# THE DEVELOPMENT OF EXPERTISE ON AN INTELLIGENT TUTORING SYSTEM

Final Report

NASA/ASEE Summer Faculty Fellowship Program--1989

Johnson Space Center


| | |
|---|---|
| Prepared By: | Debra Steele Johnson, Ph.D. |
| Academic Rank: | Assistant Professor |
| University & Department: | University of Houston<br>Department of Psychology<br>4800 Calhoun<br>Houston, TX 77204 |
| NASA/JSC | |
| Directorate: | Mission Support |
| Division: | Mission Planning and Analysis |
| Branch: | Technology Development and Applications |
| Section: | Artificial Intelligence |
| JSC Colleague: | Robert T. Savely |
| Date: | August 14, 1989 |
| Contract Number: | NGT 44-001-800 |

ABSTRACT

An initial examination was conducted of an Intelligent Tutoring System (ITS) developed for use in industry. The ITS, developed by NASA, simulated a satellite deployment task. More specifically, the PD (Payload Assist Module Deployment)/ICAT (Intelligent Computer Aided Training) System simulated a nominal Payload Assist Module (PAM) deployment. The development of expertise on this task was examined using three Flight Dynamics Officer (FDO) candidates who had no previous experience with this task. The results indicated that performance improved rapidly until Trial 5, followed by more gradual improvements through Trial 12. The performance dimensions measured included performance speed, actions completed, errors, help required, and display fields checked. Suggestions for further refining the software and for deciding when to expose trainees to more difficult task scenarios are discussed. Further, the results provide an initial demonstration of the effectiveness of the PD/ICAT system in training the nominal PAM deployment task and indicate the potential benefits of using ITS's for training other FDO tasks.

## INTRODUCTION

Intelligent Tutoring Systems (ITS's) have been developed for a variety of tasks, ranging from geometry to LISP programming.[1] However, many of these systems have been used primarily for research purposes and have not been widely used in academic or industrial settings. An examination is needed of an ITS developed for use in an industrial setting. More specifically, an examination is needed of the development of expertise on an ITS in an industrial setting.

## BACKGROUND ON PD/ICAT

Recently, an ITS was developed at NASA simulating the deployment of a specific type of satellite. Researchers at NASA developed the PD (Payload Assist Module Deployment)/ICAT (Intelligent Computer Aided Training) system.[2,3] The task selected for this ITS was unique in that it required highly specialized skills and required extensive training using traditional OJT (On the Job Training) methods. The population (i.e., FDO's) performing this task were also unique in that they tended to be well-educated and highly motivated. The PAM deployment task is one of many tasks (e.g., Ascent, Entry, Perigee Adjust, Rendezvous, IUS Deployments) performed by Flight Dynamics Officers (FDO's) working in the Mission Control Room. The training period for certifying a FDO ranges from two to four years. Due to the high costs and time required for training, researchers at NASA were charged with investigating tools to more quickly and economically train FDO's. The PAM deployment task was selected for ITS development in part because it was of moderate difficulty compared to other FDO tasks. In addition, PAM deployments were very common at that time, so training on this task was likely to be immediately useful to a FDO (although the frequency of PAM deployments has declined more recently). Moreover, the PAM deployment task had components common to several other FDO tasks, so training on this task was expected to transfer in part to performance on other FDO tasks.

The PD/ICAT system included a domain expert (i.e., an expert model), a trainee model, a training session manager, a scenario generator, and a user interface.[2] The domain expert contained information on how to perform the task. The task was described by a sequence of required and optional actions. However, it was necessary to build some flexibility into the sequence because several alternative sequences were equally acceptable for subsets of the actions. The knowledge type could be described as "flat procedural", that is, as requiring procedural knowledge without requiring subgoaling.[4] Because the PAM deployment task was a highly procedural task, the domain expert was constructed as a set of procedures. To model the trainee, the system used an overlay model and a bug library.[4] The system assumed the trainee model was similar to the expert model, but with some procedures missing. Further, the trainee model enabled the identification of incorrect procedures through the bug library. It is important to note that although the expert and trainee models were built as a set of procedures, extensive declarative knowledge was required to understand and perform those procedures. The training session manager interpreted the student's actions and reported the results in system (statement of action taken) messages or provided coaching in tutor (error, hint, or help) messages. Moreover, as recommended by other researchers,[5,6] the training session manager provided feedback at each step in the action sequence and provided different levels of help or hints depending on the frequency of specific errors. Information from the training session manager was also

incorporated into the student's performance record. Thus, the trainee model and training session manager together performed the major functions of student modelling: updating the level of student performance, providing information to the tutor, and recording student performance.[7] The training scenario generator was used to expose the student to scenarios of varying difficulty. Lastly, the user interface enabled the student to interact with the system to obtain, enter, and/or manipulate information and complete actions.

