Using Self-Explanation to Exploit Ecological Interface Design

Dianne E. Howie and Kim J. Vicente

Proceedings of the Human Factors and Ergonomics Society Annual Meeting 1997 41: 279
DOI: 10.1177/107118139704100163

USING SELF-EXPLANATION TO EXPLOIT ECOLOGICAL INTERFACE DESIGN

Dianne E. Howie
Centre for Applied Cognitive Science
Ontario Institute for Studies in Education
University of Toronto
Toronto, Ontario, Canada

Kim J. Vicente
Cognitive Engineering Laboratory
Department of Mechanical and Industrial Engineering
University of Toronto
Toronto, Ontario, Canada

This study examined the role of self-generated explanations on performance in a process control microworld, extending previous cognitive science research into a new, applied domain. The experiment was conducted using DURESS II, an interactive, thermal-hydraulic process control simulation. During this one-month experiment, participants controlled the system under normal and fault conditions on a quasi-daily basis. Participants in the self-explanation (SE) group occasionally watched a replay of their own performance immediately after completing a trial, while the control group did not. In addition, the SE group was instructed to explain aloud the reasons for their control actions while watching the replay. The SE participants were divided post-hoc into "good" and "poor" groups according to several performance criteria. An analysis of the protocols produced during self-explanation revealed that "good" SE participants included considerably more words suggesting self-explanations in their protocols than did the "poor" SE participants. Thus, self-explanation was correlated with performance within the SE group.

INTRODUCTION

Ecological interface design (EID) provides a theoretical framework for designing interfaces for complex work environments, such as nuclear power plants (NPPs; Vicente & Rasmussen, 1990, 1992). Previous research supports the assertion that an ecological interface can improve operator performance (Pawlak & Vicente, 1996; Christoffersen, Hunter, & Vicente, 1996a, 1996b, 1997; Hunter et al., 1995). Compared to participants using a traditional interface, operators using an ecological interface perform more consistently under normal conditions and better under unanticipated, abnormal conditions.

Despite the positive overall influence of EID, however, not all operators derive equal benefits from controlling the system using an ecological interface. Some operators demonstrate an exceptional level of performance on both normal and fault trials. In contrast, other operators demonstrate shallow knowledge of the system and relatively poor performance with an ecological interface. Christoffersen et al. (1996b) found that the best participants tended to reflect on the feedback provided by an ecological interface, whereas the worst participants focused only on the surface features of the interface. Does instructing operators to reflect improve their performance with an ecological interface? Research conducted by Chi and colleagues suggests that it may.

Chi, Bassok, Lewis, Reimann, and Glaser (1989) studied the self-generated explanations given by students in think-aloud protocols while studying physics problems. Chi et al. (1989) found that they were able to differentiate good and poor students based on the amounts and types of these explanations. Good students elaborated on reasons when and why a particular strategy should be used, and they were aware of gaps in their understanding to a greater extent than poor students. These findings have been replicated in several different domains including biology, mathematics, and computer programming (Ferguson-Hessler & de Jong, 1990; Lawson & Chinnappan, 1994; Pirolli & Recker, 1994). Further, self-explanation could aid performance even when the explanations were prompted rather than spontaneously generated (Chi, De Leeuw, Chiu, and La Vancher, 1994; Nathan, Mertz, & Ryan, 1994).

This study investigated the benefits of self-explanation for the domain of process control. The study was conducted with DURESS II, an interactive, thermal-hydraulic process simulation. DURESS II was controlled with an interface designed according to the principles of EID (Figure 1).

METHOD

Participants

The participants were 12 university students in science or engineering. The participants completed a demographic questionnaire to determine their age, physics background, and overall level of education; and tests of cognitive style on the holist-serialist dimension (Pask & Scott, 1972). The participants were then matched in pairs according to these criteria, and one member of each pair was assigned to the control and self-explanation (SE) groups.

Procedure

The experiment was approximately one month in duration, including a total of 67 trials per participant on the DURESS II simulation. In each trial, the participant would take the system from a shutdown state to steady state – the startup task. Steady state was achieved when the participant met the temperature and demand goals for 5 consecutive minutes. The participants gave verbal protocols for roughly half of the trials.

