Human Research Program
Human Factors and Behavioral Performance

Effects of Long-Duration Space Flight on Training Retention and Transfer

Human Factors and Behavioral Performance:
PI: Immanuel Barshi, NASA Ames Research Center
Co-I: Donna Dempsey, NASA Johnson Space Center
Co-I: Lauren Landon, Wyle Laboratories, Inc.

Exploration Medical Capability (ExMC):
Co-I: Kerry McGuire, NASA Johnson Space Center

University of Colorado Boulder:
Co-I: Alice F. Healy (ret.)

University of Northern Colorado:
Co-I: James A. Kole

Ground Phase Report 2019

Abstract

The space environment imposes on the astronaut crew significant physiological, psycho-social, and cognitive loads that can not be replicated on the ground. These loads likely impact crew performance. To date, no systematic data collection has taken place to understand the effects of such loads on crew members' ability to retain trained knowledge and skills, and to transfer such knowledge and skills to novel situations. The research described here was originally requested by HRP management to be the first such study to systematically collect data on the effects of long duration space flight on training retention and transfer. Because current theories of retention and transfer are based on results obtained in university laboratories using undergraduate students as research participants, and because crew time in space is very expensive, this study was designed to compare the performance of 4 groups of subjects: crew members in space, crew members on the ground, crew-like subjects, and university undergraduate students. Results from the ground-phase of the study reported here demonstrate that crew members' performance under cognitive load can not be predicted from the performance of university undergraduate students. It is still an open question the extent to which crew members' cognitive performance in space can be predicted from the performance of crew members on the ground.

Table of Contents

Summary of Ground Phase Study……………………………………………….………. 4

Report of Key Findings ……….……………………………………………………… 7

Future Research ……………………………………………...………………… 38

Acknowledgements ……………………………………………………………….. 40

References ……………………………………………………………………………… 41

Appendix A – Method, Results and Data Analyses of the Ground Phase Study ..…...… 50

Appendix B – Background to the Study on the Effects of Long-Duration Space Flight on Training Retention and Transfer (short name: Training Retention) ………………..….. 87

Appendix C – Brief Responses to Comments from the HRP HFBP ES ……….….…. 106

Summary of Ground Phase Study

This report describes a study involving 2 experimental tasks conducted with crew and crew-like subjects tested at NASA's Johnson Space Center, and with university undergraduate students tested at the University of Colorado, performing the same tasks under the exact same experimental paradigm.

In the first task, subjects trained in a standard data entry task, which involved typing numbers (e.g., 2147) using their right hand. At an initial test 6 months after training, subjects completed the standard task, followed by a left-hand variant (typing with their left hand) that involved the same perceptual, but different motoric, processes as the standard task. At a second test 8 months after training, subjects completed the standard task, followed by a code variant (translating letters into digits, then typing the digits with their right hand) that involved different perceptual, but the same motoric, processes as the standard task. At test, for each of the three tasks, half the trials were trained numbers (old) and half were new. Repetition priming (faster response times to old than new numbers) was found for each task. Repetition priming for the standard task reflects retention of trained numbers, for the left-hand variant reflects transfer of perceptual processes, and for the code variant reflects transfer of motoric processes. There was, thus, evidence for both specificity and generalizability of training data entry perceptual and motoric processes despite very long retention intervals.

In the second task, subjects engaged in six 124-trial sets of a continuous memory-updating paradigm. They studied name-location associations and were tested for the location most recently associated with a given name. On study trials, all responses were to be made on a right-hand map. On retrospective test trials, responses were also to be made on the right-hand map, but on prospective test trials, designated as such during study (by showing the associations in green), responses were to be made on a left-hand map. Both retrospective and prospective test trials occurred after short (2-back) and long (8-back) retention intervals. In the last session, trials occurred under conditions in which working memory was occupied with a concurrent secondary counting backwards task. Memory for the name-location associations was better with short than with long retention intervals and was better when prospective (rather than retrospective) responses were to be made, especially at the long retention interval, even with counting backwards. Thus, the intention to respond in a special way protects against forgetting associations, and this protection is not simply due to holding the information from the special trials in working memory.

Some similar results were obtained for the crew and crew-like subjects as found for university students, although the crew and crew-like subjects consistently performed at a higher level than the university students. Importantly on both tasks, the crew and crew-like subjects showed different patterns of response under cognitive load (i.e., by the extra requirements of the code task in the first task and by the secondary counting task in the second task) than the patterns shown by the university students. Thus, crew members' performance under cognitive load on the ground can not be predicted from the performance of university undergraduate students. Given the criticality of cognitive performance during space missions and given the significant cognitive loads space missions impose on the astronaut crew, this finding has important implications for HRP as many theories and research assumptions are driven by university-based research.

It is still an open question the extent to which crew members' cognitive performance during space missions can be predicted from the performance of crew members on Earth.

Key Findings of the Ground Phase Study

This portion of the report describes a study involving 2 experimental tasks conducted with crew and crew-like subjects tested at NASA's Johnson Space Center, and with university undergraduate students tested at the University of Colorado, performing the same tasks under the exact same experimental paradigm. The results obtained with university undergraduate students have been reported earlier (Healy, Kole, Schneider, & Barshi, 2019; Healy, Schneider, Buck-Gengler, Kole, & Barshi, 2019). The comparison presented here is focused on significant performance differences among the 3 subject groups. Detailed methodology and analyses, as well as presentation of all statistically reliable findings are given in Appendix A.

A Simple Perceptual-Motor Task

Although it is hoped that learners are flexible and adaptive so they can generalize the knowledge and skills they acquire to novel situations, research has shown that instead learning in many cases is highly specific to the training context, with little or no transfer of training when the test and training situations differ.

A number of theories have been proposed to account for such demonstrations of specificity. Thorndike's (1906) theory of identical elements (see also Rickard & Bourne, 1996, and Singley & Anderson, 1989) suggests that as the similarity between training and test elements increases, better performance during test will be observed. The elements in this theory are often the physical/perceptual features of a task, such as cues and responses (Carraher & Schliemann, 2002; Pan & Rickard, 2018). By Tulving and Thomson's (1973) theory of encoding specificity, information is encoded within a context, and the provision of contextual cues that were present at the time of training may facilitate recall at the time of test. The procedural reinstatement principle suggests that as the similarity between the procedures required during training and those required at test increases, better performance during test will be observed (Healy et al., 1992; Healy, Wohldmann, & Bourne, 2005; Lohse & Healy, 2012). Thus, each of these theories explicitly predicts specificity in that greater overlap between training and test, whether based on cues, responses, context, or procedures, should improve performance at test. However, collectively these theories also allow for the possibility of transfer, to the extent that there is partial overlap between training and test in cues, responses, context, or procedures.

Retention Intervals

In addition to being flexible, it is hoped that learning is durable so that individuals may retain the knowledge and skills they acquire over long periods of time, even when there is little to no use or application of the acquired knowledge and skills during the delay. For example, for long-duration space missions astronauts must train on Earth, then apply the acquired skills to tasks in space after delays as long as months or even years, and college students might be required to apply the knowledge they gained in the classroom years after graduating and securing a job. The effect of retention interval on retention has been widely documented, starting with the work of Ebbinghaus (1885/1913), who demonstrated that retention of declarative knowledge in the form of non-word trigrams declines rapidly over the first day, then asymptotes such that those items that were retrievable after the first day remained so for the next 30 days (see Murre & Dros, 2015, for a replication of Ebbinghaus' classic study). Other research has examined even longer retention intervals, from one year (Fendrich, Healy, & Bourne, 1993) to several years (8 years, Bahrick & Phelps, 1988; 3-16 years, Ellis, Semb, & Cole, 1998; 9 years, Squire, 1989), to several decades (Bahrick, Bahrick, & Wittlinger, 1975). Squire tested memory for television programs, and Bahrick and colleagues for face-name pairs and Spanish vocabulary. Even with these differences in retention intervals and materials, a forgetting function similar to Ebbinghaus' was found, with most forgetting occurring early within the retention interval and asymptotic levels of retention reached after 8 years (Squire, 1989). In general, forgetting may be characterized by a power function, with differences in the time required to reach asymptotic levels of retention due to differences in the degree of initial learning as well as in the forgetting rate (Wixted & Carpenter, 2007). However, for pragmatic reasons, most experimental research examining long-term retention of knowledge employs retention intervals that are orders of magnitude shorter (Squire, 1989).

The first task of the present study examines the long-term retention and transfer of procedural knowledge, rather than declarative knowledge as in the previously cited studies. By the procedural reinstatement principle (Healy, 2007; Healy & Bourne, 1995; Healy, Wohldmann, & Bourne, 2005; Healy et al., 2013), declarative knowledge demonstrates greater forgetting but more robust transfer, whereas procedural knowledge demonstrates greater retention but more limited transfer. Thus, strong retention is expected in the present investigation of procedural knowledge.

The Data Entry Task

In the data entry task, we focus on the specificity and generalizability of trained perceptual and motoric procedures. That is, when individuals are trained with one set of requirements on a perceptual-motor skill, can they retain what they have learned and generalize it to new requirements, or does changing either the perceptual requirements or the motoric requirements of the skill eliminate or reduce the benefits of training especially after long retention intervals? The skill we examine here involves data entry, a procedural task in which subjects typically see four-digit numbers as numerals (e.g., 2147), read them silently, and then type them into the computer without seeing their responses and without any feedback. The sequences usually remain on the screen until the subjects hit the concluding keystroke (such as the return key). This task has been used to study skill retention and transfer and has been the test bed for multiple training principles (see Healy, Kole, Wohldmann, Buck-Gengler, & Bourne, 2011, for a review).

The data entry task was used in two experiments, using university undergraduate students as participants, addressing the training, retention, and transfer of data entry perceptual and motor processes (Healy, Kole, et al. 2019). One experiment was conducted over relatively short retention intervals, and the other over very long retention intervals. In both experiments, at training, subjects were given the standard version of the data entry task, in which they were shown four-digit numbers presented as numerals in a box in the middle of the computer screen and they entered them using the number row of the keyboard with their right hand. After training, subjects were given two tests. For the first test, subjects performed the standard task along with a left-hand variant. The left-hand variant involved different motoric processes because the numbers were entered with the left hand rather than with the right hand, but the perceptual aspects of the task did not change. For the second test, subjects performed the standard task along with a code variant. The code variant involved different perceptual processes because participants saw letters in the box but entered the corresponding digits (e.g., if they saw badg, they typed 2147) with their right hand, so the motoric aspects of the task did not change (see Appendix A for details).

The four-digit numbers entered during the two tests were either the same as during training (old) or were seen for the first time during the test (new). Repetition priming (old faster than new) at test for the standard task reflects retention of the old numbers and specificity of training. Repetition priming for the left-hand task also requires perceptual transfer of the trained task (across changes in motor responses), and repetition priming for the code task also requires motoric transfer of the trained task (across changes in perceived stimuli). On the basis of the theories and results reviewed earlier explicitly documenting specificity of training, our working experimental hypothesis was that we would find repetition priming for the standard task, reflecting specificity, but either no repetition priming or (given an overlap in task parameters) at least less repetition priming for the two task variants (left-hand and code), both of which require transfer of training. Thus, the index of transfer that was used was the existence of repetition priming on the task variants, so that we are, in effect, examining an index of the transfer of specificity of learning. It should be noted that this is a novel transfer index, and that many different transfer indices have been used in past research (see, e.g., Wohldmann & Healy, 2010).

A Complex Memory Task

Memory processes underly all of learning, the application of such learning, and the performance of all cognitive tasks such as problem solving and decision making. The simple data entry task examined retention and transfer of perceptual and motoric processes; the complex memory tasks probed the fundamental cognitive process of memory. In a study by Bourne, Healy, Bonk, and Buck-Gengler (2011), it was found that responding to particular items differently than to default items protected against forgetting associations to those items. In other words, responding in a special way to items preserved associated information in memory. The second task of the present study aims to determine whether the protection against forgetting still occurs when working memory processes are otherwise occupied.

The paradigm developed by Bourne et al. (2011) allows for a continuous assessment of four types of memory in a single setting: short-term retrospective memory for items with default responses, long-term retrospective memory for items with default responses, short-term prospective memory for items with special responses, and long-term prospective memory for items with special responses. Prospective memory requires forming an intention to perform a task at a later time, whereas retrospective memory simply involves the retrieval of information encoded in memory previously. A response to an item is considered "special" when it is different from the response required on default items. Default items are encountered much more frequently than items requiring a special response.

Also, this paradigm permits separate tests of memory for previously learned associations and of memory for designated response types. Specifically, subjects were required to keep track of the color associated with a particular name as well as the response location required to demonstrate recognition of that color on a later test.

On both retrospective and prospective test trials, subjects associated colors with names. The only difference was in where the color response was to be made (default or special response location). Also, there were twice as many retrospective as prospective trials, with the same default response location for study trials and retrospective test trials, ensuring that the prospective response location was indeed special. To examine forgetting of associations, two retention intervals were compared (short and long). Forgetting of associations would be evident by a lower proportion of correct association responses after the long retention interval than after the short retention interval.

In fact, subjects were more accurate overall for color responses on prospective memory trials than on retrospective memory trials. Also, subjects were more accurate overall for color responses at short than at long retention intervals. However, importantly, the effect of retention interval was evident only for retrospective memory, not for prospective memory. That is, forgetting of name-color associations across the short and long retention intervals was much greater on retrospective test trials than on prospective test trials.

We used the Bourne et al. (2011) paradigm to study retention and transfer of memory processes, but the task we used was different. Instead of name-color associations, subjects made associations between eight astronauts named alphabetically (Alpha, Bravo, Charlie, etc.) and four cardinal-direction locations in an explored space (North, South, East, West). All study trials consisted of presenting a name-location association (e.g., Alpha North), to which subjects responded by clicking at the appropriate location in a map labeled Spacecraft on the right side of the display screen. Thus, we created a mapping task (see details in Appendix A).

Study trials and test trials were intermingled. In study trials, names and locations were presented; in test trials, names were presented whereas location and response type (default or special) had to be recalled. Most of the astronaut names shown during the study trials were in the default format (with black type and mixed font). Some astronaut names, however, were special, by being in an atypical, distinctive format (green type and all capital letters). Test trials, intermingled with study trials, required responding with the last known location for a given name and consisted of the name alone in blue surrounded by X's. On retrospective test trials the response was made in a default manner by clicking on the Spacecraft map (see Appendix A). On prospective test trials (when the name-location association had been studied in green and all capital letters) the response was made instead on the special left-side Mission Control map. These requirements can be understood if subjects think of the atypical, distinctive format as signaling importance. Subjects were instructed that, when tested, names most recently signaled as important needed to be reported to Mission Control rather than to the Spacecraft. It should be noted that on test trials the names themselves were always in blue; there was no distinctive color or capitalization for the expected special responses. To respond on the correct side on a test trial, subjects had to remember whether or not the distinctive color and capitalization occurred for a given name during the most recent time it was studied.

To examine forgetting of name-location associations, as in the study by Bourne et al. (2011), two retention intervals were compared for the time between study and test of an item. Short intervals occurred following one intervening trial (2-back), and long intervals following seven intervening trials (8-back). A concurrent counting backwards task (counting backwards by 3's from 100) was used during one transfer test block to occupy working memory, so that it would be unavailable for the primary task of associating names with locations, as a way to test transfer under cognitive load.

The experiments were designed to examine the acquisition, retention, and transfer of the skill of learning name-location associations. Multiple blocks of training trials were conducted to examine skill acquisition, and transfer test blocks were conducted to examine skill transfer. However, the results involving skill learning showed that performance on the name-location associations did not improve substantially during the training process. Hence, the training analyses reported by Healy, Schneider, et al. (2019) were averaged across blocks of training trials and were focused instead on the issues involving the forgetting of name-location associations across the short and long retention intervals during a particular block of trials as a function of retrospective and prospective memory task requirements (i.e., default side responding for retrospective and special side responding for prospective).

The Data Entry Task (Experiment 1)

Method (see further detail in appendix A and in Appendix B)

Subjects. Twenty-six undergraduate students in Aerospace Engineering at the University of Colorado, Boulder, participated for pay. In addition, 20 subjects were recruited from NASA' JSC, 11 of whom were astronaut candidates (crew) and 9 of whom were crew-like subjects.

Materials. There were two training sessions, both of which involved the standard data entry task. Both training sessions included three blocks of 100 trials, with each trial consisting of a different four-digit number. Allowable four-digit numbers were those in which no digit was repeated, and the digit 0 was not used. A block was constructed by randomly selecting, without replacement, four-digit numbers from the total set of 3,024 unique numbers meeting these criteria. The same 100 trials were used in all three blocks of both training sessions, in a different random order. Thus, each stimulus number was presented a total of six times across both training sessions.

There were two test sessions. The first test session included 100 trials of the standard data entry task followed by 100 trials of a left-hand variant of the data entry task, which involved seeing digits and typing them with the left hand. The second test session included 100 trials of the standard data entry task followed by 100 trials of a code variant, which involved seeing letters and translating them into digits, with a = 1, b = 2, etc., and typing the digits with the right hand.

For the first test session ("Onboard 1"), 50 (or half) of the 100 trials for each task (standard, left hand) were old trials; that is, they had been presented once in each block of training during each training session. The other 50 (or half) of the 100 trials for each task were new trials that had not been presented during the training sessions. The new trials were also 4-digit numbers in which no digit was repeated, and the digit 0 was not used. This method allowed for a retention test of the trained stimuli in the standard task and transfer test of the trained stimuli in the left-hand variant. More specifically, the left-hand variant tests for perceptual transfer because the motoric components of the task are different for the standard and left-hand tasks, but the perceptual requirements are the same.

For the second test session ("Onboard 2"), again 50 (or half) of the 100 trials for each task (standard, code) were old trials (numbers that were practiced during training), and the other 50 (or half) of the 100 trials for each task were new trials that had not been practiced during training and were different than those used during the first test session. The code variant tests for motoric transfer because the perceptual components of the task are different for the standard and code tasks, but the motoric requirements are the same. For the code task, the digits were replaced with corresponding letters in the alphabet; for example, if a trained number was *2154*, then the corresponding stimulus for the code task was *baed*.

Before training there was a short pretest (Baseline Data Collection; BDC) and after the tests there was an identical short posttest, both involving all task variants; these tests are not discussed here because of their abbreviated nature.

Design. The design for training and for the tests included the between-subjects variable of subject type (crew, crew-like, undergraduate), and within-subject independent variables. For training, the within-subject variables were training session (Session 1, Session 2) and block (Block 1, Block 2, Block 3). For the tests, the within-subject variables were task (standard, left hand) or (standard, code) and trial type (new, old). The dependent variables examined were accuracy as well as total response time (TRT; the time to type the four digits and the concluding *Enter* key) for correct trials (i.e., for trials in which all four digits were typed correctly).