## DEVELOPMENT OF EXPERTISE ON THE PD/ICAT SYSTEM

The experts identified a total of 57 actions (38 required; 19 optional) to perform the PAM deployment task. These actions were performed in sequence although some subsets of actions could be performed in varying orders. In addition, the experts identified 83 display fields to check on 8 different displays. Some actions were performed more than once (e.g., anchoring an ephemeris); similarly, some of the displays were viewed more than once (e.g., the Checkout Monitor display). Performance improvement was defined in terms of increasing performance speed, completing task actions in sequence, requiring less help, and checking displays fields identified as important by the experts. These performance dimensions provided a means for examining the development of expertise on the task. Other researchers[8,9,10] have similarly described the development of skill or expertise in terms of increasing performance speed and decreasing errors. More specifically, the declarative phase of skill acquisition involves acquiring knowledge about the task. Performance at this phase tends to be slow and error-prone. The knowledge compilation phase of skill acquisition involves using declarative knowledge to build procedures for performing the task. In this phase, performance speed increases and errors are reduced as productions are built and refined.

The purpose of the current project was to map the development of expertise on the PD/ICAT task. The data collected would provide an initial examination of how efficiently novices learned from the PD/ICAT system and enable recommendations for further refinements to the software. To accomplish this, the novices' performance on various dimensions was mapped across task trials and patterns of performance examined.

## METHOD

### Subjects and Procedure

Three novices performed 12 task trials on the PD/ICAT. The novices were Flight Dynamics Officer (FDO) candidates. None had previous experience with Payload Assist Module (PAM) deployments. Experience with other integrated simulation tasks ranged from a minimum of 12 hours of observing IUS (Inertial Upper Stage) Deployments to a maximum of 48 hours of observing IUS Deployments plus more than 60 hours observing and participating in other integrated simulations (e.g., Deorbit Preparation, Entry, Ascent, Perigee Adjust, Rendezvous).

Each novice agreed to work 15-20 hours on the task in approximately 3-hour blocks spaced over a few weeks. However, due to work and other constraints, each novice had a different schedule of work sessions. Also, novices performed multiple task trials in a single work session after the initial task trials (i.e., after 3 to 5 trials, depending on the novice).

Novices were asked to read the section on PAM deployments in the Spin-Stabilized Deployment section of the Procedures Manual prior to coming to their first session. At the first session novices were shown an example of the screen display and told how to use the keyboard and the mouse to enter and manipulate task information. They were asked to "think out loud" as they performed the first task trial, that is, to describe what they were doing. In addition, the novices were invited to give their comments about the task interface and to ask questions as they performed the task. Their description of their actions, comments, and questions were tape recorded. All comments on the interface and questions about the task were noted by the researcher. However, only questions about the mechanics of the task were answered. No information was provided about which actions to perform at various points in the task. The novices were also told that their comments about the interface would be discussed with the task experts and the PD/ICAT programmers. Following each session, novices were shown a computer-generated feedback report describing their performance, their comments and questions were noted, and the next work session was scheduled. They were asked to "think out loud" again for Trials 3 and 9 (Trial 8 for one subject who was available for only 11 trials). On all other trials, the novices performed the task without having their comments tape recorded. Their comments and questions were noted by the researcher, usually at the end of the task trial.

The 12 task trials were completed in 5-6 work sessions. Following the last work session, the novices were asked to complete two short, paper and pencil tests. First, novices were asked to sort a list of all task actions into the proper sequence as quickly and accurately as possible. Second, novices were asked to identify information fields on screen displays as quickly and accurately as possible. Printed copies of each screen display were provided on which novices circled or checkmarked information fields they thought they were supposed to check during the PAM deployment task. Two of the novices completed these tests 7 days after and one novice 12 days after their last work session. Finally, novices were debriefed and thanked for their participation.

## Measures

Performance measures were collected by the computer during task performance. The performance measures collected for each trial were: trial time, number of actions completed, number of errors, number of help requests, and number of display fields checked. Trial time referred to the time required (in minutes) to complete a task trial. Number of actions completed referred to the number of actions (with or without errors) completed by the novice rather than by the Training Session Manager. (The PD/ICAT system was structured such that when the novices made three consecutive errors while attempting to complete an action, the Training Session Manager used the domain expert to complete the action.) Number of errors was the sum of three types of errors: the number of actions performed in an incorrect sequence, typographical errors (i.e., inputs the computer was unable to interpret), and optional (but recommended) actions which were not performed by the novice. Number of help requests was the sum of two types of help requests: the number of times novices requested more information from a tutor message following an error and the number of requests for explanations of the current or last step of the task.