Faults were distributed randomly and infrequently throughout the trials. This simulated the unanticipated nature of faults in a natural setting. The faults consisted of valve blockages, heater failures, and reservoir leaks. Each trial had a different steady state demand pair to prevent the participants from adopting excessively simplified control methods. The participants were not informed of the results of their trials, although they could receive feedback at any point during the trial by observing the elapsed time indicated on the interface, and any system failure messages displayed on the screen.

Whenever a participant made a control action, the system recorded the component the participant adjusted, the time, and the state of all system variables. For each trial, these values were stored in a log file, making them available for further analysis.

In addition, participants in the SE group replayed one trial each day using the Dplayer trial replay module. Dplayer allowed the SE participants to view the state of the system on the interface they had used, accompanied by an arrow indicating any control actions. Further, Dplayer allowed the participants to control the pace of replay. During replay, these participants were instructed to explain aloud to the experimenter the reasons for their control actions. The experimenter provided prompted the participants to think aloud when necessary.

RESULTS

The data analysis method used in this experiment follows the examples set by Pawlak & Vicente (1996) and Christoffersen et al. (1997), but is unorthodox and thus requires justification. The emphasis in the design of the experiment was on representativeness (Brunswik, 1956) to improve the generalizability of results to operational settings. This choice greatly increased the complexity and duration of the experiment. Furthermore, many of the analyses were necessarily based on verbal protocols which are notoriously time consuming to analyze. Consequently, only a small number of participants could be included in each group in the time available for the study, resulting in low statistical power. Chi (1997) recommended that, under these conditions, validity should be demonstrated using means other than inferential statistics, preferably through multiple, converging measures. Thus, we analyzed each participant's data in great detail using various measures, and then explicitly summarized the findings bearing across measures.

Trial Completion Times

Trial completion times are the time for each participant to take the system from a shut-down state to steady state, meeting the temperature and demand goals for five consecutive minutes. In order to examine the participants' improvement with experience, the steady state times for the first and last ten normal trials were compared (Table 1).

Table 1. Steady State Times

	Steady State Times (s)			
	First 10 Trials		Last 10 Trials	
Group	Mean	SD	Mean	SD
Control	881.1	384.7	502.3	82.8
SE	821.8	343.5	556.9	212.8

The steady state time decreased between the first ten and last ten trials for both groups, showing an improvement in performance with experience (means of 873.9 seconds for the first block and 526.5 seconds for the last block). The average steady state times were relatively consistent across groups (means of 662.6 and 671.8 seconds for the control and SE groups respectively), but the rate of decrease depended upon the group. The control group showed the greatest improvement in steady state times across trials. The standard deviation of steady state times also decreased for both groups between the first and last blocks of trials. Thus, self-explanation did not produce any overall improvement in performance as measured by trial completion times. However, within the SE group, performance was associated with the amount of self-explanation, as described below.

Ranking

Previous studies have divided self-explanation groups into "good" and "poor" participants, based on a post hoc median split (Chi et al, 1994; Pirolli & Recker, 1994). Performance using DURESS II is a function of multiple criteria on both normal and fault trials, so the rankings of SE participants were determined by combining several measures. The measures were as follow: number of incomplete normal trials, steady state times for the first and last block of normal trials, number of incomplete fault trials, number of faults detected, detection times, diagnosis scores, and compensation times for the first and last block of fault trials. (For more details on these measures, see Howie, Janzen, & Vicente, 1996). The ranks for a SE participant on each measure were summed to provide an overall ranking. The "good" SE participants, in order of rank, were labeled A, B, and C, while the "poor" SE participants were labeled X, Y, and Z. These groupings were used in all of the measures described below.

Time on Task

Self-explanation studies across various domains have all found that good students spend more time working on a self-explanation task than poor students (Chi et al., 1994; Nathan et al., 1994; Pirolli & Recker, 1994). Table 2 summarizes the amount of time each SE participant spent reviewing the first and last three fault trials. On average, the good participants spent more time replaying their trials compared to the poor participants (means of 455 and 241 seconds).