It should be noted that the order of the two tests was not counterbalanced. Such a lack of counterbalancing would be a problem if the two tests or the two novel task variants (left-hand and code) were being compared, but such comparisons were not made and are not of interest in the present study. Also, it should be kept in mind that the order of the two tasks within each test (the standard and the novel) was also not counterbalanced (the standard always preceded the novel), and those two tasks are compared in the present analyses. Such an order confounding needs to be considered whenever this comparison is made, but other confounding differences between the two tasks also occurred and are a necessary feature of the design. Specifically, the standard task, which is treated as a baseline, is the only task used in training, occurs more often at test (i.e., in each test), and is simpler than the novel variants. These confounding differences are not a problem, however, for testing for a lack or presence of repetition priming in each novel task variant. Repetition priming in the left-hand task would indicate transfer of training perceptual processes and repetition priming in the code task would indicate transfer of training motoric processes, and such evidence for transfer could not be explained by either the lack of counterbalancing or the presence of confounding differences between the standard task and the novel task variants.

Scheduling of Sessions. There were six experimental sessions. The first session involved pretesting (BDC). The second and third sessions involved training on the standard task. The fourth session was the first test, which consisted of 100 trials of the standard task and then 100 trials of the left-hand task. The fifth session was the second test, which included 100 trials of the standard task and then 100 trials of the code task. The sixth session involved posttesting. The sessions were distributed over approximately 480 days (16 months). The schedule was meant to mimic that used by astronauts on a standard duration six-month mission to the International Space Station. Specifically, there were three months between the two training sessions (mean = 98.55 days, standard deviation = 6.00 days). The first test occurred six months after the second training session (mean = 188.63 days, standard deviation = 5.66 days). The second test occurred eight months after training (mean = 254.60 days, standard deviation = 6.38 days). The length of time between sessions in this experiment varied depending on subject availability. Each session lasted less than one hour.

Key Results and Discussion (see Appendix A for further detail, including results from the pre- and post-tests and specific pair-comparisons of crew vs. crew-like, crew vs. students, and crew-like vs. students)

Note: all statistical tests were mixed-factorial ANOVAs (used when there is 1 or more between-subjects independent variables, in this case subject type, and 1 or more within-subjects variables), using Statsview and SPSS.

Training. Averaging across training sessions and blocks, there were differences in accuracy between subject types; crew (M = .959) and crew-like subjects (M = .959) demonstrated higher accuracy than did undergraduate subjects (M = .926), $F(2, 41)$ = 3.566, MSE = .010, p = .037, η_p^2 = .148.

Average total response time (TRT) decreased from the first training session (M = 2.891 s) to the second training session (M = 2.831 s), $F(1, 41)$ = 4.895, MSE = .033, p = .033, η_p^2 = .107, as well as across blocks within each training session (Block 1 = 2.881 s; Block 2 = 2.864 s; Block 3 = 2.838 s), $F(2, 82)$ = 4.179, MSE = .010, p = .019, η_p^2 = .092. There was a marginally significant difference in TRT between subject types, with faster TRT for crew (M = 2.664 s) than for crew-like (M = 2.798 s) and undergraduate subjects (M = 2.963 s), $F(2, 41)$ = 3.048, MSE = .671, p = .058, η_p^2 = .129. A post-hoc test revealed that the difference between NASA subjects (combining crew and crew-like) and undergraduate subjects was significant, $F(1, 42)$ = 5.368, MSE = .667, p = .026, η_p^2 = .113.

Test 1 (Standard, Left Hand). For the standard task, accuracy was higher for old numbers than for new numbers for crew and crew-like subjects; the opposite pattern was observed for undergraduate subjects. For the left-hand task the opposite pattern was observed: accuracy was higher for new numbers than for old numbers for crew and crew-like subjects, but accuracy was higher for old numbers than for new numbers for undergraduate subjects, $F(2, 43) = 4.466$, $MSE = .001$, $p = .017$, $\eta_p^2 = .172$ (see Figure 1).

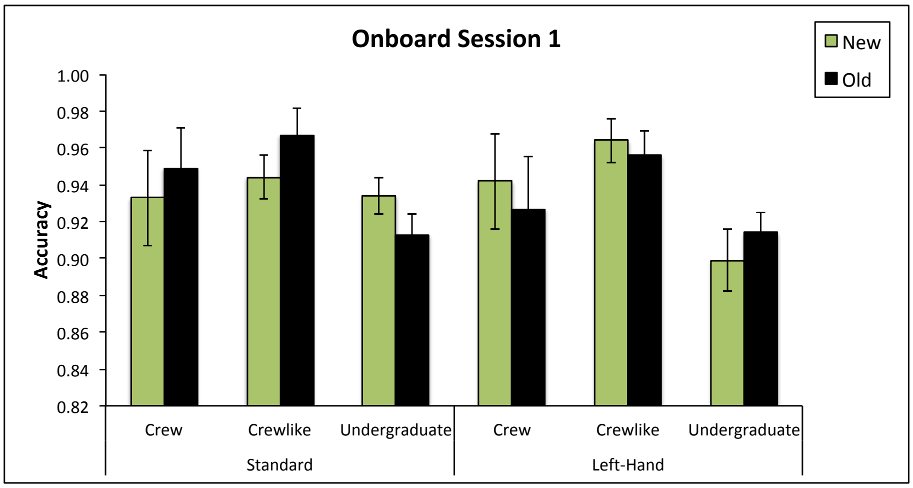


Figure 1. Test 1 Accuracy as a function of subject type and trial type. *Note:* Here and in all subsequent figures, error bars represent between-subjects standard errors of the mean.

Subjects were faster to perform the standard task (M = 2.753 s) than the left-hand task (M = 2.873 s), $F(1, 43) = 24.485$, $MSE = .025$, $p < .001$, $\eta_p^2 = .363$. There was an overall significant repetition priming effect. Averaging across task, old numbers (M = 2.787 s) were entered more quickly than new numbers (M = 2.839 s), $F(1, 43) = 25.532$, $MSE = .004$, $p < .001$, $\eta_p^2 = .373$. However, the repetition priming effect was larger for the left-hand task than for the standard task, $F(1, 43) = 3.605$, $MSE = .006$, $p = .064$, $\eta_p^2 = .077$ (see Figure 2). The repetition priming effect was marginally significant for the standard task ($p = .070$) but significant for the left-hand task ($p < .001$). (See discussion of the Code task further below.)

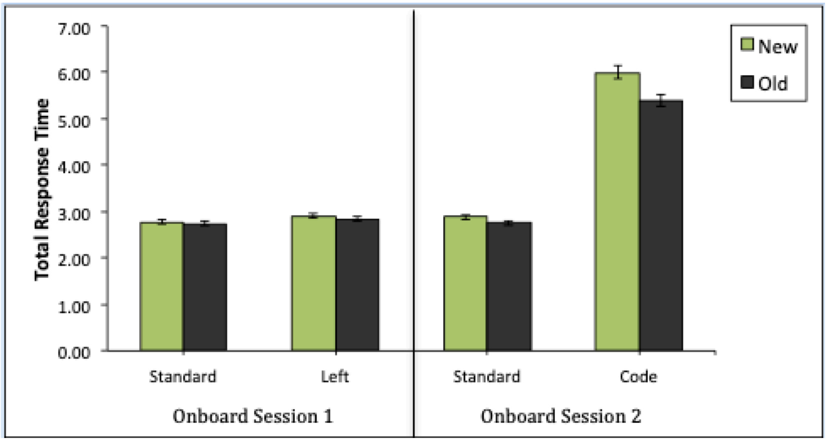


Figure 2. Test 1 TRT as a function of test and trial type.

Test 2 (Standard, Code). Overall accuracy was higher for the standard data entry task (M = .937) than for the code task (M = .829), $F(1, 43) = 41.277$, $MSE = .007$, $p < .001$, $\eta_p^2 = .490$. There were also differences in accuracy between subject types; on average, accuracy was higher for crew (M = .910) and crew-like (M = .924) subjects than for undergraduate subjects (M = .857), $F(2, 43) = 4.392$, $MSE = .019$, $p = .018$, $\eta_p^2 = .170$. However, the increase in accuracy for crew and crew-like subjects was restricted to the code task, $F(2, 43) = 4.337$, $MSE = .007$, $p = .019$, $\eta_p^2 = .168$; accuracy was similar for the three subjects types on the standard task (see Figure 3).

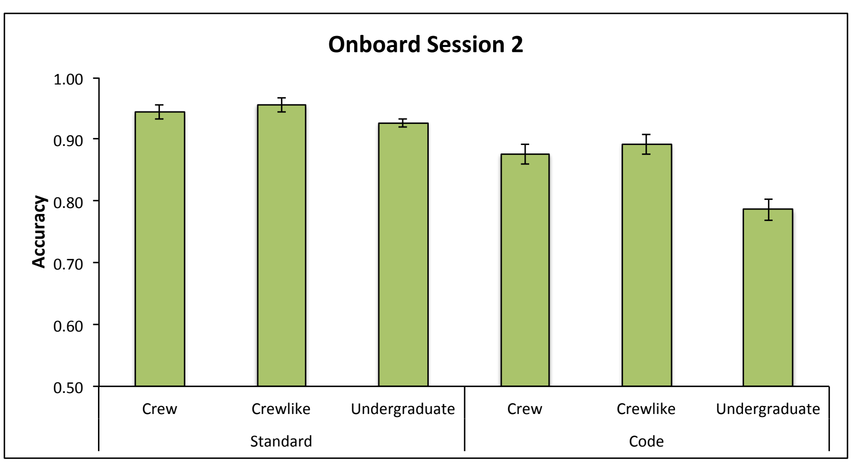


Figure 3. Test 2 Accuracy as a function of subject type and task.

Subjects were also faster to perform the standard task (M = 2.817 s) than the code task (M = 5.694 s), $F(1, 43) = 549.364$, $MSE = .542$, $p < .001$, $\eta_p^2 = .927$. There was an overall significant repetition priming effect. Averaging across tasks, old numbers (M = 4.072 s) were entered more quickly than new numbers (M = 4.439 s), $F(1, 43) = 133.616$, $MSE = .035$, $p < .001$, $\eta_p^2 = .757$. The repetition priming effect was larger for the code task than for standard task, $F(1, 43) = 43.222$, $MSE = .046$, $p < .001$, $\eta_p^2 = .501$ (see Figure 2), although it was significant for both the standard task ($p < .001$) and the code task ($p < .001$).

Discussion. As found in many earlier studies with different paradigms (e.g., Healy, Wohldmann, Parker, & Bourne, 2005; Healy et al., 2006; see Johnson & Proctor, 2017, for a recent review), there was evidence in the present experiment of a significant degree of specificity of training because the repetition priming effect found here for the standard task with the total response time measure shows that subjects were faster at responding to numbers that had occurred six times during training than they were at responding to new numbers not shown earlier. However, contrary to our initial hypothesis based on previous findings and models of specificity (e.g., procedural reinstatement; Healy et al., 1992), there was strong evidence for transfer of training in the present study because the repetition priming effect for total response time was found even when the training and testing tasks differed in the motoric aspects of the responses (in the left-hand task) and when the training and testing tasks differed in the perceptual aspects of the stimuli (in the code task). In fact, the repetition priming effects for total response time were larger for the novel task variants (left-hand and code) than for the standard task. This difference in the magnitude of the repetition priming effects might be attributed to the longer response times with the novel variants. In any event, it seems difficult to reconcile this finding with the earlier models of specificity, including the identical elements models (e.g., Thorndike, 1906) and the procedural reinstatement principle (e.g., Healy et al., 1992). The present results, thus, provide food for thought as to how to incorporate findings of generalizability into models of specificity.

In any case, there was clearly transfer of perceptual processes in the left-hand task and of motoric processes in the code task. In previous work by Healy, Schneider, and Barshi (2015), involving a navigation task and different measures of transfer, either specificity or generalizability was found, but not both for a given condition in each of their six experiments. Nevertheless, in the present study there was evidence for both specificity and generalizability of training for both perceptual and motoric processes of data entry.

An important aspect of the present findings is the very long retention intervals involved (eight months following training). It seems truly amazing that subjects could retain the 100 old four-digit sequences across such long retention intervals. It is even more impressive that subjects could demonstrate such durable memory even when they entered the sequences with a hand different from that used at training or even when they saw the sequences in a format different from that used at training.

Beyond the theoretical implications concerning models of specificity, the practical implications of the present results concerning training of people on the ground are encouraging. Despite the long retention intervals and changes in the perceptual or motoric aspects of the trained tasks, Earth-bound learners could be expected to benefit from the training they received earlier even when the learning is incidental and even when what is learned may be confusable. Furthermore, the significant transfer found for the total time measure across changes in either the perceptual or motoric aspects of the task has implications for the issue of how much fidelity to the target task is necessary for simulators and other training devices to be effective. The present results suggest that changes in perceptual or motoric aspects of the task may not detract from the usefulness of such training devices, so they may be effective without perfect fidelity to the target task. However, it is important to recognize that the data entry task is a very simple task. Thus, to truly understand retention and transfer of the kinds of tasks astronauts will have to perform on long duration space missions, ecologically valid operational tasks should also be studied.

An additional issue addressed here is whether the results obtained with college students are also found with crew and crew-like subjects. This issue is especially important given the finding that crew and crew-like subjects performed both more accurately and more rapidly than the university students overall. Although there were several significant interactions of subject type with other variables, a critical and meaningful interaction involved subject type and task (standard, code) on accuracy in Test 2. The elevated accuracy for crew and crew-like subjects was found only for the code task, not for the standard task. This interaction implies that the crew and crew-like subjects are better able than undergraduate students to cope with the increased cognitive load demanded by the code task relative to the standard task.

Of particular interest is the pattern shown in accuracy results of Test 1 (Figure 1). Undergraduate university students show an opposite pattern to that of crew and crew-like subjects. This opposite pattern is seen in both the standard task and the left-hand task.

This opposite pattern demonstrates that results obtained in studies using undergraduate students as test subjects may not predict the pattern of responses of crew members. Thus, to be able to predict the behavior of crew members on Earth, we must study crew members, or at least crew-like subjects.

The Mapping Task (Experiment 2)

Method (see Appendix A and Appendix B for further detail)

Subjects. The same subjects employed in Experiment 1 were employed in Experiment 2. At each of the six sessions, they participated in the Experiment 2 procedures immediately after the Experiment 1 procedures.

Materials and procedure. As with the data entry task of Experiment 1, the first session was a pretest, a base-line data collection (BDC) session. The second and third sessions were considered “training,” the fourth and fifth sessions as the on-board tests, with the sixth session as posttest. Each of the training sessions included two blocks of training trials, and each block consisted of 124 continuous memory-updating trials (i.e., at each trial the participant had to update the memory for the location of the given astronaut name) and used the same set of stimuli in the same order. The 124 trials in each of the four training blocks included 16 prospective test trials, with eight of each of the two retention intervals (2-back and 8-back), and 32 retrospective test trials, with 16 of each of the two retention intervals. The remaining 76 trials in each of the four training blocks were study trials.

Transfer testing occurred during the fourth and fifth sessions. There were two transfer tests (standard and counting), both involving different stimuli from those shown during training. The first transfer test, conducted during the fourth session, required no secondary task. The second transfer test, conducted during the fifth session, required the concurrent secondary counting backwards task. The change in stimuli from one test to another only involved differences in the order of the names (e.g., the initial trial was a study trial and Bravo was the name on that trial in the four training blocks, whereas Charlie was the name on the initial trial in the first test (standard), and Delta was the name on the initial trial in the second test block (counting); all of these initial study trials were shown with the location North). The pseudorandom order of the trial types (retrospective study, prospective study, retrospective test, prospective test), the pseudorandom order of the correct locations (North, South, East, West), and the pseudorandom order of the retention intervals (2-back, 8-back) were the same for each of the four training blocks and the two test blocks.

The pretest (BDC) during the first session and the identical posttest during the last session included a 37-trial sequence with no secondary task followed by the same 37-trial sequence with the concurrent counting secondary task. See Appendix A for the results of the pretest and posttest.

Design. The reported training analyses were averaged over the four training blocks (two blocks on each of two days). The design of the training analyses, thus, included the between-subjects variable of subject type (crew, crew-like, undergraduate students) and within-subject independent variables, including both trial type (retrospective, prospective) and retention interval (2-back, 8-back). Two dependent variables were examined: proportion correct location (North, South, East, West), which tests whether or not subjects remembered the location associated with the names, and proportion correct side (right, left), which tests whether or not subjects realized that it was a special trial.

The design of the transfer ("onboard") tests, which also included the between-subjects variable of subject type and within-subject independent variables, was the same except for the additional within-subject variable of test (standard, counting).

The protection against forgetting name-location associations that may result when special responses are required would be evidenced in an interaction between trial type (retrospective, prospective) and retention interval (2-back, 8-back) on the proportion of correct location responses, reflecting a smaller difference between the short (2-back) and long (8-back) retention intervals when prospective rather than retrospective memory is required. According to the working memory hypothesis (Bourne et al. 2011), the protective function of prospective memory should be eliminated, or at least reduced, when working memory is occupied with the counting backwards task (i.e., on the counting test). Thus, the working memory hypothesis leads to the prediction for the transfer tests of a three-way interaction between test, trial type, and retention interval in an analysis examining the proportion of correct location responses.

Key Results and Discussion (see Appendix A for further detail, including specific pair-comparisons of crew vs. crew-like, crew vs. students, and crew-like vs. students)

Training. For location accuracy on the training blocks, performance on prospective memory trials (M = .664) was better than on retrospective memory trials (M = .517), $F(1, 41) = 35.875$, $MSE = .082$, $p < .001$, $\eta_p^2 = .467$. Performance was also better at the short (2-back) retention interval (M = .698) than at the long (8-back) retention interval (M = .483), $F(1, 41) = 122.065$, $MSE = .059$, $p < .001$, $\eta_p^2 = .749$. However, the effect of retention interval was much larger for retrospective than for prospective trials, $F(1, 41) = 22.211$, $MSE = .024$, $p < .001$, $\eta_p^2 = .351$, thus illustrating the protective function of special responding against forgetting name-location associations across the short and long retention intervals (see Figure 4).

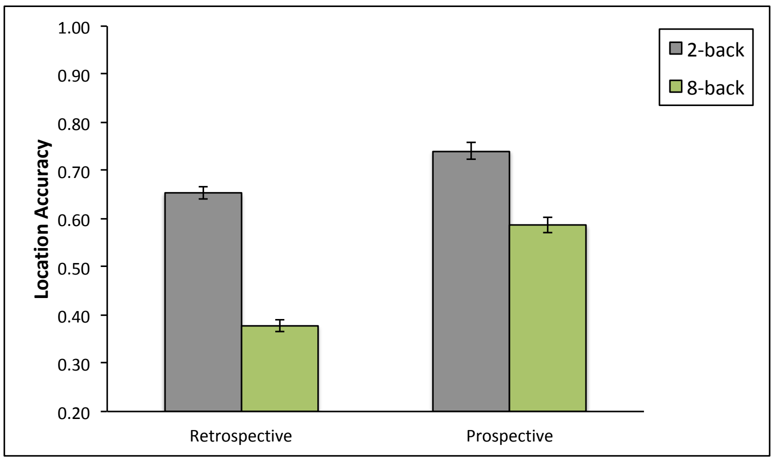


Figure 4. Location accuracy as a function of trial type and retention interval at training.