Finally, number of display fields checked was the sum of the checks made on 8 unique screen displays, some viewed multiple times (see Table 1). The maximum score was 83 display checks. Data was not available for one other display (Detailed Maneuver Table 1) because the computer did not correctly record the number of

display fields checked. Viewing any display was an optional (but recommended) action. (The PD/ICAT system was structured such that configuring and viewing each display constituted two separate actions. A display could be configured without being viewed.) The recommended sequence and frequency of viewing different displays was determined by experts and incorporated into the PD/ICAT software. The Vector Comparison Display, however, was the only display not viewed as often as recommended by the experts. Rather than penalize the novices for failing to check displays fields on a display they failed to view, an average score was calculated. The score for the Vector Comparison Display was calculated as the average number of display fields checked each time the display was viewed (e.g., the score was 5 if the novice viewed the display twice and checked 4 and 6 fields on the first and second viewings, respectively).

**TABLE 1.- DESCRIPTION OF SCREEN DISPLAYS AND DISPLAY CHECKS.**

| Display | # of Times Viewed | # of Display Fields to Check |
|---|---|---|
| Vector Comparison* | 3 | 7, 6, 6 |
| Trajectory Digitals | 1 | 2 |
| Checkout Monitor | 4 | 9, 9, 9, 9 |
| Trajectory Profile Status** | 2 | 7, 7 |
| Detailed Maneuver Table 2 | 1 | 7 |
| Weight Gain/Loss Table | 1 | 3 |
| Supersighter | 1 | 9 |
| FDO Deploy Comp | 1 | 12 |

*An average score was used calculated from the 3 viewing opportunities.
**Only the score for the 2nd viewing opportunity was used. Data was not correctly recorded by the computer for the 1st viewing opportunity.

Additional performance measures were collected using the paper and pencil tests administered after the task trials. Three performance measures were collected on the sorting task. Sorting time referred to the time (in minutes) required to sort the sequence of actions. Unacceptable reversals referred to the number of actions sorted in incorrect sequences. Acceptable reversals referred to the number of actions sorted in a sequence regarded by the experts as an acceptable alternate sequence of actions. Two performance measures were collected from the display checking task. Checking time referred to the time (in minutes) required to check display fields on the 8 displays listed in Table 1. Number of display checks recalled was the sum of the fields checked on these 8 displays.

## RESULTS

To examine how efficiently the novices learned the PD/ICAT task, their data was plotted for each performance measure. As discussed below the data indicated rapid performance improvements until Trial 5 and more gradual further improvements

through Trial 12. A logarithmic function was used to describe the data in each measure.

As shown in Figure 1, the trial time required to perform the task decreased rapidly until Trial 5. Further performance speed improvements were more gradual. In Trial 1 only one novice completed the task and required 195 minutes. The mean trial time was approximately 46 minutes by Trial 5 and decreased to approximately 26 minutes by Trial 12. The data was described by a logarithmic function ($Y = 233.95 * X^{-.83}$, $R^2 = .87$). Interestingly, the novices who were unable to complete the task in the initial task trials demonstrated a performance pattern similar to that shown by Novice 1. Two novices failed to complete the task during the first 3-hour session, and 1 novice failed to complete the task until the third session. However, these novices demonstrated trial times similar to Novice 1 by Trial 5. Finally, the data indicates that the instruction to "think out loud" while performing the task slows performance speed. The time required to perform the task increased in Trial 3 for Novice 1, in Trial 8 for Novice 2, and in Trial 9 for Novices 1 and 3.

