. 44]. The inclusion of the AH training term in the model can be accounted for in a different manner. Hunter *et al.* [7] attribute 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 in this model is education level, and the model indicates that undergraduates were generally able to detect fewer faults than other subjects. The second predictor is holist SRT score. The model indicates that subjects 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 subjects. The model indicates that subjects in the Dplayer review and self-explanation group were able to detect a greater number of faults than subjects with all other types of training. This

Regression on Incomplete Normal Trials (INT)

INT = 6.10 – 3.04 HOLIST + 2.87 GT2 – 1.88 GENMALE + 1.66 TRAH + 1.66 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 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

*model adjusted r^2 = 0.49

Table 4. Prediction equation for Incomplete Normal Trials.

Regression on Number of Fault Detections (NFD)

NFD = 5.83 – 4.52 EDUG + 4.27 HOLIST + 2.14 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 r^2*	0.24	0.13	0.06
F	13.6	8.7	4.2
p	0.001	0.005	0.05

*model adjusted $r^2 = 0.39$

Table 5. Prediction equation for Number of Fault Detections.

result confirms that obtained previously by Howie *et al.* [6], who attributed this to a deeper knowledge often gained by subjects who engage in self-explanation [2].

Notably absent from this model is a term for interface. In absolute terms, subjects using the P+F interface were able to detect more faults than either the P or the divided interfaces. Nonetheless, the 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. The first analysis included all observations, and the resulting model indicates what individual differences affect overall ability to detect and diagnose faults. The second analysis includes only those observations for which DA ≥ 1. Thus, the resulting model indicates what 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. As this is the case, r^2 and normalized beta weights have little meaning, and 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 subject's diagnosis score. An interaction between the P+F interface and a subject'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 subjects 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 subjects had a more difficult time diagnosing the second fault of a non-routine fault sequence than either routine faults or the first fault of

Regression on Diagnosis Accuracy (DA) for all observations

DA = 1.87 + 0.76 PFXHOL + 0.60 LSDEEP – 0.90 TRREV – 0.46 SEQUENCE - 0.86 PXCSS – 0.89 PXCSH

where PFXHOL = 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
PXCSS = interaction between P interface and serialist cognitive style
PXCSH = interaction between P+F interface and holist cognitive style

	Variable					
	PFXHOL	*LSDEEP*	*TRREV*	*SEQUENCE*	*PXCSS*	*PXCSH*
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 6. Prediction equation for Diagnosis Accuracy for all observations.

a non-routine fault. As the second faults of the non-routine fault sequences were designed to be "difficult to diagnose and compensate for" [6, vol. 2, p. 24], this result is not surprising. The negative impact of Dplayer review training on diagnosis score confirms the results of Howie et al., who reported that "the [Dplayer review] group had the largest proportion of diagnosis scores of 0" [6, vol. 2, p. 49].

A second regression on DA was performed for all observations that had DA ≥ 1. This equation included three predictors and is 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, $p < 0.001$).
- 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 subject's holist SRT score is the most significant predictor of performance. In this case, given that they had detected the fault, subjects 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.

Subjects 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. This may be the result of a poorly worded question on the demographic test administered to all subjects 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 subject.

Regression on Diagnosis Accuracy (DA) for observations with DA ≥ 1

DA = 1.442 + 0.83 PFXHOL + 0.48 LSDEEP + 0.45 PHYSICS

where PFXHOL = interaction between P+F interface and holist SRT score
LSDEEP = 1 if deep learning style, 0 otherwise
PHYSICS = number of physics courses taken

	Variable		
	PFXHOL	*LSDEEP*	*PHYSICS*
F	25.7	25.8	20.6
p	<0.001	<0.001	<0.001

Table 7. Prediction equation for Diagnosis Accuracy for observations with DA ≥ 1.

3.6: Regression on fault detection time

Residual analyses of all of the analyses 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 [8]. Accordingly, the data for fault detection and diagnosis time 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 by this model 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 subject's holist SRT score and the P+F interface is associated with a reduced fault detection time. The significant fault variable indicates that fault detection times improved with experience. These two trends can be seen in Figure 1, which depicts fault detection time as a function of holist SRT score and fault

Regression on Fault Detection Time (TDETECT)

TDETECT = exp (4.21 – 1.01 PFXHOL – 0.058 FAULT)

where PFXHOL = interaction between P+F interface and holist SRT score
FAULT = fault number

	Variable	
	PFXHOL	*FAULT*
Normalized Beta Weight	-0.27	-0.16
*Partial r^2**	0.07	0.03
F	25.5	10.1
p	< 0.001	0.002