There were also subject type differences in location accuracy, with accuracy higher for crew (M = .691) and crew-like (M = .632) subjects than for undergraduate subjects (M = .535), $F(2, 41) = 8.402$, $MSE = .184$, $p < .001$, $\eta_p^2 = .291$.

For side accuracy on the training blocks, performance on retrospective memory trials (M = .882) was better than on prospective memory trials (M = .553), $F(1, 41) = 73.502$, $MSE = .146$, $p < .001$, $\eta_p^2 = .642$. Performance was also better at the short (2-back) retention interval (M = .741) than at the long (8-back) retention interval (M = .634), $F(1, 41) = 55.078$, $MSE = .035$, $p < .001$, $\eta_p^2 = .573$. The interaction between retention interval and trial type was also significant, $F(1, 41) = 28.515$, $MSE = .046$, $p < .001$, $\eta_p^2 = .410$, however, the pattern was in the opposite direction for side (better for retrospective) than for location (better for prospective) accuracy, reflecting the greater frequency of retrospective trials (see Figure 5).

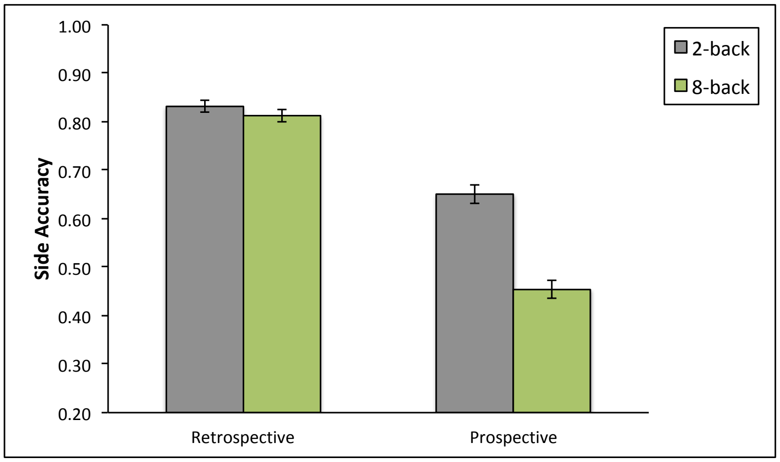


Figure 5. Side accuracy as a function of trial type and retention interval at training.

As for location accuracy, there were subject type differences with side accuracy higher for crew (M = .768) and crew-like (M = .718) subjects than for undergraduate subjects (M = .634), $F(2, 41) = 4.919$, $MSE = .192$, $p = .012$, $\eta_p^2 = .194$.

Transfer tests. The same effects of trial type, $F(1, 43) = 29.885$, $MS = .043$, $p < .001$, $\eta_p^2 = .410$, retention interval, $F(1, 43) = 68.371$, $MS = .031$, $p < .001$, $\eta_p^2 = .614$, and the interaction between the two, $F(1, 43) = 22.514$, $MS = .017$, $p < .001$, $\eta_p^2 = .344$, were significant during the transfer test sessions. Location accuracy was higher for prospective trials (M = .555) than for retrospective trials (M = .440), and for the shorter retention interval (M = .578) than for the longer retention interval (M = .416). The effect of retention interval was larger for retrospective trials than for prospective trials (see Figure 6).

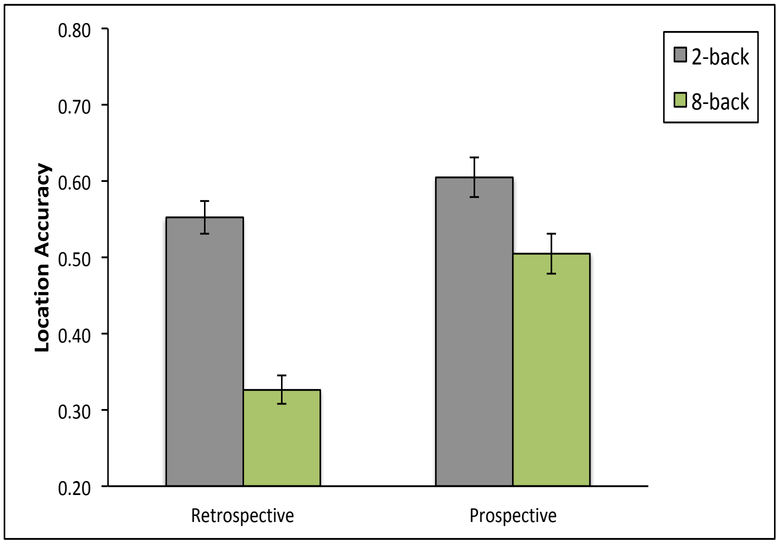


Figure 6. Location accuracy as a function of trial type and retention interval at test.

Overall, location accuracy was higher for the first transfer test, Onboard 1 (M = .608), than for the second transfer test, Onboard 2 (M = .386), during which there was a concurrent counting backwards task, $F(1, 43) = 59.747$, $MS = .052$, $p < .001$, $\eta_p^2 = .581$. However, the effect of the concurrent secondary task was stronger for short-retention interval trials than for long-retention interval trials; the interaction between test and retention interval was significant, $F(1, 43) = 7.572$, $MS = .028$, $p = .009$, $\eta_p^2 = .150$ (see Figure 7).

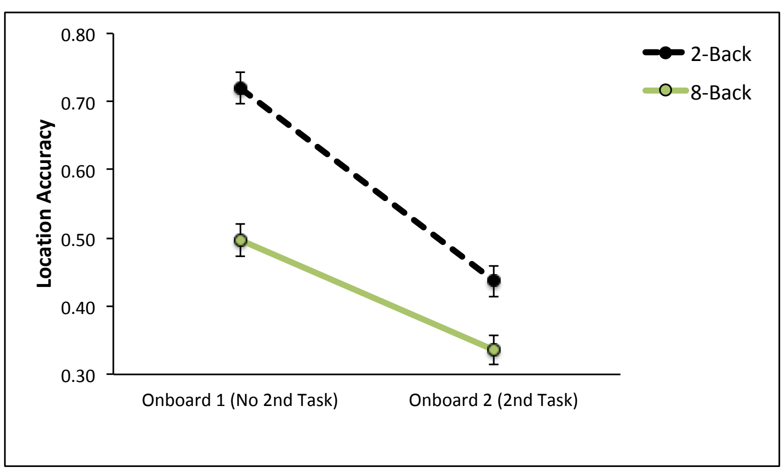


Figure 7. Location accuracy as a function of trial type and retention interval at test.

Location accuracy was higher for crew (*M* = .599) and crew-like (*M* = .551) subjects than for undergraduate students subjects (*M* = .435), *F*(2, 43) = 10.059, *MS* = .096, $p < .001$, $\eta_p^2 = .319$. However, the differences between subject types depended on test session and trial type, *F*(2, 43) = 4.346, *MS* = .022, $p = .019$, $\eta_p^2 = .168$. For the first test session, location accuracy was higher on prospective memory trials than on retrospective memory trials for all subject types. During the second onboard session, the same pattern was evident, but weaker for undergraduate subjects, mostly likely due to the fact that performance was approaching the floor for these subjects (see Figure 8).

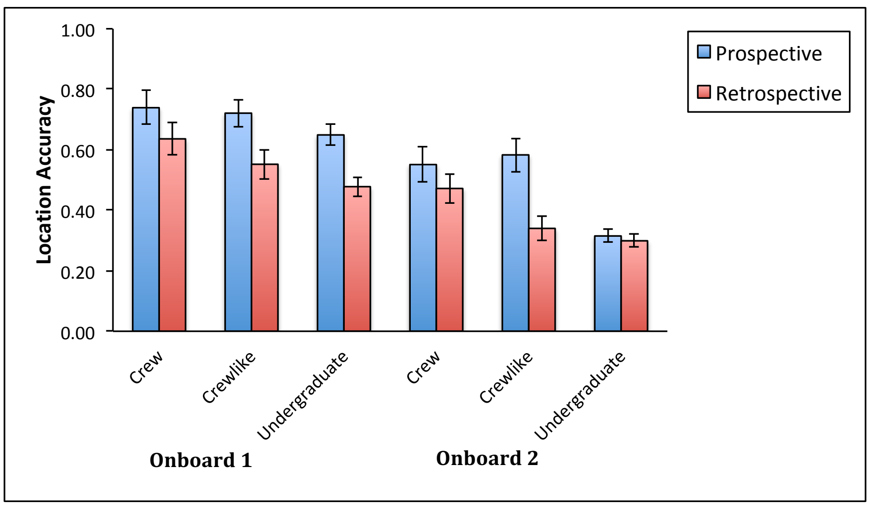


Figure 8. Location accuracy as a function of test session, subject type, and trial type.

For side accuracy at test, performance on retrospective memory trials (*M* = .845) was better than on prospective memory trials (*M* = .442), *F*(1, 43) = 111.572, *MSE* = .902, $p < .001$, $\eta_p^2 = .722$. Performance was also better at the short retention interval (*M* = .671) than at the long retention interval (*M* = .616), *F*(1, 43) = 23.841, *MSE* = .014, $p < .001$, $\eta_p^2 = .357$. The interaction between retention interval and trial type was also significant, *F*(1, 43) = 38.121, *MSE* = .015, $p < .001$, $\eta_p^2 = .470$, with side accuracy similar for short and long retention intervals on retrospective memory trials, but higher for short retention intervals on prospective memory trials (see Figure 9).

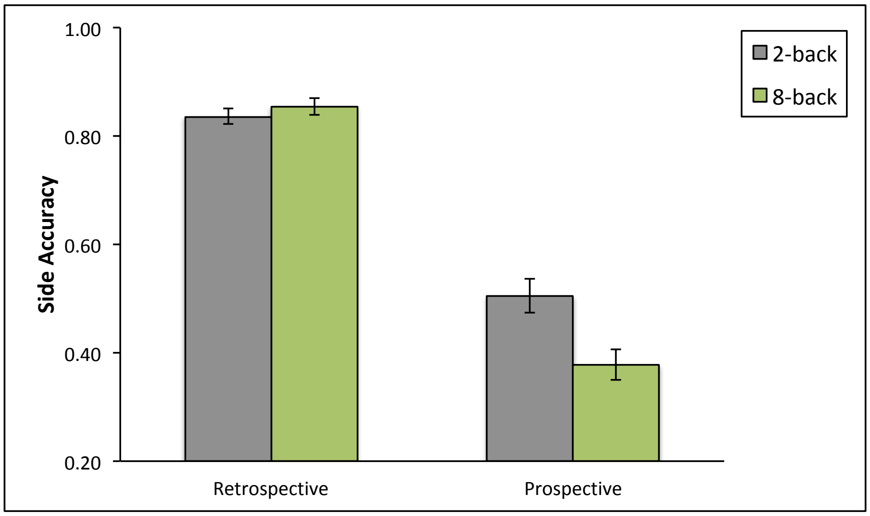


Figure 9. Side accuracy as a function of trial type and retention interval at test.

Side accuracy was higher for crew (M = .725) and crew-like (M = .707) subjects than for undergraduate students subjects (M = .587), , $F(2, 43) = 10.959$, $MS = .071$, $p < .001$, $\eta_p^2 = .338$. Side accuracy was also higher for the first test (M = .708) than for the second test (M = .579), which involved the concurrent secondary task (M = .579), $F(1, 43) = 43.096$, $MS = .025$, $p < .001$, $\eta_p^2 = .501$. The three-way interaction between subject type, test session, and trial type was also significant, $F(2, 43) = 5.702$, $MS = .045$, $p = .006$, $\eta_p^2 = .210$, reflecting the fact that during the first onboard session, side accuracy was higher on retrospective memory trials than on prospective memory trials for all subject types. However, during the second onboard session, the same pattern was evident, but was especially strong for undergraduate subjects (see Figure 10). An additional aspect of this interaction is that undergraduate subjects were particularly impacted negatively by the concurrent secondary task, primarily on prospective memory trials.

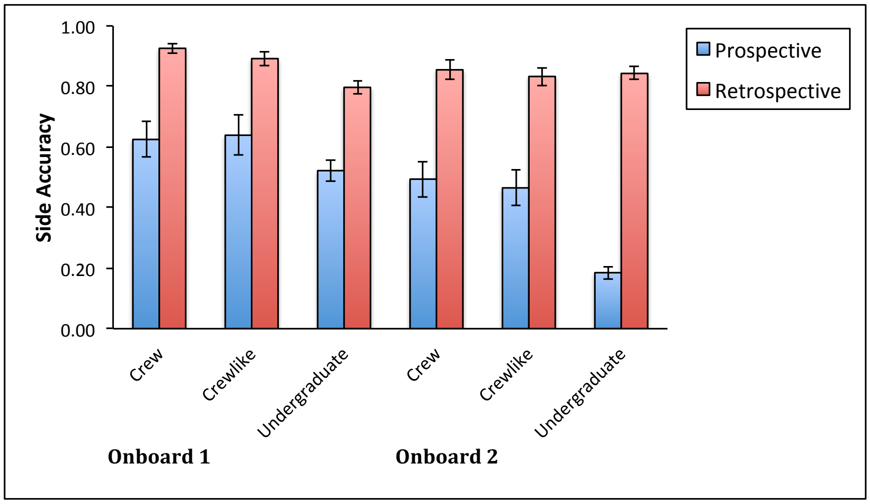


Figure 10. Side accuracy as a function of test session, subject type, and trial type.

Discussion. This experiment replicated the finding from Bourne et al. (2011) that special responding protects against forgetting associations. Specifically, an intention to respond in a special way not only preserved that intention (the link between the name and the side) but also sustained the associated information (the link between the name and the location).

The hypothesis raised by Bourne et al. (2011) for explaining the protective function of special responses concerns working memory through a mechanism of consciously refreshing the targets and their associations. However, there was no evidence that the protective function is due to holding the associations from the special trials in working memory (Baddeley, 2007). Instead the protection occurred even when working memory was unavailable for the primary task of associating names with locations because it was used for the concurrent secondary counting task. This finding importantly extends the generality of the protective function to situations in which working memory is unavailable, as occurs often in everyday life due to concurrent cognitive activities and task sharing. Understanding the cause of the protective function for special responding is, thus, an important challenge for future theorizing and research.

One hypothesis involves kinesthetic or spatial processing of location associations. Although the counting task should hinder processing of the phonological store of working memory (the "phonological loop") through its use of articulatory suppression (e.g., Papagno, Valentine, & Baddeley, 1991) and might also impact working memory's central executive (Marsh & Hicks, 1998), it should not influence the visuospatial sketchpad of working memory (Baddeley, 2007). During study, subjects saw location information as words (North, South, East, West), so subjects could code the stimuli verbally, using the phonological store. However, on each study trial, subjects responded to location by clicking in the appropriate square on the screen. Thus, subjects might have learned location kinesthetically (e.g., remembering an upward movement trajectory for North) or spatially/visually (e.g., remembering the top-most location for North), rather than verbally in terms of words they read. A similar set of processes could account for evidence for the protective function from Bourne et al. (2011) because subjects in that study also responded by clicking on specific locations (representing different colors).

Whether subjects use phonological, kinesthetic, or spatial working memory processes, these accounts all imply that subjects consciously maintain information from previous trials within a given block in working memory. In contrast, subjects, when given the cue for a special (prospective) trial (i.e., they see a name that had recently occurred in green font and all capital letters), might retrieve the associated cardinal-direction location information spontaneously, as suggested by the spontaneous retrieval hypothesis for prospective memory proposed by Einstein et al. (2005), with no conscious resources devoted to these monitoring processes (but also see Marsh, Hicks, & Cook, 2006, for a demonstration that the intention to respond in a subsequent memory test can interfere with performing other concurrent ongoing activities). The continuous memory updating procedures used in the present study (and by Bourne et al., 2011) are not the same as those used previously to investigate prospective memory -- "remembering to do something in the future" (Marsh & Hicks, 1998, p. 336). Nevertheless, the present findings might involve cognitive processes similar to those involved in earlier studies of prospective memory. The present study, unlike most previous studies of prospective memory, compared directly prospective and retrospective responding (with a substantial number of both types of responses at both short and long retention intervals for every subject). Thus, the present findings might shed light on the longstanding issue of whether prospective and retrospective memories have different forgetting functions (see, e.g., Hicks, Marsh, & Russell, 2000), suggesting a reduced rate of forgetting for prospective relative to retrospective memory.

Alternative processes that might account for the present findings and that also do not require conscious monitoring are automatic binding processes conjoining name and location information (e.g., Chalfonte & Johnson, 1996; Treisman & Zhang, 2006; Udale, Farrell, & Kent, 2018). However, an account involving automatic binding would need to specify why such processes would apply only to prospective memory, not to retrospective memory.

In conclusion, if special items signal importance and if individuals remember that a given item was important, information associated with that item should be protected against forgetting across short and long retention intervals, possibly even when conscious thoughts are being devoted to other activities. Although we have drawn these conclusions based on findings from a particular laboratory task involving the learning of name-location associations and given relatively short retention intervals, we speculate that the same conclusions would hold both in other laboratory tasks and in tasks occurring in everyday life even over longer retention intervals. For a sample task common in everyday life (as well as in space), consider the need to take some medicine at breakfast. Thus, breakfast can replace the name as a cue. And the particular medicine to take can replace the location association. Our results suggest that if you can remember the need to take your medicine at breakfast, which medicine to take should be protected against forgetting, even when your conscious thoughts are directed elsewhere.

As with the data entry task, results from the mapping task also address the question of whether the results obtained with college students on the complex memory task would be found with crew and crew-like subjects as well. Again, it was found that crew and crew-like subjects performed more accurately than the university students overall in terms of both location and side accuracy, with the crew subjects also performing somewhat better than the crew-like subjects. Although there were a few significant interactions of subject type with other variables, the most critical and meaningful interaction in this experiment was the three-way interaction of subject type, test session, and trial type on side accuracy at test. Much reduced side accuracy on prospective memory was found for undergraduate subjects relative to crew and crew-like subjects in the presence of the concurrent secondary task, although not in the absence of that task. This interaction implies that the crew and crew-like subjects are better able than undergraduate students to cope with the increased cognitive load demanded by the secondary counting task, and it resembles the finding in Experiment 1 that the crew and crew-like subjects are better able than the undergraduate subjects to cope with the high cognitive load demanded by the code task. Again, this pattern demonstrates that results obtained in studies using undergraduate students as test subjects may not predict the pattern of responses of crew members, especially under cognitive load. Thus, to be able to predict the cognitive performance of crew members on Earth, we must study crew members, or at least crew-like subjects.