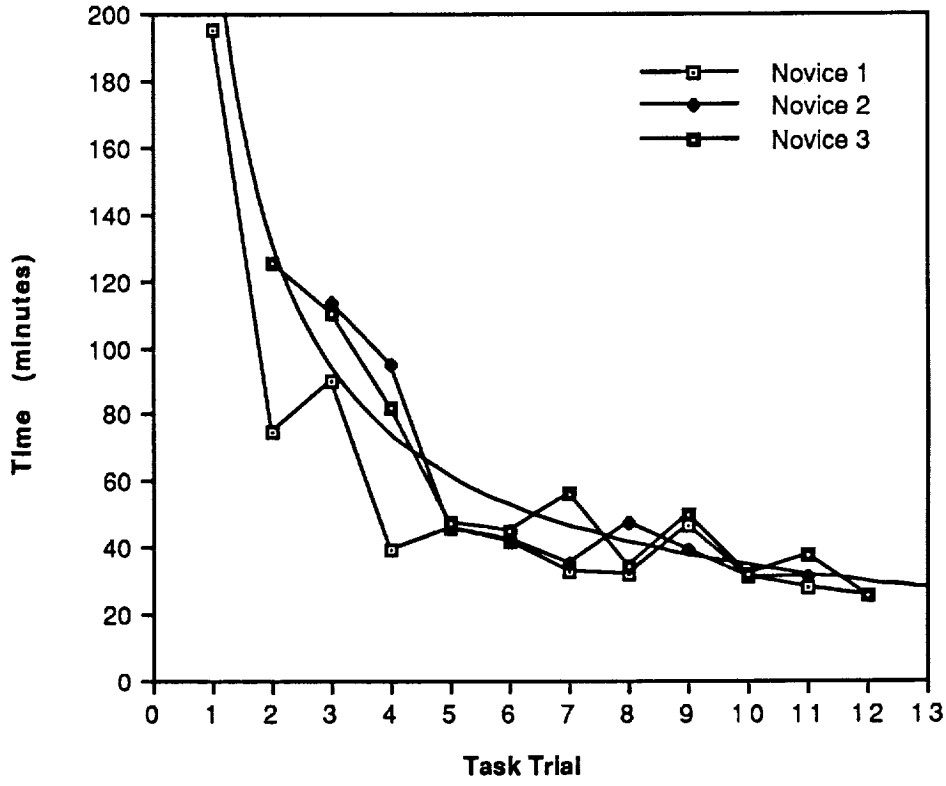


**Figure 1.- Performance speed in Trials 1 through 12.**

Number of actions completed also demonstrated rapid performance improvements until Trial 5 and then gradual further improvements. A logarithmic function (Y =

34.49 * $X^{.22}$) accounted for 63% of the variance (see Figure 2). In Trial 1, Novice 1 completed 43 actions out of the 57 possible actions. The remaining 14 actions were completed by the Training Session Manager, using the domain expert. Novices 2 and 3 completed only 28 and 26 actions, respectively. An additional 5 actions were completed by the Training Session Manager. Thus, Novice 1 completed 75% of the actions he attempted and Novices 2 and 3 completed 85% and 84% of the actions they attempted. However, one should note that Novices 2 and 3 completed or attempted to complete only 60% of the possible actions during Trial 1 while Novice 1 completed or attempted to complete all possible actions. The novices completed a mean of 52.33 actions in Trial 5 and a mean of 53.67 actions in Trial 12. Further, the novices completed at least 96% of the actions they attempted in Trial 5 and at least 98% in Trial 12. None of the novices attempted to complete more than 55 actions. Thus, novices chose not to perform at least 2 of the optional actions in every trial.