Table 2. Time Spent Replaying Trials

	Good Participants			Poor Participants		
Trial	A	B	C	X	Y	Z
1	665	267	297	357	311	235
2	424	504	1118	326	260	209
3	298	495	535	307	230	184
4	243	220	387	209	168	70
5	359	320	547	377	245	85
6	256	240	476	435	198	70
Average	374 s	361 s	560 s	335 s	235 s	142 s

Amount Spoken

Previous research has also found that good students speak more in their protocols than poor students do (Chi et al., 1994; Nathan et al., 1994; Pirolli & Recker, 1994). Table 3 gives the total number of words in each Dplayer verbal protocol. The good participants uttered almost twice as many words as the poor participants (means of 976 and 564 words).

Table 3. Number of Words in Verbal Protocol

	Good Participants			Poor Participants		
Trial	A	B	C	X	Y	Z
1	1421	519	672	593	443	313
2	-	1036	2145	664	377	364
3	637	915	959	561	414	322
4	549	521	878	414	261	89
5	771	721	1248	801	413	130
6	581	831	1153	913	337	132
Average	792	757	1176	658	374	225

Explanations

Previous studies of self-explanation have found that the verbal protocols of the good students contained more "explanations" than those of poor students (Chi et al., 1989). Broadly, a self-explanation in the context of DURESS II was considered as an utterance that went beyond a direct statement of a control action to explain the reason for that action (for example, "I increased the heat for heater 1 *because* I realized that I hadn't met the objective and that it was climbing at such a slow rate" [emphasis added]). In order to provide a more objective count of self-explanations, a list of key words that indicated an explanation were compiled and counted for each replay verbal protocol. The self-explanation key words were: *because, so, since,* and *reason.* Table 4 shows the frequency with which these words occurred in the verbal protocols.

Table 4. Number of Explanations

	Good Participants			Poor Participants		
Trial	A	B	C	X	Y	Z
1	38	11	18	18	10	6
2	-	40	73	19	7	7
3	21	32	33	14	7	9
4	15	16	21	15	9	0
5	22	31	49	25	10	4
6	15	37	36	32	10	5
Total	111	167	230	123	53	31

The number of self-explanation words for the good and poor participants in the SE group were compared. The good participants included over twice as many self-explanation key words as the poor participants – totals of 508 and 207 words respectively (or averages of 32.9 and 11.6 words per trial respectively).

Monitoring

Another characteristic of good students is that they tend to monitor their understanding of a topic as they study (Pirolli & Recker, 1994). This should be reflected in the replay verbal protocols for DURESS II. Three kinds of monitoring statements were considered:

1. The participant notices a mistake that they had made during the trial (for example, "I made a mistake as well in VO2 and I readjusted it smaller").
2. The participant questions their actions during the trial (for example, "what did I do there?").
3. The participant mentions reversing Dplayer to go back to check a control action (for example, "I don't remember what I was thinking ... I'm going to go back again").

A list of key words suggesting monitoring was compiled. These monitoring words include: *mistake, error, wrong, accident, forgot, goofed up, not right, don't know, go back, check,* and any questions. Table 5 summarizes the number of these key words contained in the replay verbal protocols for the good and poor participants in the SE group.

Table 5. Number of Monitoring Statements

	Good Participants			Poor Participants		
Trial	A	B	C	X	Y	Z
1	2	1	4	2	2	0
2	-	1	4	0	1	6
3	1	3	2	0	0	0
4	1	0	2	2	0	0
5	1	0	5	5	0	0
6	4	3	3	3	5	0
Total	9	8	20	12	8	6

The good participants used slightly more monitoring key words than the poor participants (total of 37 versus 26 words, or an average of 2.2 and 1.6 words per trial respectively).

DISCUSSION

Chi et al. (1989) found that there were substantial differences in the verbal protocols of good and poor students, where the groupings were determined by a post-hoc median split. Analyses similar to Chi et al.'s (1989) were conducted in this study. All measures revealed differences in the expected direction, as summarised in Table 6. As far as we know, this is the first replication of Chi et al.'s results in the domain of process control.