Given the criticality of cognitive performance during space missions and given the significant cognitive loads space missions impose on the astronaut crew, the findings reported here have important implications for HRP as many theories and research assumptions are driven by university-based research. The research problem is compounded by our inability to replicate on the ground the space environment and all its loads. If we want to be able to predict the cognitive performance of crew members on space missions, we must study crew members in space.

Future Research

The original study had 3 specific aims (see Background in Appendix B):

> *Aim A***.** Test the retention and transfer of specific motor, perceptual and cognitive processes learned pre-launch to assess the need for (and possible schedule of) onboard refresher and just-in-time (JIT) training.
>
> *Aim B.* Compare the process of knowledge/skill decay on orbit with that of closely matched participants on Earth and with that of undergraduate university students to determine the applicability of existing Earth-based theories of learning to space operations.
>
> *Aim C*. Collect naturalistic data from onboard crew and ground-control personnel on training-related crew performance including performance errors, requests for ground support, need to review previously learned material, and training success stories.

The portion of the study reported here only addresses a portion of Aim A. This ground-based portion of the study is now completed, but this portion of the overall study is of very limited utility without the space-based portion. All three aims need to be addressed in full to be able to respond the original reason for which the HRP Program Scientist requested this study.

If the on-orbit portion of the study were to be re-instated, as promised when the study was removed from the crew timeline, progress could be made towards the Training Risk Gaps closure. The results of the Ground Phase alone do not by themselves provide input towards gap closure, because we still don't know the effects of long duration space flight on training retention and transfer. On the contrary! The current results suggest that the gaps are in fact wider than originally assumed.

The results of the full research study would allow us to develop laws of acquisition (learning), retention (memory), and transfer (generalization) for the astronaut population and more importantly for astronauts in space flight. These results would inform the design of training for NASA's future deep space missions, including a manned mission to Mars, and will inform the need for and possible scheduling of onboard refresher and JIT training.

Acknowledgements

The ground-phase of the study was conducted in collaboration with Alice Healy, Vivian Schneider and Carolyn Buck-Gengler at the University of Colorado, Boulder, and with James Kole at the University of Northern Colorado. Our university collaborators participated in the design of the study, the development of the software programs used to collect and analyzed data, as well as in writing up the results. In addition, our university collaborators conducted the portion of the study involving university undergraduate students, have presented results from that portion of the study at scientific conferences (Healy, Kole, Schneider, & Barshi, 2019b; Healy, Schneider, Buck-Gengler, Kole, & Barshi, 2018; Kole, Schneider, Healy, & Barshi, 2017; Schneider, Healy, Buck-Gengler, Kole, & Barshi, 2016; Schneider, Healy, Kole, & Barshi, 2016; Schneider, Healy, Kole, & Barshi, 2015) and published articles in scientific peer-reviewed journals (Healy, Kole, Schnieder & Barshi, 2019a; Healy, Schneider, Buck-Gengler, Kole, & Barshi, under review).

References

(Note, the reference section here includes references cited in the appendices below)

Baddeley, A. (2007). *Working memory, thought, and action*. New York, NY: Oxford University Press. doi: 10.1093/acprof:oso/9780198528012.001.0001

Bahrick, H. P., & Phelps, E. (1988). The maintenance of marginal knowledge. In U. Neisser & E. Winograd (Eds.), Remembering reconsidered: Ecological and traditional approaches to the study of memory (pp. 178-192). New York: Cambridge University Press. doi: 10.1017/CBO9780511664014.008

Bahrick, H. P., Bahrick, P. O., & Wittlinger, R. P. (1975). Fifty years of memory for names and faces: A cross-sectional approach. *Journal of Experimental Psychology: General, 104,* 54-75. doi: 10.1037/0096-3445.104.1.54

Baldwin, T.T & Ford, J.K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

Barshi, I. (2006). Adaptable Training and Decision Support. DVD submitted to NASA/ESMD, January 25, 2006.

Barshi, I. (2011). Crew Training: Gaps and Challenges. White Paper submitted to NASA/SHFE, September 15, 2011.

Barshi, I. (2015). From Healy's training principles to training specifications: The case of the Comprehensive LOFT. *American Journal of Psychology, 128,* 219-227.

Barshi, I., & Loukopoulos, L. D. (2012). Training for real world job performance. In A. F. Healy & L. E. Bourne, Jr. (Eds.). *Training cognition: Optimizing efficiency, durability, and generalizability* (pp. 287-306). New York: Psychology Press.

Bourne, L. E., Jr., Healy, A. F., Bonk, W. J., & Buck-Gengler, C. J. (2011). Intention to respond in a special way offers some protection against forgetting associations. *American Journal of Psychology, 124,* 23-36. doi: 10.5406/amerjpsyc.124.1.0023

Bourne, L. E., Jr., Healy, A. F., Parker, J. T., & Rickard, T. C. (1999). The strategic basis of performance in binary classification tasks: Strategy choices and strategy transitions. *Journal of Memory and Language, 41,* 223-252. doi: 10.1006/jmla.1999.2647

Byrne, V., Schmid, J., & Barshi, I. (2010). Performance Support Tools for Space Medical Operations. Poster presented at the Aerospace Medical Association conference, Phoenix, AZ.

Carraher, D., & Schliemann, A. D. (2002). The transfer dilemma. *Journal of the Learning Sciences, 11,* 1-24. doi: 10.1207/S15327809JLS1101_1

Chalfonte, B. L., & Johnson, M. K. (1996). Feature memory and binding in young and older adults. *Memory & Cognition, 24,* 403–416. doi: 10.3758/BF03200930

Chapman, M. J., Healy, A. F., & Kole, J. A. (2016). Memory load as a cognitive antidote to performance decrements in data entry. *Memory*, *24,* 1182-1196 . doi: 10.1080/09658211.2015.1086380

Delaney, P. F., Reder, L. M., Staszewski, J. J., & Ritter, F. E. (1998). The strategy-specific nature of improvement: The power law applies by strategy within task. *Psychological Science, 9*, 1-7. doi: 10.1111/1467-9280.00001

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology*. New York: Teachers College, Columbia University. (Original work published 1885). doi: 10.1037/10011-000

Einstein, G. O., McDaniel, M. A., Thomas, R., Mayfield, S., Shank, H., Morrisette, N., & Breneiser, J. (2005). Multiple processes in prospective memory retrieval: Factors determining monitoring versus spontaneous retrieval. *Journal of Experimental Psychology: General, 134,* 327–342. doi: 10.1037/0096-3445.134.3.327

Ellis, J. A., Semb, G. B., & Cole, B. (1998). Very long-term memory for information taught in school. *Contemporary Educational Psychology, 23,* 419-433. doi: 10.1006/ceps.1997.0976

Evans, C. A., Robinson, J. A., Tate-Brown, J., Thumm, T., & Crespo-Richey, J. (2009). *International Space Station science research accomplishments during the assembly years: An analysis of results from 2000-2008*. NASA Technical Paper TP-2009-213146.

Fendrich, D. W., Healy, A. F., & Bourne, L. E. (1993). Mental arithmetic: Training and retention of multiplication skill. In C. Izawa (Ed.), *Cognitive psychology applied* (pp. 111-133). Hillsdale, New Jersey: Erlbaum.

Fendrich, D. W., Healy, A. F., & Bourne, L. E., Jr. (1991). Long-term repetition effects for motoric and perceptual procedures. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17,* 137-151. doi: 10.1037/0278-7393.17.1.137

Fischer, U., & Mosier, K. (2014). The Impact of Communication Delay and Medium on Team Performance and Communication in Distributed Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *58*(1),115-119. https://doi.org/10.1177/1541931214581025

Flynn, C. F. (2005). An Operational Approach to Long-Duration Mission Behavioral Health and Performance Factors. *Aviation, Space, and Environmental Medicine, 76*(6), B42-b51.

Hah, S., Willems, B., Ayaz, H., Bunce, S., Izzetoglu, K., & Deshmukh, A. (2011, May). The comparison of air traffic controllers' to college students' memory performance and brain activities. Paper presented at the 16th International Symposium on Aviation Psychology, Dayton, OH.

Healy, A. F. (2007). Transfer: Specificity and generality. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 271-275). New York: Oxford University Press.

Healy, A. F., & Bourne, L. E., Jr. (2012b). Basic research on training principles. In A. F. Healy & L. E. Bourne, Jr. (Eds.), *Training cognition: Optimizing efficiency, durability, and generalizability* (pp. 40-66). New York: Psychology Press.

Healy, A. F., & Bourne, L. E., Jr. (Eds.). (1995). *Learning and memory of knowledge and skills*: *Durability and specificity*. Thousand Oaks, CA: Sage. doi: 10.4135/9781483326887

Healy, A. F., & Bourne, L. E., Jr. (Eds.). (1998). *Foreign language learning: Psycholinguistic studies on training and retention*. Mahwah, NJ: Erlbaum.

Healy, A. F., & Bourne, L. E., Jr. (Eds.). (2012a). *Training cognition: Optimizing efficiency, durability, and generalizability*. New York: Psychology Press.

Healy, A. F., & Wohldmann, E. L. (2012). Specificity and transfer of learning. In B. H. Ross (Ed.), *The psychology of learning and motivation* (Vol. 57, pp. 227-253). San Diego: CA: Elsevier Academic Press.

Healy, A. F., Clawson, D. M., McNamara, D. S., Marmie, W. R., Schneider, V. I., Rickard, T. C., Crutcher, R. J., King, C., Ericsson, K. A., & Bourne, L. E., Jr. (1993). The long-term retention of knowledge and skills. In D. Medin (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 30, pp. 135-164). New York: Academic Press.

Healy, A. F., Fendrich, D. W., Crutcher, R. J., Wittman, W. T., Gesi, A. T., Ericsson, K. A., & Bourne, L. E., Jr. (1992). The long-term retention of skills. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 87-118). Hillsdale, NJ: Erlbaum.

Healy, A. F., Fendrich, D.W., Crutcher, R .J., Wittman, W. T., Gesi, A. T., Ericsson, K. A., & Bourne, L. E., Jr. (1992). The long-term retention of skills. In A. F. Healy, S. M. Kosslyn, & R. M. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes, Volume 2* (pp. 87-118). Hillsdale, NJ: Erlbaum.

Healy, A. F., Kole, J. A., Buck-Gengler, C. J., & Bourne, L. E., Jr. (2004). Effects of prolonged work on data entry speed and accuracy. *Journal of Experimental Psychology: Applied, 10,* 188-199. doi: 10.1037/1076-898X.10.3.188

Healy, A. F., Kole, J. A., Schneider, V. I., & Barshi, I. (2019). Training, retention, and transfer of data entry perceptual and motor processes over short and long retention intervals. *Memory & Cognition, 47, 1606-1618*. doi: 10.3758/s13421-019-00955-z

Healy, A. F., Kole, J. A., Schneider, V. I., & Barshi, I. (2019b, November). *Training, retention, and transfer of data entry perceptual and motor processes over short and long retention intervals*. Presented at the 60th Annual Meeting of the Psychonomic Society, Montreal, Canada.

Healy, A. F., Kole, J. A., Wohldmann, E. L., Buck-Gengler, C. J., & Bourne, L. E., Jr. (2011). Data entry: A window to principles of training. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 277-296). New York: Psychology Press.

Healy, A. F., Schneider, V. I., & Barshi, I. (2015). Specificity and transfer in learning how to follow navigation instructions. In J. G. W. Raaijmakers, A. H. Criss, R. L. Goldstone, R. M. Nosofsky, & M. Steyvers (Eds.), *Cognitive modeling in perception* *and memory: A Festschrift for Richard M. Shiffrin* (pp. 259-273). New York: Psychology Press.

Healy, A. F., Schneider, V. I., Buck-Gengler, C. J., Kole, J. A., & Barshi, I. (under review). Intention to respond in a special way protects against forgetting associations even when working memory is occupied. *Journal of Memory*.

Healy, A. F., Schneider, V. I., Buck-Gengler, C. J., Kole, J. A., & Barshi, I. (2018, November). *Intention to respond in a special way protects against forgetting associations even when working memory is occupied*. Presented at the 59th Annual Meeting of the Psychonomic Society, New Orleans, LA.

Healy, A. F., Schneider, V. I., Buck-Gengler, C. J., Kole, J. A., & Barshi, I. (2019). *Intention to respond in a special way protects against forgetting associations even when working memory is occupied*. Manuscript submitted for publication.

Healy, A. F., Wohldmann, E. L., & Bourne, L. E., Jr. (2005). The procedural reinstatement principle: Studies on training, retention, and transfer. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 59-71). Washington, DC: American Psychological Association. doi: 10.1037/10895-005

Healy, A. F., Wohldmann, E. L., Kole, J. A., Schneider, V. I., Shea, K. M., & Bourne, L. E., Jr. (2013). Training for efficient, durable, and flexible performance in the military. In W. Arthur, Jr., E. A. Day, W. Bennett, Jr., & A. M. Portrey (Eds.), *Individual and team skill decay: The science and implications for practice* (pp. 176-204). New York: Routledge.

Healy, A. F., Wohldmann, E. L., Parker, J. T., & Bourne, L. E., Jr. (2005). Skill training, retention, and transfer: The effects of a concurrent secondary task. *Memory & Cognition, 33,* 1457-1471. doi: 10.3758/BF03193378

Healy, A.F., Kole, J.A., Schneider, V.I., Barshi, I. (2019a) Training, retention, and transfer of data entry perceptual and motor processes over short and long retention intervals. *Memory & Cognition, 47,* 1606-1618. https://doi.org/10.3758/s13421-019-00955-z

Hicks, J. L., Marsh, R. L., & Russell, E. J. (2000). The properties of retention intervals and their affect on retaining prospective memories. *Journal of Experimental*

Psychology: Learning, Memory, and Cognition, 26, 1160–1169. doi: 10.1037/0278-7393.26.5.1160

Holden, K., Muson, B., Russi-Vigoya, N., Dempsey, D., Adelstein, B. (in press). *Human capabilities assessment for autonomous missions (HCAAM) Phase II: Development and validation of an autonomous operations task list – Final report*. NASA/TP-2019-XXXXX.

Holland, A., Hysong, S., & Galarza, L. (2007). A review of training methods and instructional techniques: Implications for behavioral skills training in U.S. Astronauts (Technical Publication): NASA/TP-2007-21372.

Johnson, A., & Proctor, R. W. (2017). *Skill acquisition and training: Achieving expertise in simple and complex tasks*. New York, NY: Routledge.

JSC Committee for Protection of Human Subjects (CPHS): Guidelines for Investigations Proposing Human Research for Space Flight and Related Investigations (JSC20483, http://irb.nasa.gov).

Kole, J. A., Healy, A. F., & Bourne, L. E., Jr. (2008). Cognitive complications moderate the speed-accuracy tradeoff in data entry: A cognitive antidote to inhibition. *Applied Cognitive Psychology, 22*, 917-937. doi: 10.1002/acp.1401

Kole, J.A., Schneider, V. I., Healy, A. F., & Barshi, I. (2017, November). *Training, retention, and transfer of data entry perceptual and motoric processes over long retention intervals*. Presented at the 58th Annual Meeting of the Psychonomic Society, Vancouver, British Columbia, Canada.

Lohse, K. R., & Healy, A. F. (2012). Exploring the contributions of declarative and procedural information to training: A test of the procedural reinstatement principle. *Journal of Applied Research in Memory and Cognition, 1,* 65-72. doi: 10.1016/j.jarmac.2012.02.002

Lohse, K. R., & Healy, A. F. (2012). Exploring the contributions of declarative and procedural information to training: A test of the procedural reinstatement principle. *Journal of Applied Research in Memory and Cognition, 1,* 65-72. doi: 10.1016/j.jarmac.2012.02.002

Loukopoulos, L. D., Dismukes, R. K., & Barshi, I. (2009). The Multitasking Myth: Handling complexity in real-*world operations*. Burlington, VT: Ashgate.

Marsh, R. L., & Hicks, J. L. (1998). Event-based prospective memory and executive control of working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 336–349. doi: 10.1037/0278-7393.24.2.336

Marsh, R. L., Hicks, J. L., & Cook, G. I. (2006). Task interference from prospective memories covaries with contextual associations of fulfilling them. *Memory & Cognition, 34,* 1037–1045. doi: 10.3758/BF03193250

McDaniel, M. A. (2007). Rediscovering transfer as a central concept. In H. L. Roediger, III, Y. Dudai, & S. M. Fitzpatrick (Eds.), *Science of memory: Concepts* (pp. 267-270). New York: Oxford University Press.

Murre, J. M. J., & Dros, J. (2015). Replication and analysis of Ebbinghaus' forgetting curve. *PLoS One, 10* (7): e0120644. doi: 10.1371/journal.pone.0120644

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the power law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56,* 132-143. doi: 10.1037/h0057488

Pachella, R. G. (1974). The interpretation of reaction time in information-processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41-82). Hillsdale, NJ: Erlbaum.

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*, 710-756. doi: 10.1037/bul0000151

Papagno, C., Valentine, T., & Baddeley, A. D. (1991). Phonological short-term memory and foreign-language vocabulary learning. *Journal of Memory and Language, 30,* 331-347. doi: 10.1016/0749-596X(91)90040-Q

Performance Factors. *Aviation, Space, and Environmental Medicine, 76*(6), B42-b51.

Rickard, T. C. (1997). Bending the power law: A CMPL theory of strategy shifts and the automatization of cognitive skills. *Journal of Experimental Psychology: General, 126,* 288-311. doi: 10.1037/0096-3445.126.3.288

Rickard, T. C., & Bourne, L. E., Jr. (1996). Some tests of an identical elements model of basic arithmetic skills. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22,* 1281–1295. doi: 10.1037/0278-7393.22.5.1281

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734-760. doi: 10.1037/0033-295X.103.4.734

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207-217. doi: 10.1111/j.1467-9280.1992.tb00029.x

Schneider, V. I., Healy, A. F., & Bourne, L. E., Jr. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language, 46,* 419-440. doi: 10.1006/jmla.2001.2813

Schneider, V. I., Healy, A. F., Buck-Gengler, C. J., Kole, J. A., & Barshi, I. (2016, November). *Distinctive responding protects against forgetting associations: Role of working memory*. Presented at the 57th Annual Meeting of the Psychonomic Society, Boston, MA.

Schneider, V. I., Healy, A. F., Kole, J. A., & Barshi, I. (2015, November). *Training, retention, and transfer of data entry perceptual and motor processes*. Presented at the 56th Annual Meeting of the Psychonomic Society, Chicago, IL.