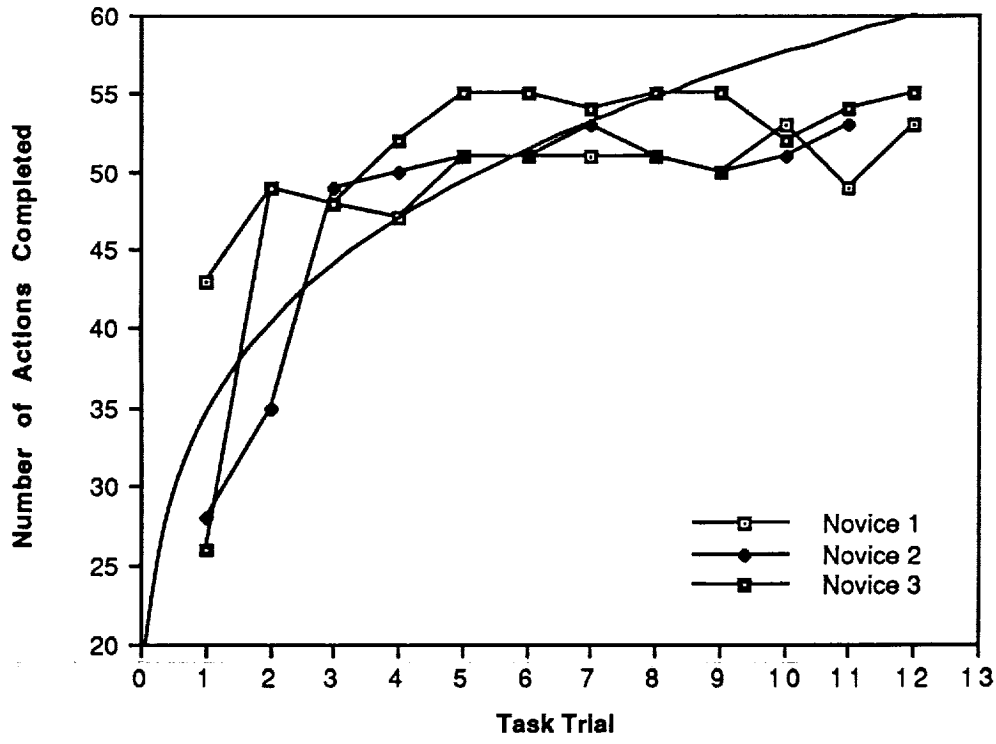


**Figure 2.- Number of actions completed in Trials 1 through 12.**

Number of errors demonstrated a similar pattern of performance. A logarithmic function ($Y = 40.46 * X^{-1.00}$) accounted for 69% of the variance (see Figure 3). In Trial 1, the novices made a mean of 24 errors. Novice 1, however, made .54 errors/action attempted while Novices 2 and 3 made .58 and .87 errors/action attempted, respectively. By Trial 5, the novices made a mean of 4.33 errors and

further reduced their errors to a mean of 3.5 by Trial 12. Thus, by Trial 5 the novices made a mean of only .08 errors/action attempted. By Trial 12, further performance improvements resulted in a mean of only .06 errors/action attempted.

Similar to other performance measures, number of help requests demonstrated rapid reductions from Trial 1 to Trial 5, but there were few help requests following Trial 5. A logarithmic function ($Y = 46.44 * X^{-1.58}$) accounted for 91% of the variance (see Figure 4). (Note: The data in Figure 4 reflect a transformation of [X + 1] to enable a logarithmic function to be fit. Data reported in the text are in their original, untransformed units.) However, the novices showed much greater variability in their help requests than in other performance measures, especially in Trials 1 and 2. In Trial 1, the number of help requests ranged from 7 to 32 requests. The number of help requests varied even more in Trial 2, ranging from 1 to 49 requests. By Trial 3, however, the novices made similar numbers of requests with a mean of 7.33 requests. In Trial 5, the novices made a mean of .67 help requests and only one help request was made from Trial 8 through 12.

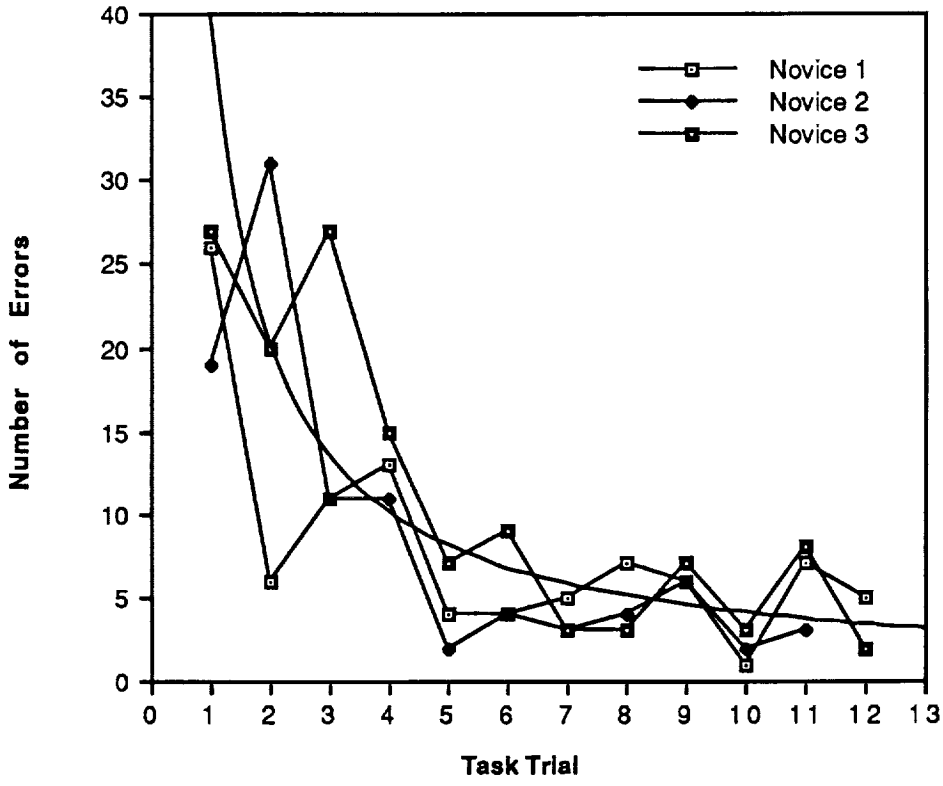


**Figure 3.- Number of errors made in Trials 1 through 12.**

Finally, number of display fields checked demonstrated rapid performance improvements from Trial 1 to Trial 5 and more gradual improvements through Trial