Table 1. Summary of Spontaneous Self-Explanation Results

Measure	Rank of Group	
	Good	Poor
Time	1	2
Words	1	2
Explanations	1	2
Monitoring	1	2
Total	4	8

The good participants spent almost twice as long reviewing their trials in Dplayer as the poor participants. Self-explanation studies across domains ranging from biology, to mathematics, to computer programming have all found that good students spend more time working on a self-explanation task than poor students (Chi et al., 1994; Nathan et al., 1994; Pirolli & Recker, 1994). These studies have also found that good students speak more in their protocols than poor students. The current study found a similar trend; the good participants uttered over twice as many words as the poor participants.

The good participants included significantly more explanations in their Dplayer protocols. This supports the findings from previous studies (Chi et al., 1989; Ferguson-Hessler & de Jong, 1990). The more explanations participants produce, the more they are reflecting deeply on the reasons behind their actions and how those reasons might be revised in later trials to improve performance. When a poor participant avoids explaining the reasons behind their control actions, they have little cause to change their actions in the future to improve their performance.

Chi et al. (1989) found that good students tended to monitor their understanding of a problem and detect gaps in their understanding more often than poor participants did. The good participants in this study also included more monitoring statements in their Dplayer protocols than the poor participants, but only barely. The difference between the groups may have been minimised because the poor participants made more mistakes that merited comment. Also, mistakes may be more visible in the dynamic, interactive DURESS II environment, particularly with the P+F interface, as compared to in a non-interactive, text-based learning environment. Every control action that a participant makes has an impact on the state of the system, so the effects of an action cannot be as easily ignored as when studying text. This may explain why the difference between good and poor SE participants was not as noticeable as in previous research.

However, overall, asking the participants to self-explain did not improve their performance on normal trials. This differs from the findings of Chi et al. (1994). A separate analysis of the fault trials may show greater evidence of the influence of self-explanation. Chi et al. (1994) found that their self-explanation group demonstrated the greatest improvement relative to the control group on the most difficult questions -- those that required deeper domain knowledge. These more difficult questions would seem to be analogous to the fault trials in DURESS II.

Further, it may be necessary to introduce some form of explicit training to maximize the influence of self-explanation. This might take the form of theoretical training (Hunter et al., 1995) or an introduction to strategies and heuristics (Denning, 1995).

ACKNOWLEDGMENTS

This research was sponsored by a research contract with the Japan Atomic Energy Research Institute (Dr. Fumiya Tanabe, contract monitor), and by grants from the Natural Science and Engineering Research Council of Canada. We would like to thank Lisa Orchanian, Michael Janzen, and Darryl Minard for their contributions to this research. We also thank all of the participants for their patience and help.

REFERENCES

Brunswik, E. (1956). *The Representative Design of Psychological Experiments* (2nd edn.). Berkeley, CA: University of California Press.

Chi, M. T. H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6,* 271-315.

Chi, M. T. H., Bassok, M., Lewis, M. W., Reimann, P., & Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive Science, 13,* 145-182.

Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18,* 439-477.

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

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

Christoffersen, K., Hunter, C. N., & Vicente, K. J. (1996b). *A longitudinal study of the effects of ecological interface* *design on deep knowledge.* Submitted for publication.

Denning, R. (1995). *A case study in the design and evaluation of an interactive learning environment to teach problem-solving skills.* Unpublished manuscript, Center for Cognitive Science, The Ohio State University.

Ferguson-Hessler, M. G. M., & de Jong, T. (1990). Studying physics texts: Differences in study processes between good and poor performers. *Cognition and Instruction, 7*, 41-54.

Howie, D. E., Janzen, M. E., & Vicente, K. J. (1996). *Research on Factors Influencing Human Cognitive Behavior (III)* (CEL 96-06). Toronto, Ontario, Canada: University of Toronto, Cognitive Engineering Laboratory.

Hunter, C. N., Janzen, M. E., & Vicente, K. J. (1995). *Research on Factors Influencing Human Cognitive Behavior (II)* (CEL 95-08). Toronto, Ontario, Canada: University of Toronto, Cognitive Engineering Laboratory.

Lawson, M. J., & Chinnappan, M. (1994). Generative activity during problem solving: Comparison of the performance of high-achieving and low-achieving high school students. *Cognition and Instruction, 12*, 61-93.

Nathan, M. J., Mertz, K., & Ryan, R. (1994). *Learning through self-explanation of mathematics examples: Effects of cognitive load.* Paper presented at the 1994 Annual Meeting of the American Educational Research Association.