Schneider, V. I., Healy, A. F., Kole, J. A., & Barshi, I. (2016, April). *Training data entry: Specificity and generalizability of perceptual and motor processes*. Invited paper presented in the symposium Studies from the Center for Research on Training, University of Colorado Boulder. Rocky Mountain Psychological Association Convention, Denver, CO.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317-1323. doi: 10.1126/science.3629243

Singley, M. K., & Anderson, J. R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.

Squire, L. R. (1989). On the course of forgetting in very long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 241-245. doi: 10.1037/0278-7393.15.2.241

Thorndike, E. L. (1906). *The principles of teaching based on psychology*. New York: A. G. Seiler. doi: 10.1037/11487-000

Treisman, A., & Zhang, W. (2006). Location and binding in visual working memory. *Memory & Cognition, 34,* 1704-1719. doi: 10.3758/BF03195932

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352-373. doi: 10.1037/h0020071

Udale, R., Farrell, S., & Kent, C. (2018). No evidence of binding items to spatial configuration representations in visual working memory. *Memory & Cognition, 46,* 955-968. doi: 10.3758/s13421-018-0814-8

Urban, D., Griffin, D., Ruff, G., Cleary, T., Yang, J., Mulholland, G., & Yuan, Z. (2005) *Detection of smoke from microgravity fires*. Proceedings of the International Conference on Environmental Systems, 2005-01-2930.

Wenzel, B. M., Barshi, I., Dempsey, D. L., & Feldman, J. M. (2019). Human capabilities assessments for autonomous missions (HCAAM) integration: Just-in-time training - Final Report. NASA/TP- 2019-XXXXX.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition, 2,* 775–780. doi: 10.3758/BF03198154

Wixted, J. T., & Carpenter, S. K. (2007). The Wickelgren power law and the Ebbinghaus savings function. *Psychological Science, 18,* 133-134. doi: 10.1111/j.1467-9280.2007.01862.x

Wohldmann, E. L., & Healy, A. F. (2010). Exploring specificity of speeded aiming movements: Examining different measures of transfer. *Memory & Cognition, 38,* 344-355. doi: 10.3758/MC.38.3.344

Wu, S., & Vera, A. H. (2019). *Supporting crew autonomy in deep space exploration: Preliminary onboard capability requirements and proposed research questions. Technical Report of the Autonomous Crew Operations Technical Interchange Meeting.* NASA Technical Memorandum (NASA/TM-2019-220345).

Appendix A

Details of the Method, Results and Data Analyses of the Ground Phase Study

Performance on the Data Entry Task: Pre-Test and Post-Test

I. Method

•Twenty-six undergraduate students at the University of Colorado, 20 subjects from NASA (11 crew and 9 crew-like).

•Six experimental sessions distributed over 480 days:

-Session 1: *Pretest*. Subjects performed 50 trials of the standard data entry task (typing 4-digit numbers, e.g., 2154, using right hand), 50 trials of a word task (typing 4-digit numbers presented as words, e.g., two one five four, using right hand), 50 trials a left-hand task (typing 4-digit numbers, e.g., 2154, using left hand), 50 trials of a 3-digit task (typing 3-digit numbers, e.g., 215, using right hand), and 50 trials of a code task (typing positions of letters within the alphabet, e.g., baed, using right hand).

-Session 2: *Training Session 1*. Subjects competed three blocks of trials of the standard data entry task. Each block consisted of 100 trials, and the same 100 numbers were repeated in each of the three blocks.

-Session 3: *Training Session 2*. Identical to Training Session 1; subjects competed three blocks of trials of the standard data entry task. Each block consisted of 100 trials, and the same 100 numbers were used as during the first training session.

-Session 4: *Onboard Session 1*. Subjects completed 100 trials of the standard task followed by 100 trials of the left-hand task. For both the standard task and the left-hand task, fifty trials were on old numbers (numbers that had been practiced during training), and fifty trials were on new numbers.

-Session 5: *Onboard Session 2*. Subjects completed 100 trials of the standard task followed by 100 trials of the code task. For both the standard task and the code task, fifty trials were on old numbers (numbers that had been practiced during training), and fifty trials were on new numbers. To create old numbers for the code task, digits for each old 4-digit number were replaced with letters (a = 1, b = 2, etc.).

-Session 6: *Post-test*. The post-test was identical to the pre-test.

•Dependent measures included: accuracy; initiation time (time elapsed from stimulus presentation to entering the first digit; primarily a measure of cognitive processes such as perception and response planning); execution time (time required to enter the second, third, and fourth digits; primarily a measure of motoric processes); conclusion time (time required to enter the concluding *Enter* key for the trial; also thought to be primarily a measure of motoric processes); and total response time (time required to enter all 4 digits and the concluding *Enter* key).

II. Pre-Test and Post-Test Analyses

•Comparison of pre-test to post-test for 5 tasks (standard, word, left-hand, 3-digit, and code). Subject type (crew, crew-like, undergraduate) also included as a factor.

•5 dependent measures: TRT, response time components (initiation time, execution time, conclusion time), and accuracy. Response time components generally showed the same patterns as TRT; thus, TRT and accuracy analyses are reported here and response time component analyses are reported in *Pre-Test/Post-Test Response Time Component Analyses*.

•Pre-test data only were also analyzed to check for initial differences between crew, crew-like, and undergraduate subjects.

III. Pre-Test & Post-Test Results: Accuracy

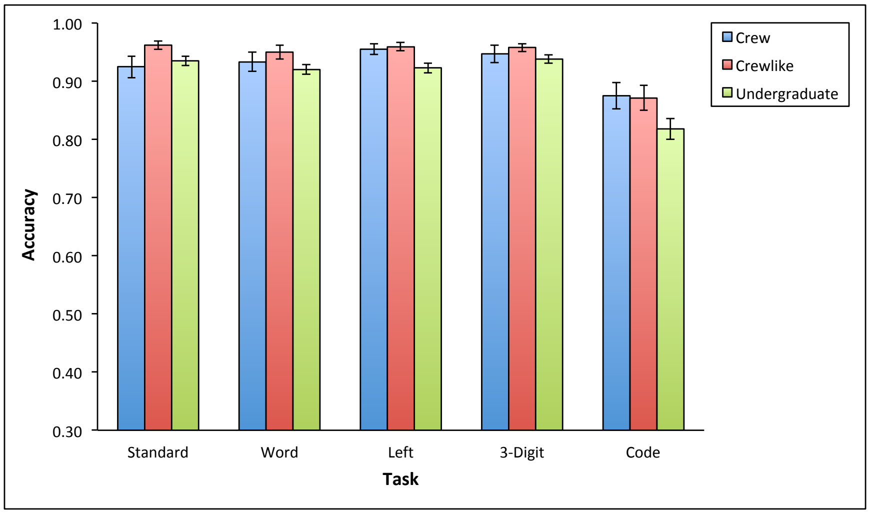


Figure 1. Accuracy as a function of task and subject type averaging across pre-test and post-test. *Note*: Here and in all subsequent figures, error bars represent between-subjects standard errors of the mean.

•There were overall differences in accuracy between tasks. Averaging across subject type, accuracy was similar for standard data entry (.938), word (.928), left-hand (.937), and 3-digit (.944) tasks, but lower for the code task (.841), $F(4, 168) = 28.747$, $MSE = .004$, $p < .001$.

•In the analysis restricted to the pre-test to check for pre-existing differences between groups, there were overall differences in accuracy between subject types during the pre-test (not shown in graph, which includes both pre-test and post-test). Accuracy was higher for crew (.942) and crew-like subjects (.945) than for undergraduate subjects (.911), $F(2, 42) = 4.039$, $MSE = .007$, $p = .025$.

Crew-Crew-like, $p = .856$
Crew-Undergraduate, $p = .028$
Crew-like-Undergraduate, $p = .031$

•An analysis restricted to the post-test revealed that the differences between subject types were not significant, $F(2, 42) = .894$, $MSE = .018$, $p = .417$.

IV. Pre-Test & Post-Test Results: Total Response Time

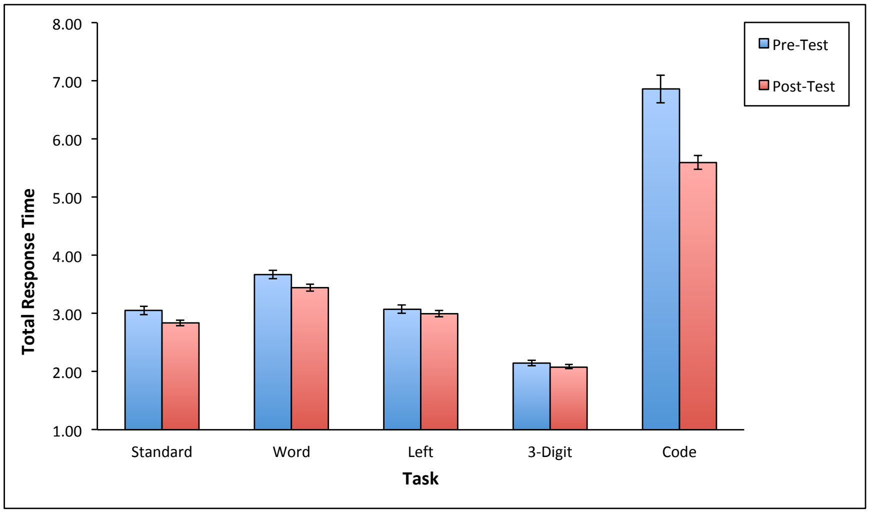


Figure 2. Total response time as a function of task and test (pre-test, post-test) averaging across subject type.

•There were overall differences in TRT between tasks. Averaging across pre- and post-test, TRT was slowest for code (6.227 s), fastest for 3-digit (2.111 s), and intermediate for standard (2.945 s), word (3.556 s), and left-hand (3.013 s) tasks, $F(4, 168) = 376.139$, $MSE = .439$, $p < .001$.

•TRT decreased from pre-test (3.759 s) to post-test (3.381 s), averaging across tasks, $F(1, 42) = 35.050$, $MSE = .298$, $p < .001$.

•The decrease in TRT from pre-test to post-test was greater for the code task than for the other tasks, $F(4, 168) = 28.299$, $MSE = .118$, $p < .001$.

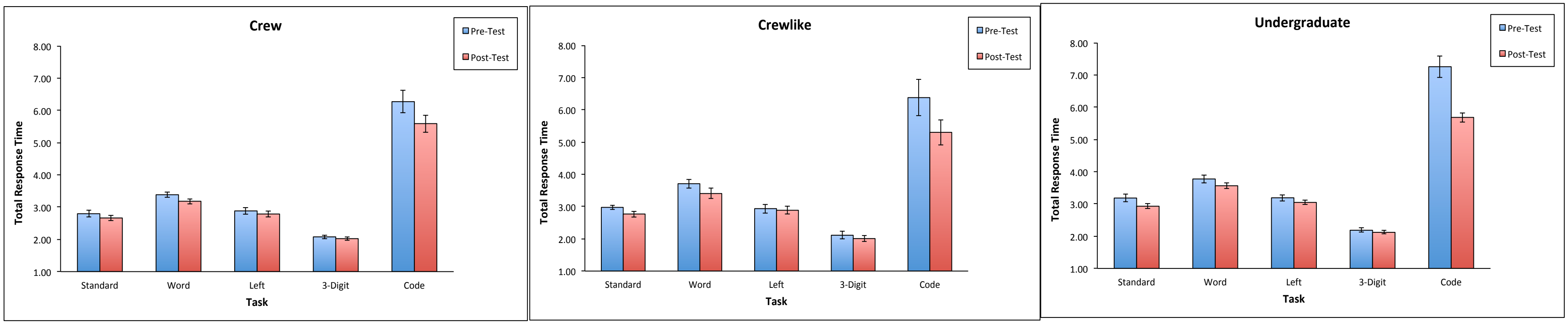


Figure 3. Total response time as a function of task and test (pre-test, post-test) for crew (left panel), crew-like (middle panel), and undergraduate (right panel) subjects.

•The interaction between task and session depended on subject type. For all subject types (crew, crew-like, undergraduate), TRT decreased from pre-test to post-test for all 5 tasks. However, crew subjects showed a less dramatic decrease for the code task relative to the other tasks than did crew-like and undergraduate subjects. This result might be partly due to the fact that crew subjects were faster initially than other groups, $F(8, 168) = 2.594$, $MSE = .118$, $p = .011$.

•Marginally significant differences in TRT between subject types during the pre-test. Crew (3.483 s) subjects were significantly faster than undergraduate subjects (3.918 s), $F(2, 42) = 2.692$, $MSE = .306$, $p = .079$.

Crew-Crew-like, $p = .588$

Crew-Undergraduate, $p = .034$

Crew-like-Undergraduate, $p = .194$

V. Summary of Pre- and Post-Test Results

Pre-test to post-test analyses reveal the extent to which subjects maintain or improve learned skills over the 480-day period as a result of practice received during pre-test as well as training and onboard sessions. There were significant differences between groups during the pre-test, with both crew and crew-like subjects demonstrating higher accuracy than undergraduate subjects, and crew subjects demonstrating faster speed than undergraduate subjects. These significant differences between subject types were eliminated at the post-test; with increased experience over two training sessions and two test sessions, undergraduate subjects were able to reach similar levels of performance as crew and crew-like subjects.

For both pre-test and post-test, there were differences between tasks in total response time and accuracy, reflecting differences in task difficulty. Overall, subjects maintained/improved skills over the 480-day period: for all tasks there was a significant improvement (in speed) from pre-test to post-test. However, this improvement was most pronounced for the code task, presumably because subjects were closer to a performance floor for the other tasks.

Interestingly, the decrease in TRT from pre-test to post-test occurred for the word task ($p = .003$) as well as the 3-digit task ($p = .054$) even though subjects did not perform either of these tasks during training or onboard sessions. Two possible explanations are: (1) the practice received on the standard, left-hand, and code tasks during the training and onboard sessions transfer to the word and 3-digit tasks; or (2) the experience received with the word and 3-digit tasks during pre-test was retained over the 480-day period to the post-test.

Performance on the Data Entry Task: Training Sessions

I. Method

•Subjects competed three blocks of trials of the standard data entry task. Each block consisted of 100 trials, and the same 100 numbers were repeated in each of the three blocks during both Training Session 1 and Training Session 2.

II. Training Sessions Analyses

•Standard data entry task only; comparison between training sessions (session 1, session 2), and block within each training session (block 1, block 2, block 3). Subject type (crew, crew-like, undergraduate) also included as a factor.

•5 dependent measures: TRT, response time components (initiation time, execution time, conclusion time), and accuracy. Response time components generally showed the same patterns as TRT unless otherwise noted. TRT and accuracy analyses are reported here and response time component analyses are reported in the section *Training Response Time Component Analyses*.

III. Training Sessions Results: Accuracy

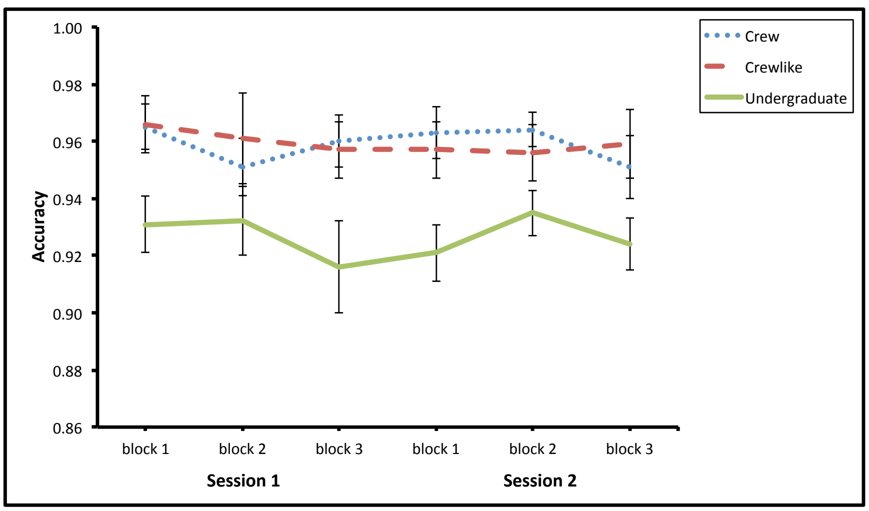


Figure 4. Accuracy as a function of training session, block, and subject type.

•Averaging across training sessions and blocks, there were overall differences in accuracy between subject groups. Crew (.959) and crew-like subjects (.959) demonstrated higher accuracy than did undergraduate subjects (.926), $F(2, 41) = 3.566$, $MSE = .010$, $p = .037$.

Crew-Crew-like, $p = .997$
Crew-Undergraduate, $p = .036$
Crew-like-Undergraduate, $p = .43$

IV. Training Sessions Results: Total Response Time

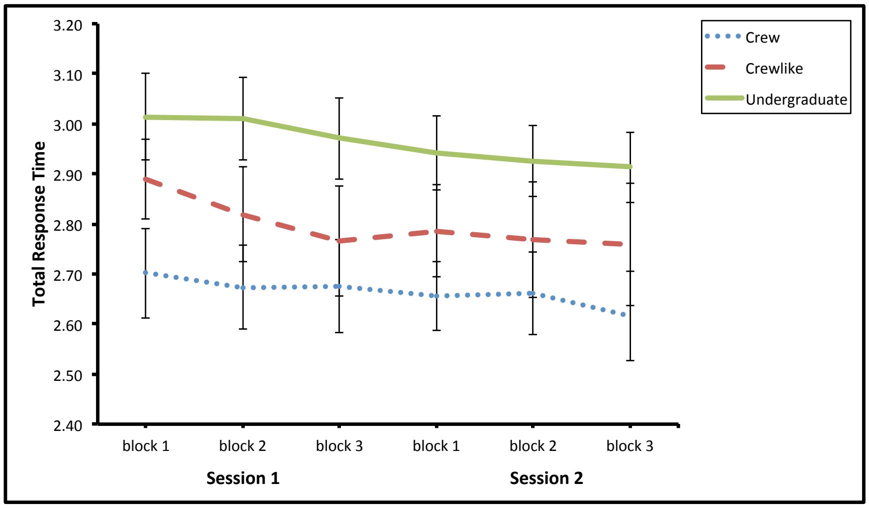


Figure 5. Total response time as a function of training session, block, and subject type.

•Averaging across blocks within each session, TRT decreased from the first training session to the second training session, $F(1, 41) = 4.895$, $MSE = .033$, $p = .033$, and decreased across blocks within each training session, $F(2, 82) = 4.179$, $MSE = .010$, $p = .019$, both of which reflect practice effects. Note that from the third block of Training Session 1 to the first block of Training Session 2, subjects maintained or even improved TRT, despite the fact that there was a 3-month retention interval between the two sessions.