12. A logarithmic function ($Y = 25.94 * X^{.50}$) accounted for 70% of the variance (see Figure 5). In Trial 1, the novices checked a mean of 10.33 display fields. However, only Novice 1 had the opportunity to check all 83 display fields because the other two novices did not complete the task in Trial 1. Thus, Novice 1 checked 13% of the appropriate display fields. Novice 2 checked 12% of the 34 display fields he viewed, and Novice 3 checked 59% of the 27 display fields he viewed. Although Novice 3 checked a higher percentage of display fields than the other novices, it is not clear that he understood which fields should be checked. He may have checked numerous fields because he was unsure which were important. The task software did not record checks of any display fields other than those identified as important by the experts. Thus, following Trial 1, the novices were instructed to check only those fields they considered important in each display. In Trial 5, the novices checked a mean of 69.22 fields which was 80% of the identified display fields. By Trial 12, the novices checked a mean of 79.89 fields, checking 96% of the identified fields.

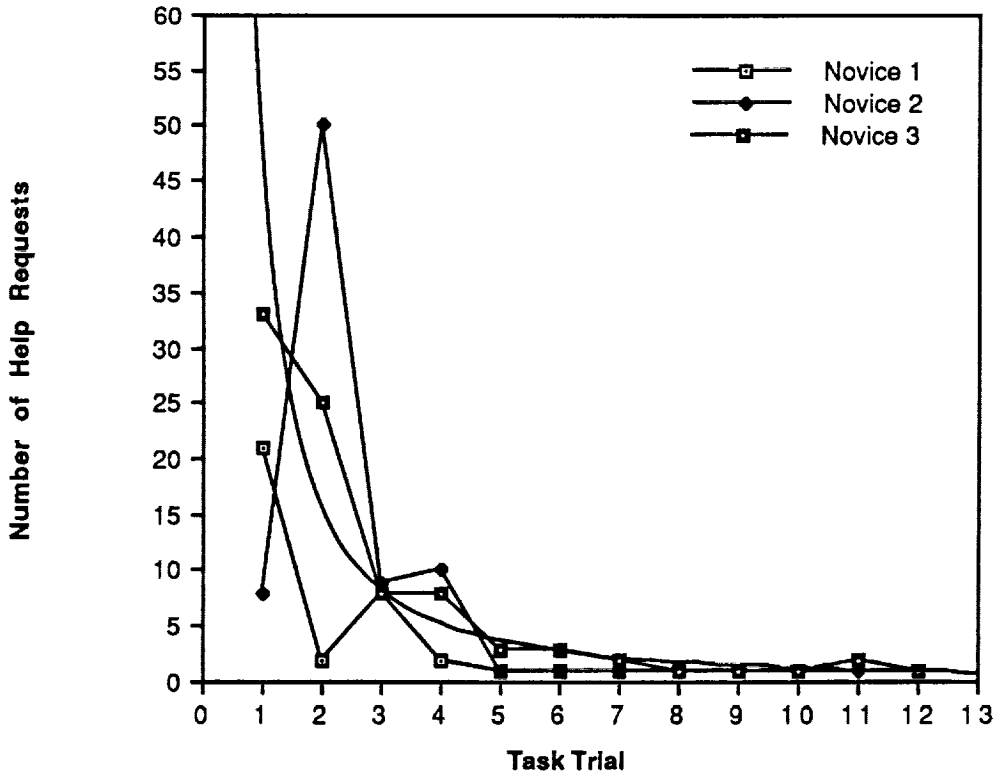


**Figure 4.- Number of help requests in Trials 1 through 12.**

The results from the two paper and pencil tests were examined to determine whether the novices knew the correct sequence of actions in the task and whether they knew which display fields were important to check, as identified by the experts. The

results of the sorting test indicated that Novices 2 and 3 required 17.08 and 17.92 minutes, respectively, to sort the task actions into the correct sequence. These two novices made 5 and 2 reversals, respectively, in how they sequenced the actions, but both reversals reflect alternate sequences regarded as acceptable by the experts. Novice 1 required 29.25 minutes to sort the task actions and made 4 acceptable and 3 unacceptable reversals. Of the 3 unacceptable reversals, one action was placed too soon, a second action too late, and the third action omitted from the task sequence. Thus, Novices 2 and 3 were able to correctly sort the task actions even after a 7-day delay. However, Novice 1 made 3 errors in sorting the task actions after a 12-day delay.