*model adjusted r^2 = 0.09

Table 8. Prediction equation for Fault Detection Time

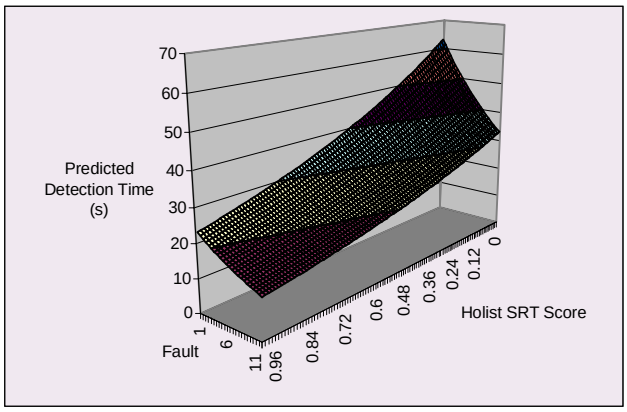

Figure 1. Predicted Fault Detection Time as a function of Holist SRT Score and Fault Number.

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=0.009).
- 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).

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. Subjects using the P+F interface have significantly shorter fault detection times than subjects using either the P or the divided interface. This confirms previous findings [3,6].

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 r^2**	0.09	0.04	0.03
F	14.7	7.0	4.9
p	< 0.001	0.009	0.03

*model adjusted r^2 = 0.13

Table 9. Prediction equation for Fault Diagnosis Time.

Finally, 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 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 subjects seemed to have been inconsistent in their interpretation of what constituted one physics course (see above).

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 certainly be worthy of our attention. Fortunately, 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 was a statistically significant predictor in five of the eight regression analyses we presented. 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 before, 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 that individuals that 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 basic and applied 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. It follows 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. What is not so straightforward is whether such systems thinking can be taught or whether it is a stable trait of an individual.

Fortunately, existing research on cognitive style can help us address this issue. The holist/serialist cognitive style distinction [10] tells us that some individuals (i.e., holists) have a stronger natural tendency to engage in this type of relational thinking than others. Thus, we might expect that a strong holist ability is required to take full advantage of the functional information in the P+F interface. The set of empirical findings presented above support this conjecture. The stronger the holist score, the better an individual performed with the P+F interface.

5: Conclusions

The individual differences analyses described in this paper were motivated by two concerns, one basic and one applied. From a basic research perspective, we were interested in uncovering the root cause of the substantial differences in performance between individuals we had observed in previous studies. From an applied operations perspective, 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. The findings described above shed light on both of these issues.

Our results show that the holist/serialist cognitive style is a statistically significant root cause of the performance variability we had observed between participants. As with most studies of individual differences [13], the proportion of variance accounted for is not large, in our case ranging from approximately 5% to 21%. Nevertheless, having identified the source of some of between-subject variability provides a basis for designing more sensitive experiments in the future. This can be accomplished in two ways. First, we can minimise the variance across groups by making sure that different experimental groups are matched in terms of holist/serialist cognitive style by administering the SRT to all participants. Second, we can minimise the variance within groups by screening subjects according to their holist/serialist cognitive score (e.g., by not including subjects with very high or very low scores). Either of these measures should result in an increase in statistical power.

Perhaps even more importantly, the linear regression 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 generalizable, 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 [6,7] we showed that two different types of instruction can 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, we should make sure that operators have strong holist tendencies, in addition to receiving suitable training.

References

[1] Biggs, J. B. (1987). Student approaches to learning. Hawthorn, Australia: Australian Council for Educational Research.

[2] Chi, M. T. H., DeLeeuw, N., Chiu, M. H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive Science, 18, 439-477.

[3] Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1994). Research on factors influencing human cognitive behaviour (I) (CEL 94-05). Toronto: University of Toronto, Cognitive Engineering Laboratory.

[4] Crossman, E. R. F. W., & Cooke, J. E. (1962/1974). Manual control of slow response systems. In E. Edwards and F. P. Lees (Eds.), The human operator and process control (pp. 51-66). London: Taylor and Francis.

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

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

[7] 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.

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

[9] Neter, J., Wasserman, W., & Kutner, M. H. (1990). Applied Linear Statistical Models. Toronto: Richard D. Irwin.

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

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

[12] Rasmussen, J., Pejtersen, A.M., & Goodstein, L.P. (1994). Cognitive Systems Engineering. New York: John Wiley & Sons.

[13] Stanton, N., & Ashleigh, M. (1996). Selecting personnel in the nuclear power industry. In N. Stanton (Ed.), Human factors in nuclear safety (pp. 159-186). London: Taylor & Francis.

[14] Vicente, K. J. (1996). Improving dynamic decision making in complex systems through ecological interface design: A research overview. Systems Dynamics Review, 12, 4, 251-279.

[15] Vicente, K. J. (1997). Operator adaptation in process control: A three-year research program. Control Engineering Practice, 5, 3, 407-416.