•There were marginally significant differences in TRT between subject types during training, averaging across training sessions and blocks (crew = 2.664 s; crew-like = 2.798 s; undergraduate = 2.963 s), $F(2, 41) = 3.048$, $MSE = .671$, $p = .058$.

Crew-Crew-like, $p = .389$
Crew-Undergraduate, $p = .022$
Crew-like-Undergraduate, $p = .212$

V. Summary of Training Sessions Results

Training analyses reveal the extent to which subjects benefit from further experience with the standard data entry task. The analyses demonstrate that training does not impact accuracy (likely due to a ceiling effect), but does impact total response time across blocks as well as across training sessions. The response time component analyses (initiation, execution, and conclusion time) illuminate specifically which aspects of the data entry task are benefitted by practice (see section *Training Response Time Component Analyses*).

As during the pre-test, there were overall differences between crew, crew-like, and undergraduate subjects in terms of speed and accuracy. However, despite these overall differences, all groups' speed benefitted equally from practice.

Lastly, as in previous studies of data entry, there appears to be very little forgetting of the data entry task. For TRT, a comparison of block 3 of the first training session and block 1 of the second training session revealed no significant differences for any subject group, even though the training sessions were separated by three months, For a further discussion of retention including all sessions in this study, see section *Analysis of Retention Across All Sessions*.

Performance on the Data Entry Task: Onboard Sessions

<u>I. Method</u>

- For Onboard Session 1, subjects completed 100 trials of the standard task followed by 100 trials of the left-hand task. For both the standard task and the left-hand task, fifty trials were on old numbers (numbers that had been practiced during training), and fifty trials were on new numbers.

- For Onboard Session 2, subjects completed 100 trials of the standard task followed by 100 trials of the code task. For both the standard task and the code task, fifty trials were on old numbers (numbers that had been practiced during training), and fifty trials were on new numbers.

- Comparison of old numbers to new numbers on the standard data entry task reveals whether or not subjects retained what was learned during training. Comparison of old numbers to new numbers on the left-hand task reveals whether or not perceptual processes transfer from the standard data entry task to the left-hand task as the motoric requirements change due to the use of a different hand, but perceptual processes are the same between tasks because 4-digit numbers are presented. Comparison of old numbers to new numbers on the code task reveals whether or not motoric processes transfer from the standard data entry task to the code task as the perceptual requirements change due to the use of letters rather than numbers, but motoric processes are the same between tasks because subjects use their right hand.

<u>II. Onboard Sessions Analyses</u>

- Comparison between tasks (standard vs. left-hand for onboard session 1 and standard vs. code for onboard session 2), and number type (old, new). The variable of subject type (crew, crew-like, undergraduate) is also included as a factor.

- 5 dependent measures: TRT, response time components (initiation time, execution time, conclusion time), and accuracy. TRT and accuracy analyses are reported here and response time component analyses are reported in the section *Onboard Response Time Component Analyses*.

III. Onboard Sessions Results: Accuracy

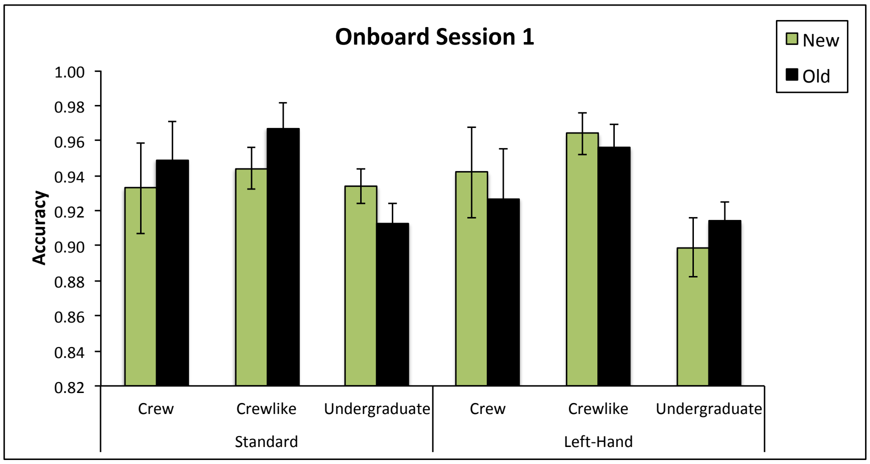


Figure 6. Accuracy as a function of task, subject type, and number type for Onboard Session 1.

•For the standard task, accuracy was higher for old numbers than for new numbers for crew and crew-like subjects; the opposite pattern was observed for undergraduate subjects. For the left-hand task the opposite pattern was observed: accuracy was higher for new numbers than for old numbers for crew and crew-like subjects, but accuracy was higher for old numbers than for new numbers for undergraduate subjects. The three-way interaction of task, number type, and subject type was significant, $F(2, 43) = 4.466$, $MSE = .001$, $p = .017$.

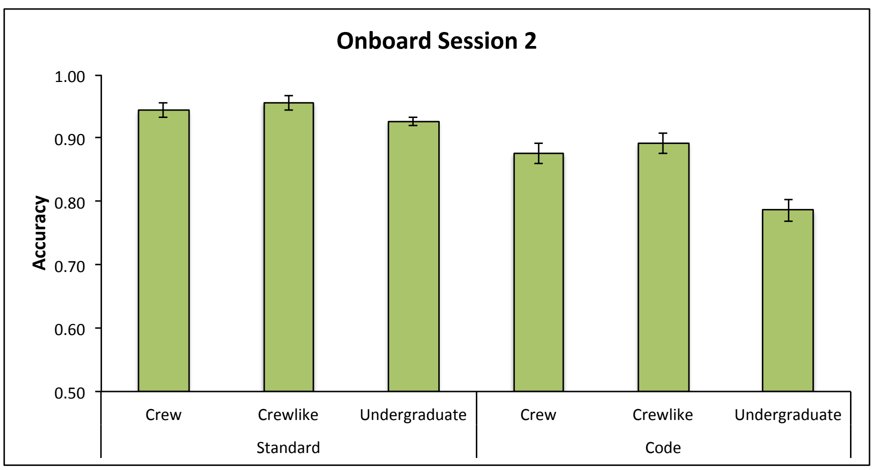


Figure 7. Accuracy as a function of task and subject type for Onboard Session 2.

•Significant differences in accuracy between subject types. Averaging across tasks, accuracy was higher for crew (.910) and crew-like (.924) subjects than for undergraduate subjects (.857), $F(2, 43) = 4.392$, $MSE = .019$, $p = .018$.

Crew-Crew-like, $p = .653$
Crew-Undergraduate, $p = .036$
Crew-like-Undergraduate, $p = .015$

•Significant differences in accuracy between tasks. Averaging across subject types, accuracy was higher for the standard data entry task (.937) than for the code task (.829), F(1, 43) = 41.277, MSE = .007, p < .001.
•For the standard task, accuracy was similar for crew, crew-like, and undergraduate subjects. For the code task, accuracy was similar for crew and crew-like subjects, but lower for undergraduate subjects; the interaction between task and subject type was significant, F(2, 43) = 4.337, MSE = .007, p = .019.

IV. Onboard Sessions Results: Total Response Time

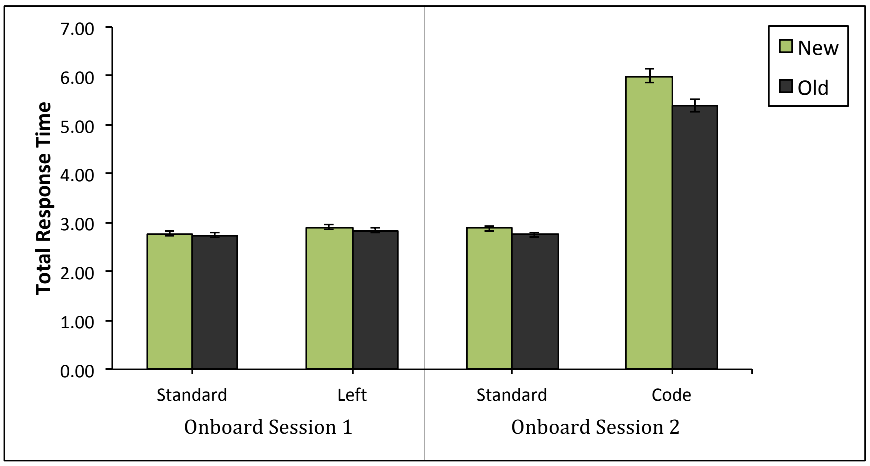


Figure 8. Total response time as a function of task and number type for Onboard Session 1 (left panel) and Onboard Session 2 (right panel).

Onboard Session 1
•Overall significant repetition priming effect. Averaging across task, old numbers (2.787 s) entered more quickly than new numbers (2.839 s), F(1, 43) = 25.532, MSE = .004, p < .001.

•Overall difference in TRT between tasks. Averaging across number type, shorter TRT for standard (2.753 s) than for left-hand task (2.873 s), F(1, 43) = 24.485, MSE = .025, p < .001.

•Larger repetition priming for the left-hand task than for the standard task (marginal), F(1, 43) = 3.605, MSE = .006, p = .064.
Repetition priming marginal for standard task (p = .070), significant for left-hand task (p < .001).

Onboard Session 2
•Overall significant repetition priming effect. Averaging across task, old numbers (4.072 s) were entered more quickly than new numbers (4.439 s), F(1, 43) = 133.616, MSE = .035, p < .001.

•Overall difference in TRT between tasks. Averaging across number type, shorter TRT for standard (2.817 s) than for code task (5.694 s), F(1, 43) = 549.364, MSE = .542, p < .001.

•Larger repetition priming for code task than for standard task, F(1, 43) = 43.222, MSE = .046, p < .001.

•Repetition priming significant for both the standard task ($p < .001$) and the code task ($p < .001$).

V. Summary of Onboard Sessions Results

For the first onboard session, repetition priming (in total response time) was found for both the standard task, which reflects retention, and the left-hand task, which reflects transfer of perceptual processes. Response time component analyses (see *Onboard Response Time Component Analyses*) suggest different underlying mechanisms for the repetition priming effect observed for each task. Specifically, for the standard task, there was positive repetition priming for initiation time, negative repetition priming (new faster than old) for execution time, and no repetition priming for conclusion time, which suggests that the overall positive repetition priming effect for TRT is based primarily on the cognitive aspects of the standard data entry task. For the left-hand task, there was positive repetition priming for initiation time and execution time, and negative repetition priming for conclusion time, which suggests that the positive repetition priming effect for TRT is based on both cognitive and motoric aspects of the task.

For the second onboard session, repetition priming (in total response time) was found for both the standard task, which reflects retention, and the code task, which reflects transfer of motoric processes. For the standard task, there was positive repetition priming for initiation and execution time, and negative repetition priming for conclusion time, which suggests that the overall positive repetition priming effect for TRT is based on both cognitive and motoric aspects of the standard data entry task. For the code task, there was positive repetition priming for initiation time and execution time, and no repetition priming for conclusion time, which suggests that the overall positive repetition priming effect for TRT is based on both cognitive and motoric aspects of the task.

Analysis of Retention Across All Sessions for each Task

I. Standard Data Entry Task

Average total response time (top panel) and accuracy (bottom panel) showed little forgetting across all sessions. For total response time, there was an effect of session, $F(5, 205) = 14.263$, $MSE = .027$, $p < .001$. Subjects were, on average, faster during the first training session than during the pre-test, and faster during the second training session than during the first. From the second training session to the first onboard session there was a slight increase in total response time for crew subjects, and for all subject groups there was an increase in total response time from the second onboard session to the post-test, although these increases are small. All subject types were faster during the post-test than during the pre-test.

For all subject types, accuracy was relatively stable across sessions; the effect of session was not significant, $F(5, 205) = 1.386$, $MSE = .003$, $p = .231$.

In sum, all subject types generally maintained or improved their performance across sessions despite the extended retention intervals between sessions.

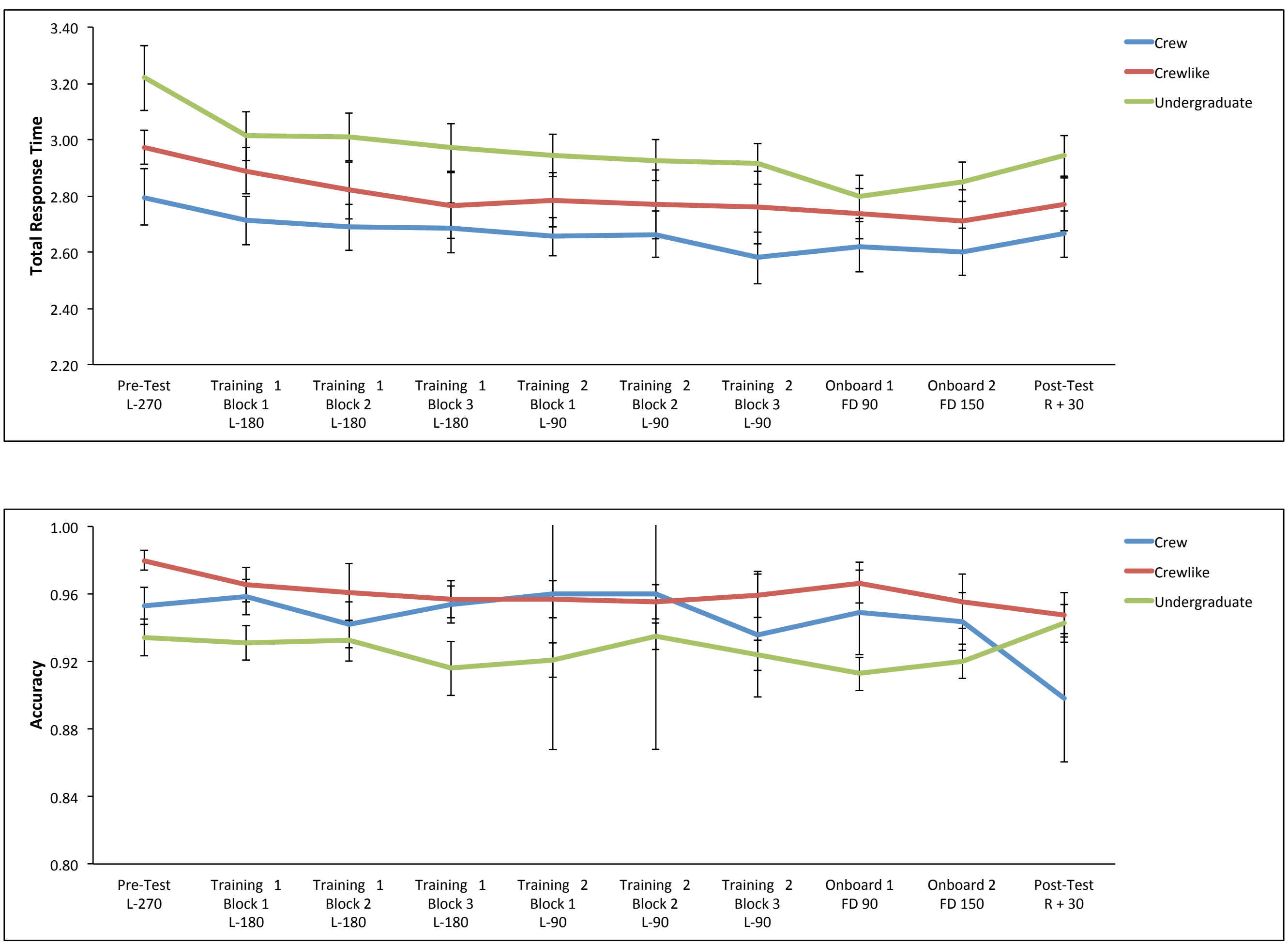


Figure 9. Total response time (top panel) and accuracy (bottom panel) as a function of session and subject type for the standard data entry task.

II. Left-Hand Task
Statistical analyses of the left-hand task were not conducted because subjects only performed this task during the pre-test (L-270), first onboard session (FD 90), and post-test (R+30). However, examination of the average total response time (top panel) and accuracy (bottom panel) reveals very little forgetting even with the prolonged retention intervals. For total response time, subjects were actually faster during the first onboard session than during the pre-test. Although there appears to be some forgetting from the first onboard session to the post-test, all subject groups were faster during the post-test than during the pre-test.

For accuracy, there were minor decrements for all subject types from pre-test to the first onboard session, but an increase from the first onboard session to the post-test.

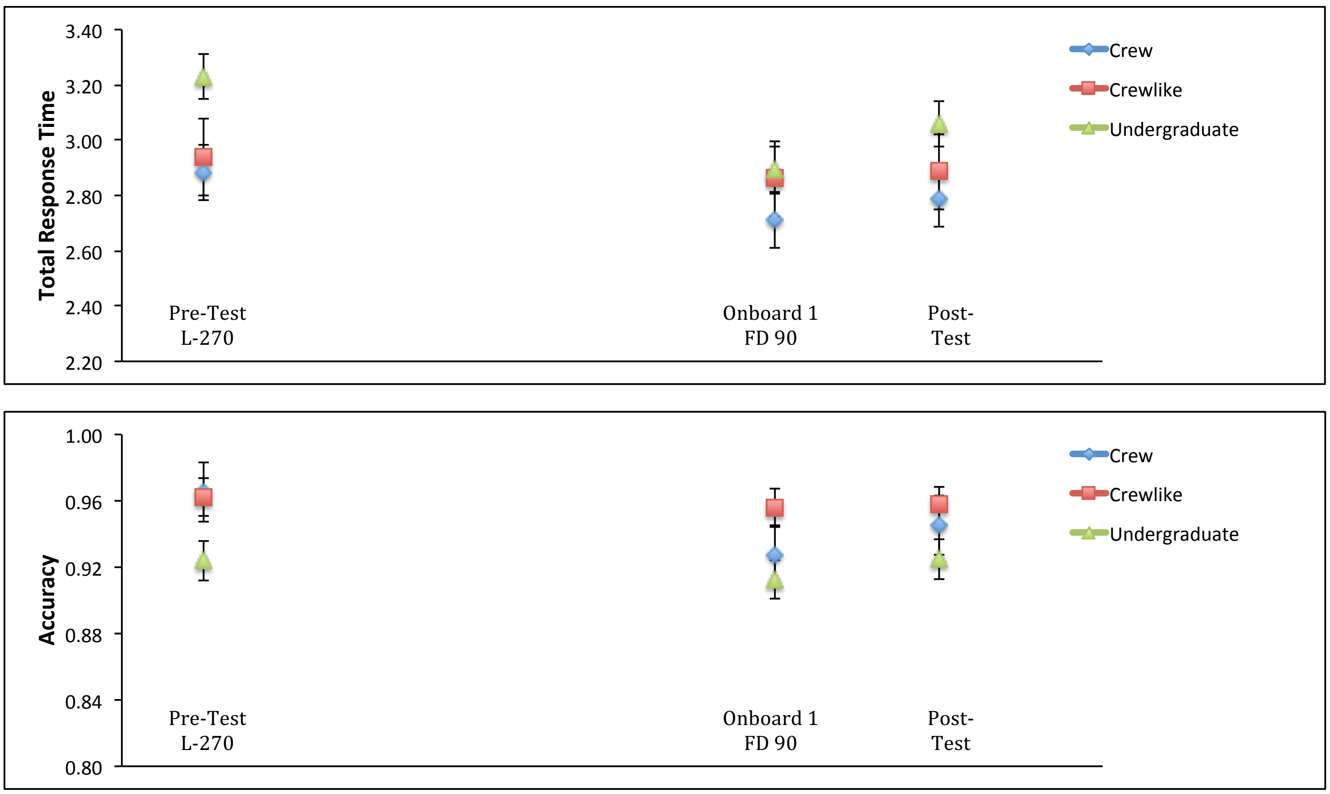


Figure 10. Total response time (top panel) and accuracy (bottom panel) as a function of session and subject type for the left-hand task.