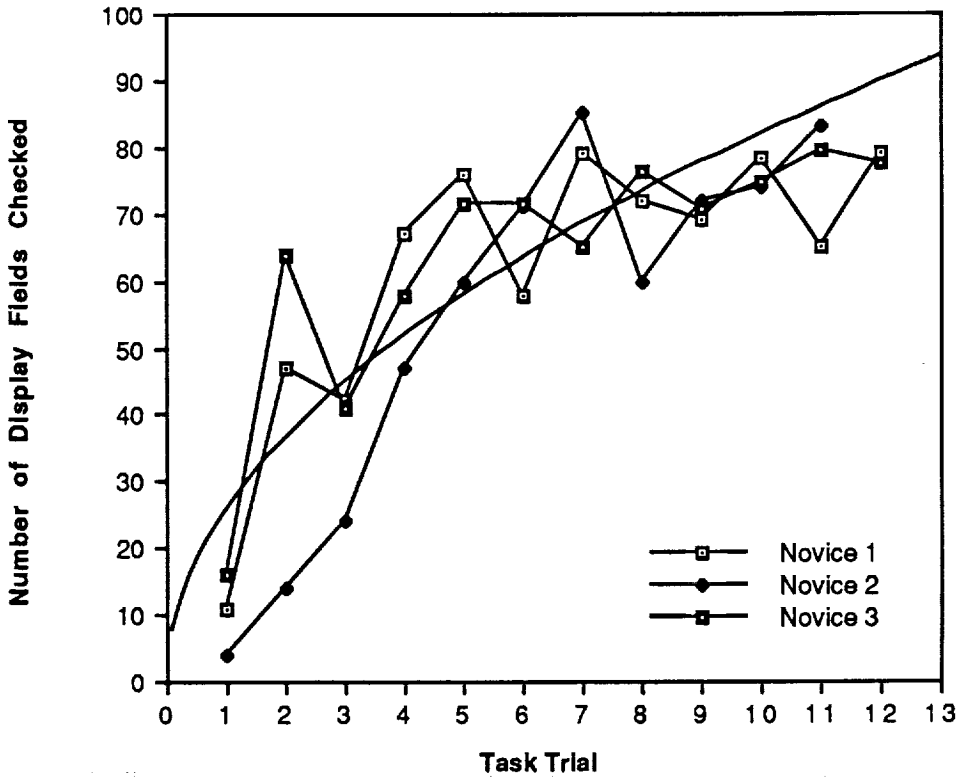


**Figure 5.- Number of display fields checked in Trials 1 through 12.**

The results of the display checking task indicated that the novices required between 3.18 and 6.58 minutes to complete the task. They checked between between 62 and 68 display fields, with a mean of 64.33. Of the total fields checked, between 40 and 51 (with a mean of 44.67) of the display fields were those identified as important by the experts. Thus, the novices checked 77% of the 58 identified display fields. However, the novices also checked between 17 and 22 (with a mean of 19.67) display

fields not identified as important. This indicated that 31% of the fields the novices checked were not identified as important by the experts.

## DISCUSSION

The results indicated that performance improved most rapidly from Trial 1 to Trial 5 on the PD/ICAT task. Additional task trials showed smaller, more gradual improvements. This suggests that the novices had developed effective procedures for performing the task by Trial 5. Additional task trials enabled the novices to refine these procedures, increasing performance speed and decreasing errors. If the goal is to train the novices to perform this specific task version as efficiently as possible, additional practice in Trials 6 through 12 may be warranted. However, the novices performed only the nominal PAM deployment task on the PD/ICAT. They also need to learn how to deal with problems that can occur during a PAM deployment, e.g., an OMS (Orbital Maneuvering Subsystem) propellant leak. So, given the smaller improvements following Trial 5, it may be reasonable after Trial 5 to expose the novices to more problematic PAM deployment scenarios.

Prior to making this decision, though, criteria should be identified for each performance dimension. That is, one needs to identify acceptable levels of performance in terms of time (in minutes) required to complete a task trial, number of completed actions (both required and optional), number of errors made, number of help requests, and number of display fields checked. These criteria, rather than a trial number, could then be used to determine when to expose a novice to a more difficult task scenario.

The results of the two tests administered after task performance indicated that the novices were able to recall the appropriate sequence of task actions a week after performing the last task trial, although there may be some decrements in recall for delays of more than a week. Similarly, the novices recalled 77% of the display fields to check after a week delay. However, decisions also need to be made here regarding 1) how many display fields should be recalled and 2) the potential benefits or costs of checking display fields not identified as important by the experts. In the nominal PAM deployment task the novices performed, no costs were associated with checking fields other than those identified. One needs to determine under what conditions it is acceptable and perhaps even desirable to check additional display fields. Experts may need to rank order the importance of checking different displays.