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

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

Pirolli, P. & Recker, M. (1994). Learning strategies and transfer in the domain of programming. *Cognition and Instruction, 12*, 235-275.

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

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

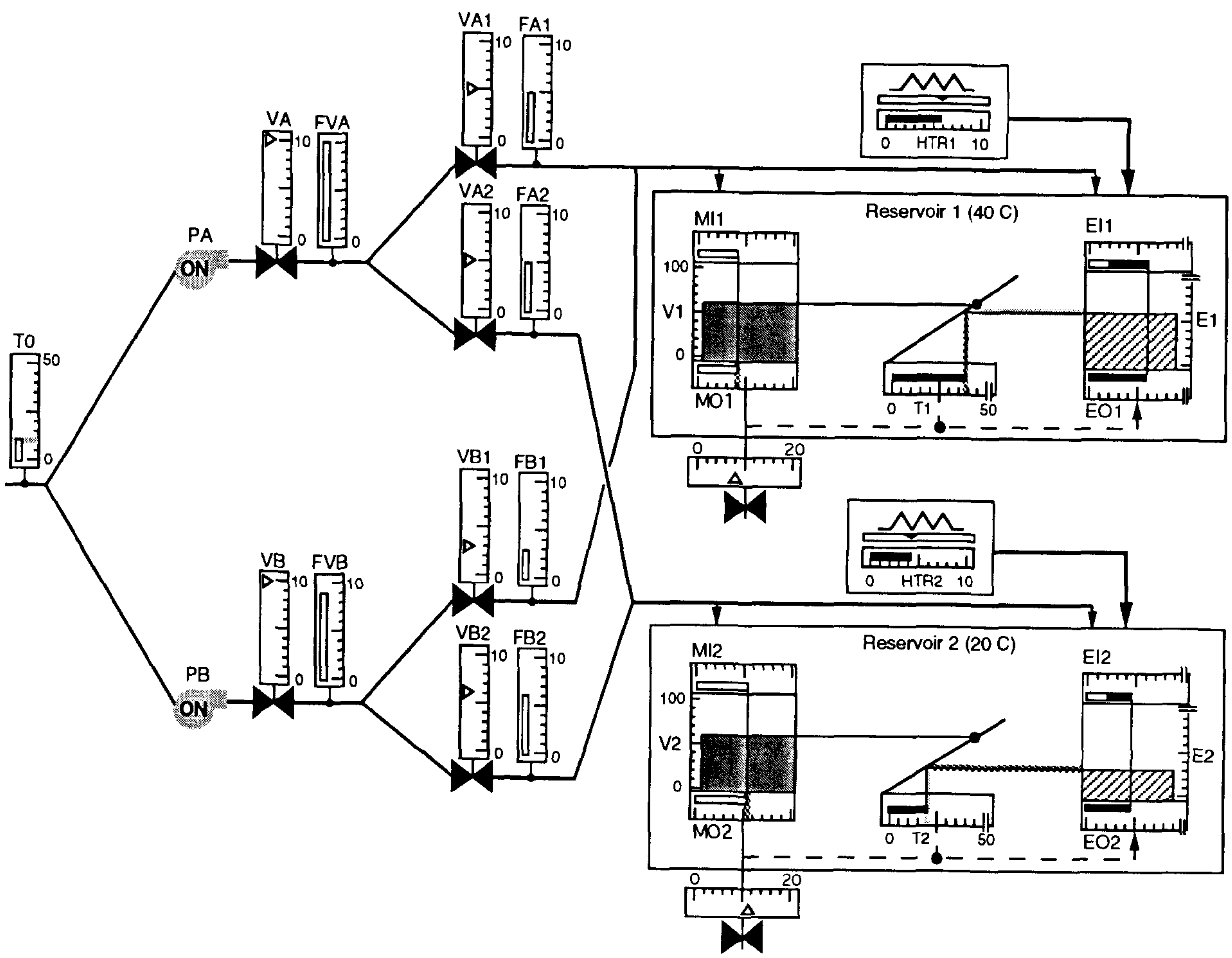

Figure 1. P+F Interface (adapted from Pawlak & Vicente, 1996).