III. Code Task

Statistical analyses of the code task were not conducted because subjects only performed this task during the pre-test (L-270), second onboard session (FD 150), and post-test (R+30). As for the left-hand task, examination of the average total response time (top panel) and accuracy (bottom panel) reveals very little forgetting even with the prolonged retention intervals. For total response time, subjects were again faster during the second onboard session and the post-test than during the pre-test.

For accuracy, crew and crew-like subjects maintained level accuracy across sessions, whereas for undergraduate subjects accuracy tended to decrease.

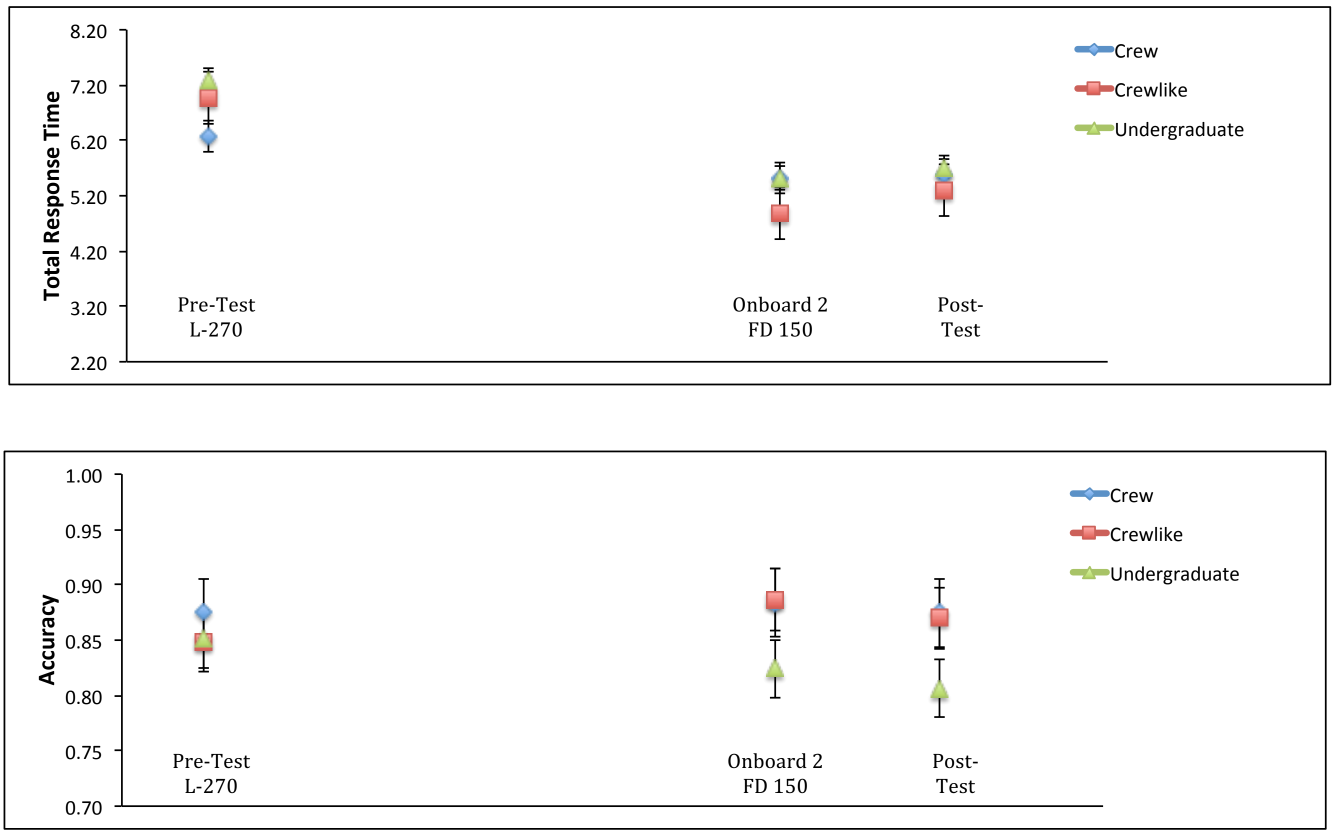
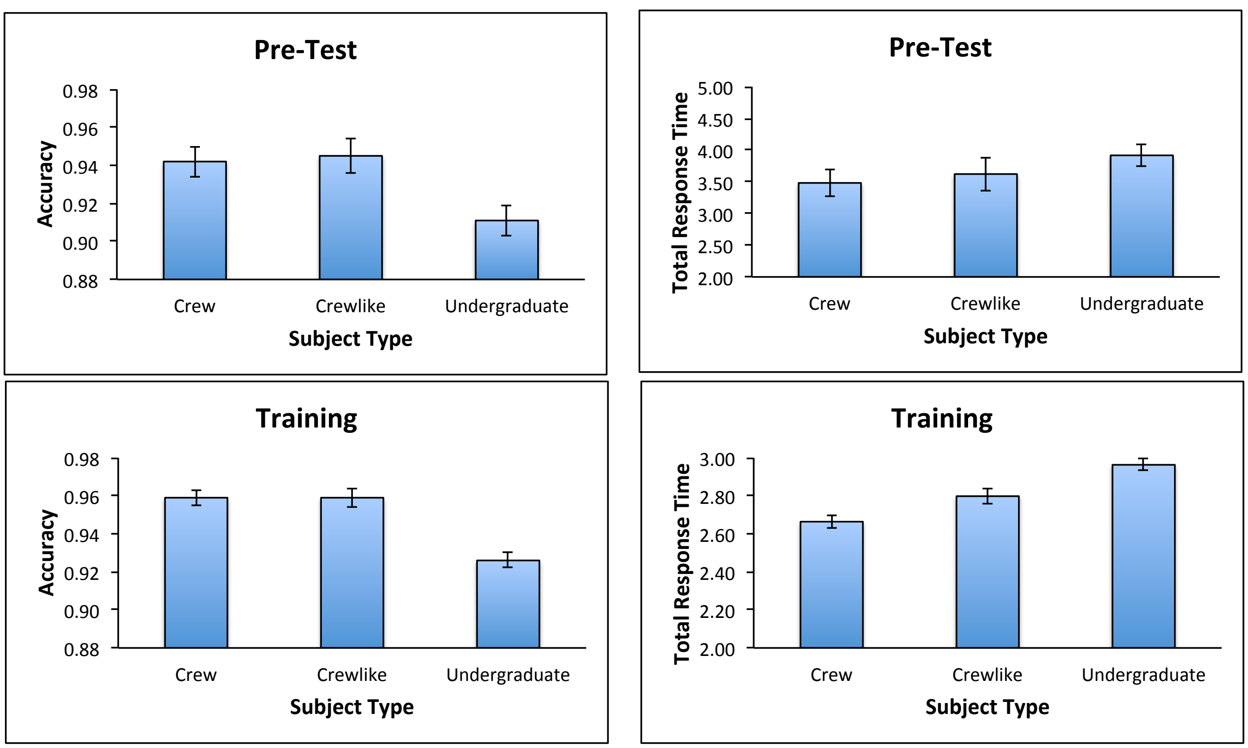


Figure 11. Total response time (top panel) and accuracy (bottom panel) as a function of session and subject type for the code task.

Summary of Subject Type (Crew, Crew-like, Undergraduate) Effects

In this section, a summary is provided of the differences observed between crew, crew-like, and undergraduate subjects. These differences often manifest in *subject-type effects*, in which there are differences in the overall levels of performance between crew, crew-like, and undergraduate subjects for a given performance measure. However, these differences may also be evident in what is termed here *magnitude effects*, in which a variable might have a similar impact on performance for all subject types, but a stronger effect for one subject type than another. Differences may also be evident in *direction effects*, in which a variable has opposite impacts on performance for different subject types. Magnitude effects are important if trying to predict with precision the impact of a variable (e.g., percent performance will increase), whereas direction effects are important if trying to predict more generally the type of impact a variable will have (e.g., increase or decrease performance).

Subject-type effects were found throughout most phases of the study. Generally speaking, crew and crew-like subjects performed better (high accuracy, quicker response time) than undergraduate subjects, but crew and crew-like subjects did not differ (significantly) from each other. These subject-type effects were found even during the pre-test, which suggests that they aren't based entirely on differential impacts of training.



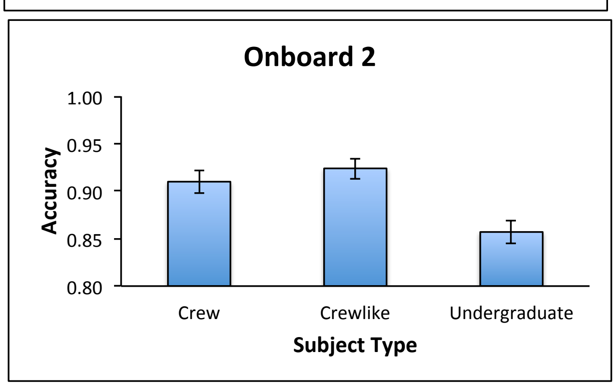


Figure 12. Total response time (right panels) and accuracy (left panels) as a function of session and subject type, averaging across tasks.

From pre-test to post-test there was a magnitude effect in the amount of improvement demonstrated for total response time. For all subject types (crew, crew-like, undergraduate), total response time decreased from pre-test to post-test. However, the decrease in total response time for the code task was stronger for undergraduate subjects than for crew and crew-like subjects. This greater improvement might be due to the fact that undergraduate subjects were slower initially.

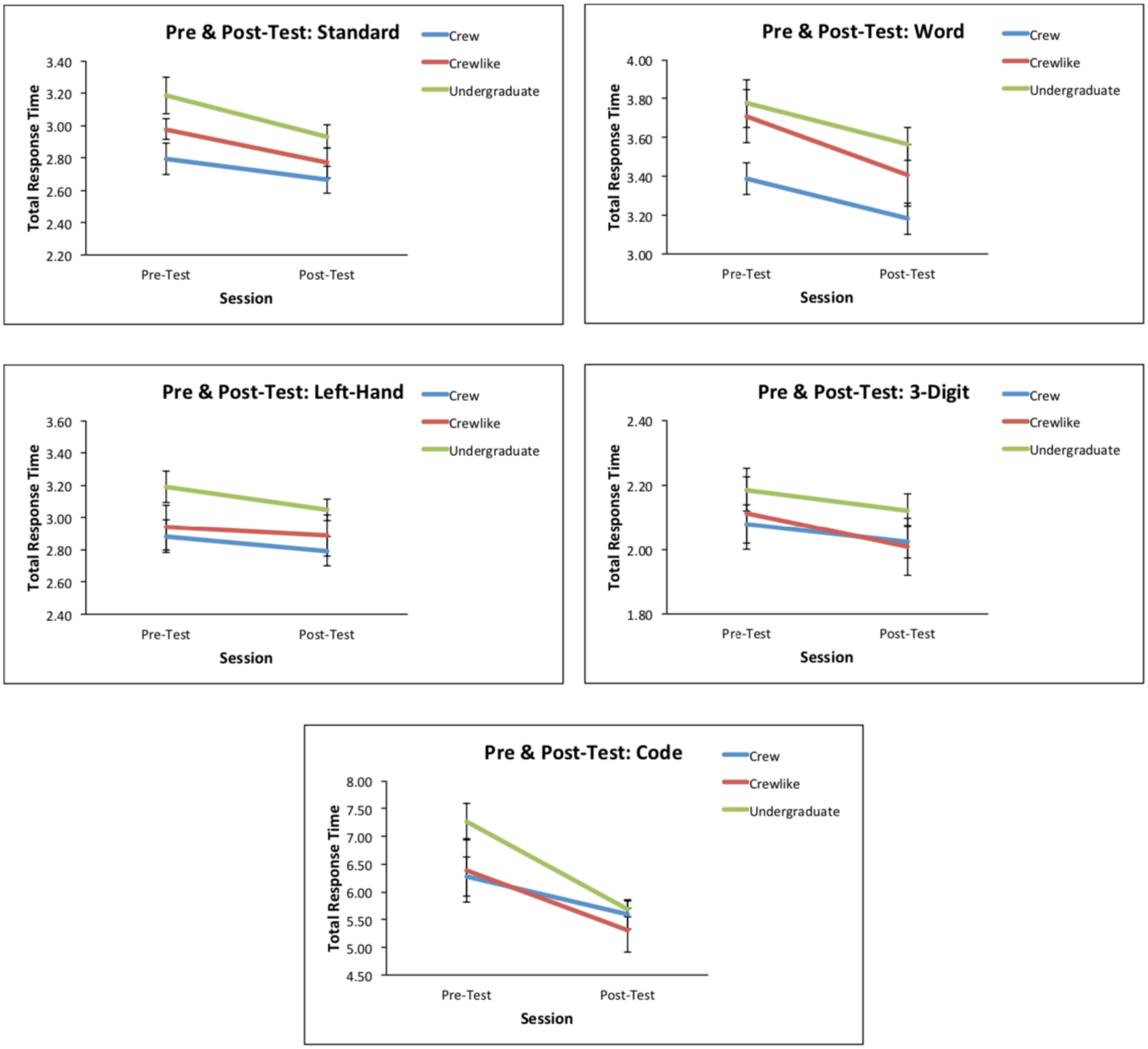


Figure 13. Total response time as a function of test (pre-test, post-test), subject type, and task.

During the first onboard session, there was an interesting direction effect for accuracy. For the standard task, accuracy was higher for old stimuli than for new stimuli for crew and crew-like subjects (p = .06); the opposite pattern was observed for undergraduate subjects (p = .03). The advantage for old stimuli vs. new is a form of repetition priming. For the left-hand task, accuracy was higher for new stimuli than for old stimuli for crew and crew-like subjects (p = .02); the opposite pattern was observed for undergraduate subjects, although not significant (p = .29). Although not definitive, it is speculated that crew and crew-like subjects form stronger associations between percepts and motor patterns, so that performance would be facilitated when there is a match between the percept (4-digit number) and the motor pattern (as in the standard task), but disrupted when the previously practice percept-motor pattern is changed (as in the left-hand task, which necessitates a change in motor pattern).

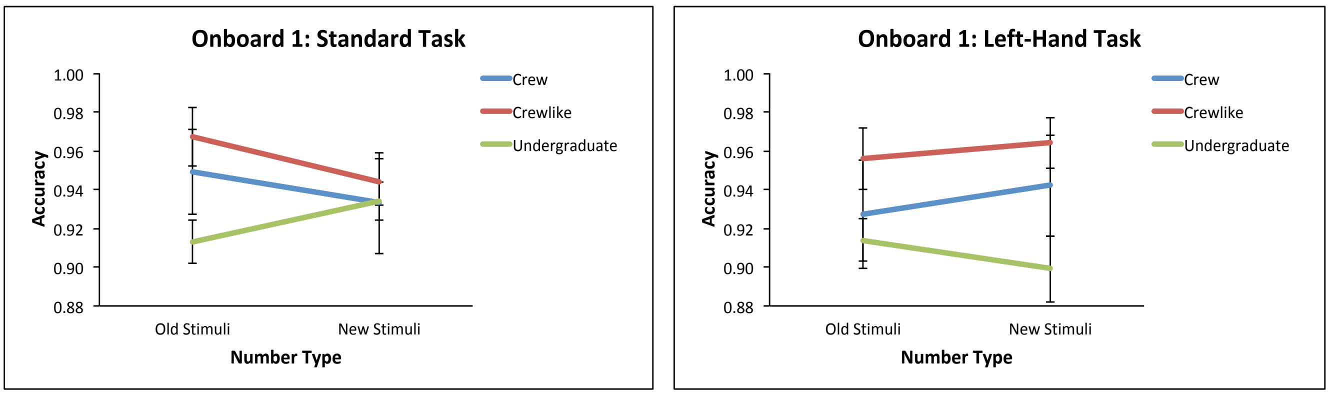


Figure 14. Accuracy as a function of number type and subject type for the standard task (left panel) and the left-hand task (right panel) for Onboard Session 1.

During the second onboard session, there was a magnitude effect for accuracy. More specifically, accuracy was similar for all subject types for the standard task, but significantly higher for crew and crew-like than for undergraduate subjects for the code task. This pattern suggests that crew and crew-like subjects are better able to handle more cognitively demanding tasks (such as the code task).

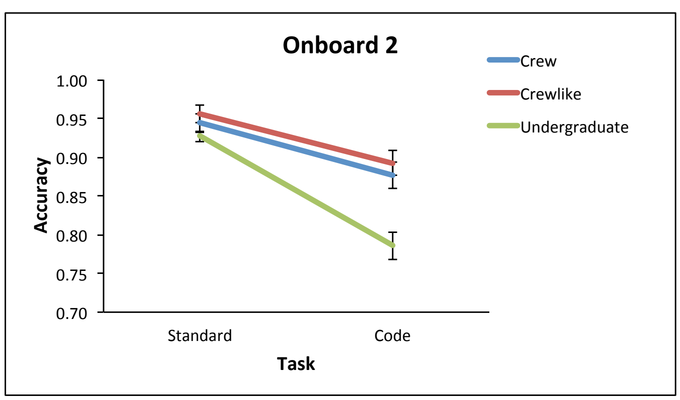


Figure 15. Accuracy as a function of task and subject type for Onboard Session 2.