Finally, a few comments on the task interface are needed. These comments are based on comments and problems reported to the researcher by the novices. First, the novices experienced difficulty in beginning the task during Trial 1. All three novices were unsure what the first step should be. Consequently they received multiple error messages and may have become frustrated. To alleviate this problem, it may be appropriate to provide novices with additional information prior to performing Trial 1. This information could be in the form of task instructions, an example of the task sequence performed by the computer as the novice observes, or perhaps step by step help in completing the task sequence in the first task trial.

Second, the novices reported that some displays should be accessible at any point in the task. The PD/ICAT task as currently designed allows the novice to request displays only at specific points in the task. The novices' report should be clarified with experts and modifications made to the software to either provide novices with greater access to displays or more explanation about why they should or should not need to view a display at a specific point in time.

Third, all three novices had difficulty interpreting the error messages provided. Further refinements of the PD/ICAT task should include improvements in the tutoring (i.e., error messages) provided.

Finally, more consideration needs to be given to the data collected from novices' task performance. Observing the novices performing the task indicated that they often attempted to perform actions out of sequence, primarily in the initial task trials. However, while the PD/ICAT software currently records whether an action has been completed and number of errors associated with that action, no record is made of the specific sequence in which the actions were attempted. Further refinements to the software should enable the recording of sequencing information. Similarly, the current PD/ICAT software records only checks of identified display fields. Thus, a possible task strategy for a novice would be to check every field in a display to ensure that the machine recorded s/he had checked the important fields. A future enhancement of the software should include recording all display fields checked and perhaps providing information to the novice on why the identified fields are important to check.

## CONCLUSIONS

Novices can efficiently learn to perform the PD/ICAT task which simulates a nominal PAM deployment. Additional work is needed to more clearly identify performance criteria and expand the PD/ICAT software to include more problematic PAM deployment scenarios. Finally, refinements are needed to improve the tutoring (error messages) provided and to assist the novice in performing the first task trial. The generally positive results of this project provide an initial demonstration of the effectiveness of the PD/ICAT software in teaching novices a nominal PAM deployment task and indicates the potential benefits of future refinements and expansions of the PD/ICAT software.

## REFERENCES


1. Wenger, E. (1987). Artificial Intelligence and Tutoring Systems. Los Altos, CA: Morgan Kaufmann.

2. Loftin, R. B. (1987). A General Architecture for Intelligent Training Systems. Final Report, NASA/ASEE Summer Faculty Fellowship Program, Johnson Space Center, Contract No. 44-001-800.

3. Wang, L., Baffes, P., Loftin, R. B., & Hua, G. (1989). An intelligent training system for space shuttle flight controllers. Proceedings of the 1989 Conference on Innovative Applications of Artificial Intelligence.

4. VanLehn, K. (1988). Student modelling. In M. C. Polson & J. J. Richardson (Eds.), Foundations of Intelligent Tutoring Systems. Hillsdale, NJ: Erlbaum, 55-78.

5. Burton, R. R. & Brown, J. S. (1982). An investigation of computer coaching for informal learning activities. In D. Sleeman & J. S. Brown (Eds.), Intelligent Tutoring Systems. NY: Academic Press, 79-98.

6. Reiser, B. J., Anderson, J. R., & Farrell, R. G. (1985). Dynamic student modelling in an intelligent tutor for LISP programming. Proceedings of the 9th International Joint Conference on Artificial Intelligence (Vol. 1, pp. 8-14).

7. Biegel, J. E., Interrante, L. D., Sargeant, J. M., Bagshaw, C. E., Dixon, C. M., Brooks, G. H., Sepulveda, J. A., & Lee, C. H. (1988). Input and instruction paradigms for an intelligent simulation training system. Proceedings of the 1st Florida Artificial Intelligence Research Symposium (pp. 250-253).

8. Anderson, J. R. (1985). Cognitive Psychology and its Implications (2nd Edition). NY: W. H. Freeman.

9. Chi, M. T. H., Glaser, R. , & Rees, E. Expertise in problem solving. In R. J. Sternberg (Ed.), Advances in the Psychology of Human Intelligence (Vol. 1). Hillsdale, NJ: Erlbaum, 7-76.

10. Stevens, A., Collins, A., & Goldin, S. E. (1982). Misconceptions in students' understanding. In D. Sleeman & J. S. Brown (Eds.), Intelligent Tutoring Systems. NY: Academic Press, 13-24.