Pre-Test/Post-Test Response Time Component Analyses

I. Initiation Time

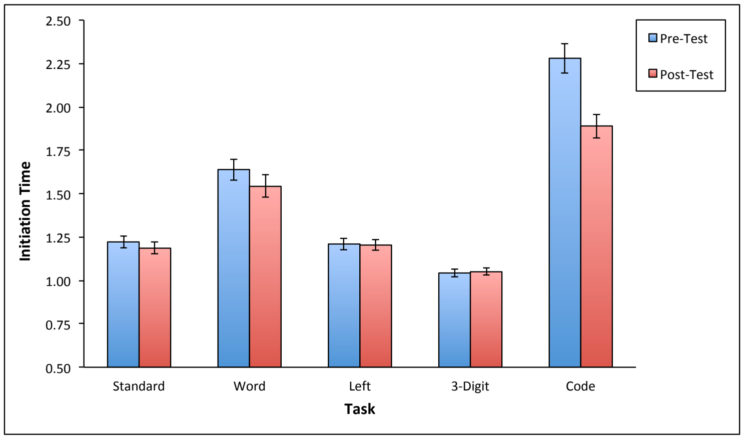


Figure 16. Initiation time as a function of task and test (pre-test, post-test).

•Initiation time decreased from pre-test to post-test, averaging across task, $F(1, 42) = 12.138$, $MSE = .075$, $p = .001$.
•Initiation time slower for code than for other tasks, averaging across session, $F(4, 168) = 109.027$, $MSE = .109$, $p < .001$.
•Decrease in initiation time from pre-test to post-test greater for code task than for other tasks, $F(4, 168) = 11.047$, $MSE = .031$, $p < .001$.
•No main effects (or interactions involving) subject type.
•An analysis of pre-test only did not show an effect of subject type.

II. Execution Time

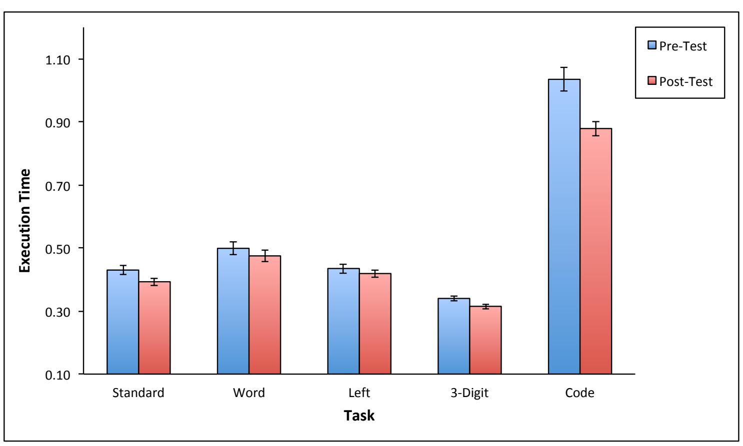


Figure 17. Execution time as a function of task and test (pre-test, post-test).

•Execution time decreased from pre-test to post-test, averaging across task, $F(1, 42) = 16.203$, $MSE = .011$, $p < .001$.
•Execution time slower for code than for other tasks, averaging across session, $F(4, 168) = 279.239$, $MSE = .015$, $p < .001$.
•Decrease in execution time from pre-test to post-test greater for code task than for other tasks, $F(4, 168) = 16.442$, $MSE = .003$, $p < .001$.
•No main effects (or interactions involving) subject type.

•An analysis of pre-test only showed a marginally significant effect of subject type, F(2, 42) = 2.661, MSE = .050, p = .082, with execution time faster for crew than for undergraduate subjects.
Crew-Crew-like, p = .341, Crew-Undergraduate, p = .028, Crew-like-Undergraduate, p = .361

Pre-Test/Post-Test Response Time Component Analyses, cont.

III. Conclusion Time

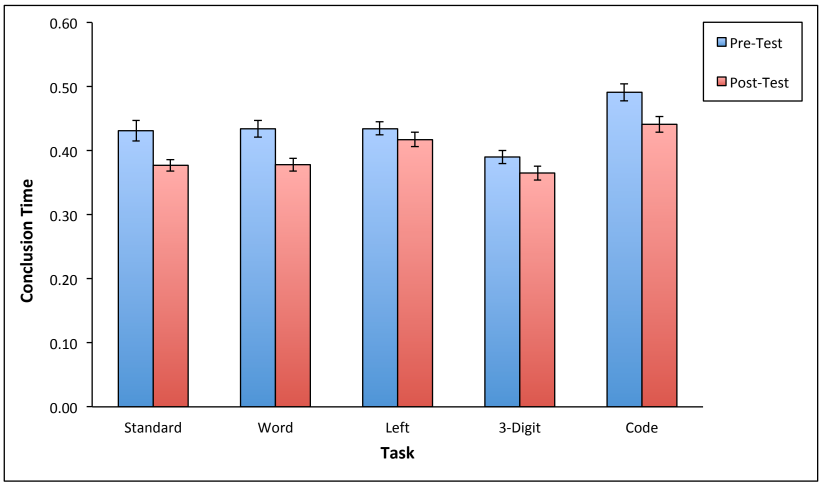


Figure 18. Conclusion time as a function of task and test (pre-test, post-test).

•Conclusion time decreased from pre-test to post-test, averaging across task, F(1, 42) = 12.615, MSE = .010, p = .001.
•Conclusion time slower for code than for other tasks, averaging across session, F(4, 168) =30.994, MSE = .003, p < .001.
•Decrease in conclusion time from pre-test to post-test greater for standard, word, and code tasks than for left and 3-digit tasks, F(4, 168) = 4.590, MSE = .001, p = .002.
•An analysis of pre-test only did not show an effect of subject type.

Training Response Time Component Analyses

I. Initiation Time

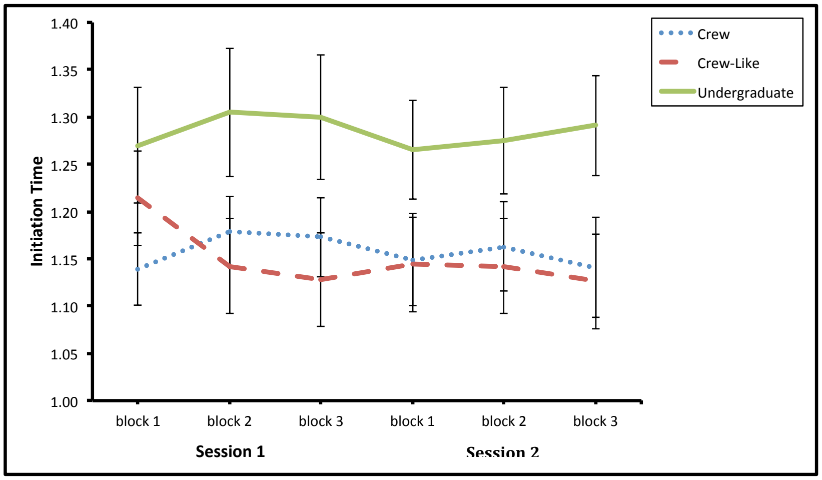


Figure 19. Initiation time as a function of training session, block, and subject type.

•No significant results for initiation time.

II. Execution Time

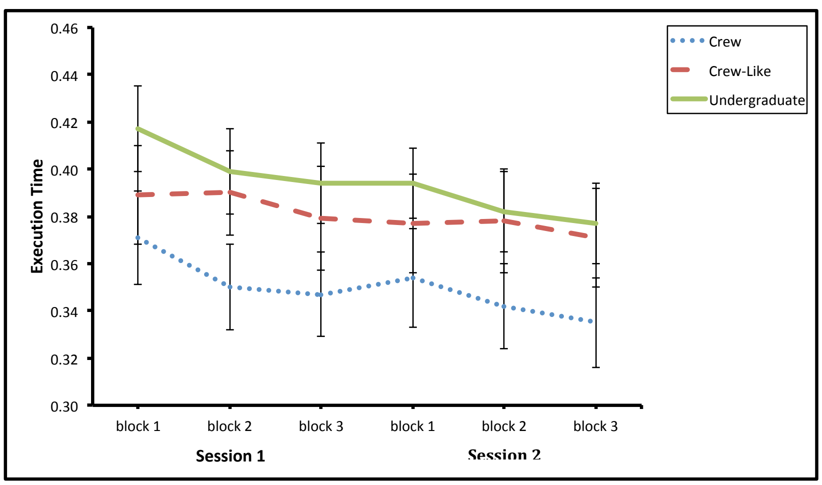


Figure 20. Execution time as a function of training session, block, and subject type.

•Execution time decreased from the first training session to the second training session, averaging across blocks.
$F(1, 41) = 16.060$, $MSE = .002$, $p = .018$

•Execution time decreased across blocks, averaging across training session.
$F(2, 82) = 6.473$, $MSE = .001$, $p = .003$

Training Response Time Component Analyses, cont.

III. Conclusion Time

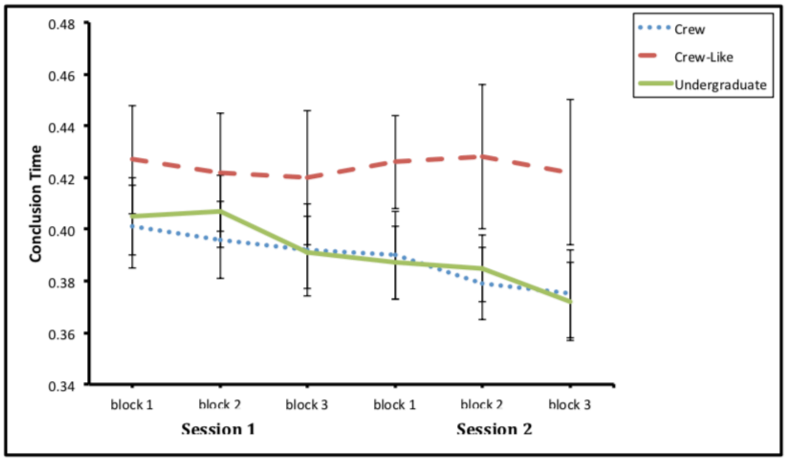


Figure 21. Conclusion time as a function of training session, block, and subject type.

•Conclusion time decreased across blocks, averaging across training session, $F(2, 82) = 3.168$, $MSE = .001$, $p = .047$.

Onboard Response Time Component Analyses (Initiation Time)

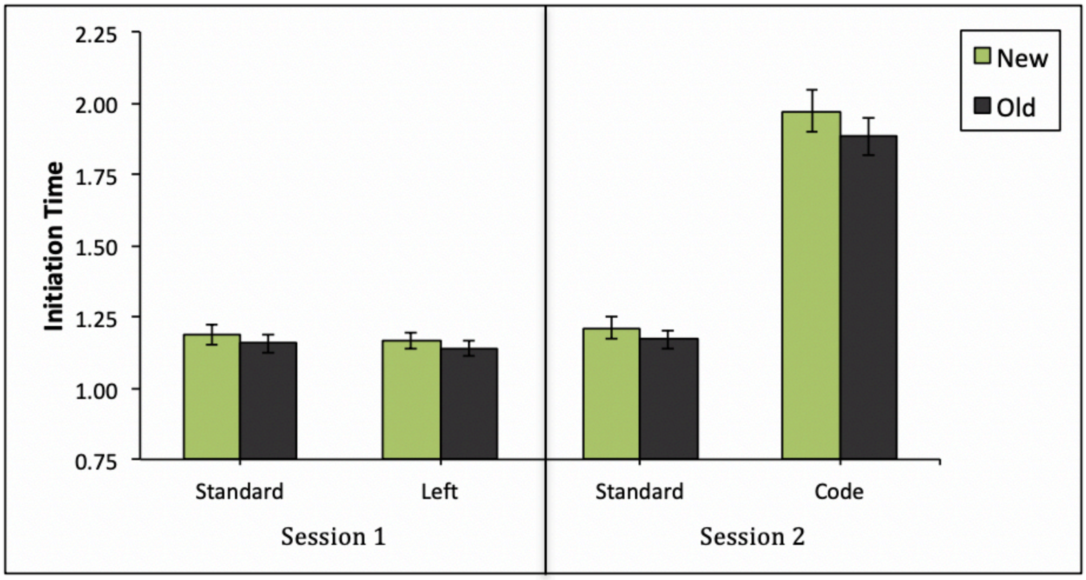


Figure 22. Initiation time as a function of onboard session, task, and number type.

<u>Onboard Session 1</u>

•Overall repetition priming effect; initiation time faster for old numbers than for new numbers, averaging across task, $F(1, 43) = 19.723$, $MSE = .002$, $p < .001$.

-Significant repetition priming for standard task ($p < .001$).

-Significant repetition priming for left-hand task ($p = .010$).

<u>Onboard Session 2</u>

•Overall repetition priming effect; initiation time faster for old numbers than for new numbers, averaging across task, $F(1, 43) = 8.991$, $MSE = .018$, $p = .005$.

-Significant repetition priming for standard task ($p = .002$).

-Significant repetition priming for code task ($p = .010$).

•Initiation time faster for standard task than for code task, averaging across number type, $F(1, 43) = 127.941$, $MSE = .158$, $p < .001$.

Onboard Response Time Component Analyses (Execution Time)

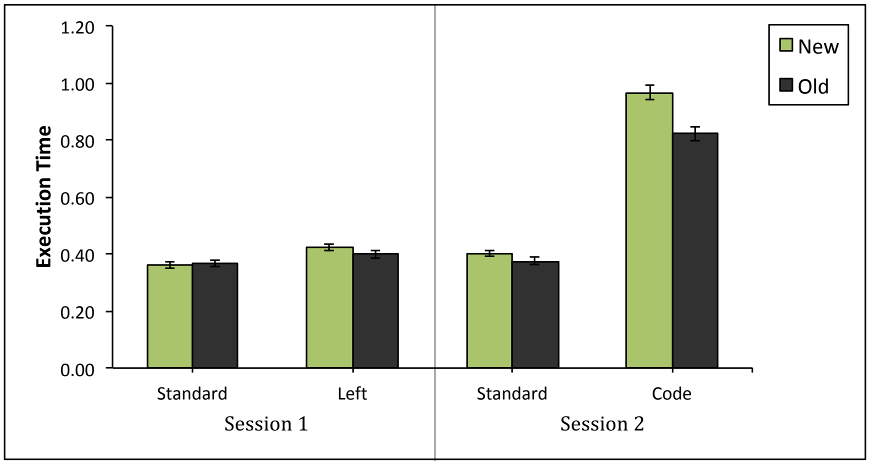


Figure 23. Execution time as a function of onboard session, task, and number type.

<u>Onboard Session 1</u>

•Overall repetition priming effect; execution time faster for old numbers than for new numbers, averaging across task.
F(1, 43) = 10.405, MSE = .0002, p = .002

•Execution time faster for standard task than for left-hand task, averaging across number type.
F(1, 43) = 34.445, MSE = .002, p < .001

•Larger repetition priming for left-hand task than for standard task.
F(1, 43) = 44.700, MSE = .0002, p < .001

-Significant **<u>negative</u>** repetition priming for standard task (p = .020)
-Significant **<u>positive</u>** repetition priming for left-hand task (p < .001)
-For the left-hand task, Crew subjects showed a smaller repetition priming effect (old = .377s, new = .388 s) than Crew-like subjects (old = .418 s, new = .453 s) and undergraduate subjects (old = .400 s, new = .429 s)

<u>Onboard Session 2</u>

•Overall repetition priming effect; execution time faster for old numbers than for new numbers, averaging across task, F(1, 43) = 134.028, MSE = .002, p < .001.

•Execution time faster for standard task than for code task, averaging across number type, F(1, 43) = 442.824, MSE = .020, p < .001.

•Larger repetition priming for code task than for standard task, F(1, 43) = 68.188, MSE = .002, p < .001

-Significant repetition priming for standard task (p < .001)
-Significant repetition priming for code task (p < .001)

Onboard Response Time Component Analyses (Conclusion Time)

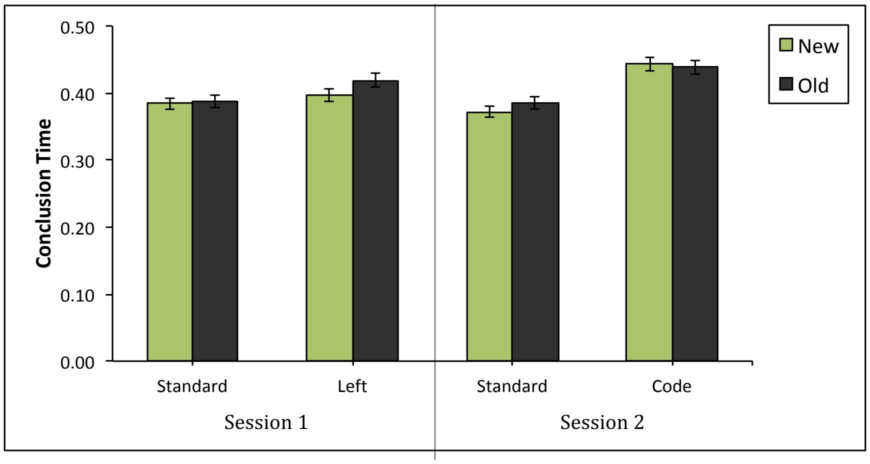


Figure 24. Conclusion time as a function of onboard session, task, and number type.

<u>Onboard Session 1</u>

•Overall **<u>negative</u>** repetition priming effect; conclusion time faster for new numbers than for old numbers, averaging across task, $F(1, 43) = 40.940$, $MSE = .0001$, $p < .001$.

•Conclusion time faster for standard task than for left-hand task, averaging across number type, $F(1, 43) = 10.257$, $MSE = .003$, $p = .003$.

•Larger **<u>negative</u>** repetition priming for left-hand task than for standard task, $F(1, 43) = 26.212$, $MSE = .0001$, $p < .001$.

-No repetition priming for standard task ($p = .189$)

-Significant **<u>negative</u>** repetition priming for left-hand task ($p < .001$)

<u>Onboard Session 2</u>

•Overall **<u>negative</u>** repetition priming effect; conclusion time faster for new numbers than for old numbers, averaging across task, $F(1, 43) = 4.789$, $MSE = .0002$, $p = .034$.

•Conclusion time faster for standard task than for code task, averaging across number type, $F(1, 43) = 56.820$, $MSE = .002$, $p < .001$.

•Larger **<u>negative</u>** repetition priming for standard task than for code task, $F(1, 43) = 68.188$, $MSE = .002$, $p < .001$.

-Significant **<u>negative</u>** repetition priming for standard task ($p < .001$)

-No repetition priming for code task ($p = .368$)

Performance on the Mapping Task: Pre-Test and Post-Test

I. Method

•Twenty-six undergraduate students at the University of Colorado, 20 subjects from NASA (11 crew and 9 crew-like). The subjects were the same as those who participated in the study on data entry.

•In the mapping task, subjects learn associations between eight astronauts (Alpha, Bravo, Charlie, etc.) and four locations (North, South, East, West). The locations associated with a given name continuously changes, thus subjects must update the location of each astronaut.

•On study trials, subjects are presented with a name-location association (e.g., Alpha North), and they respond by clicking the appropriate location on a map on the right-hand side of a display screen labeled *Spacecraft*. The name is presented either in black font or green font (see Figure 25). Regardless of font color, subjects respond in the same way by clicking the appropriate location.

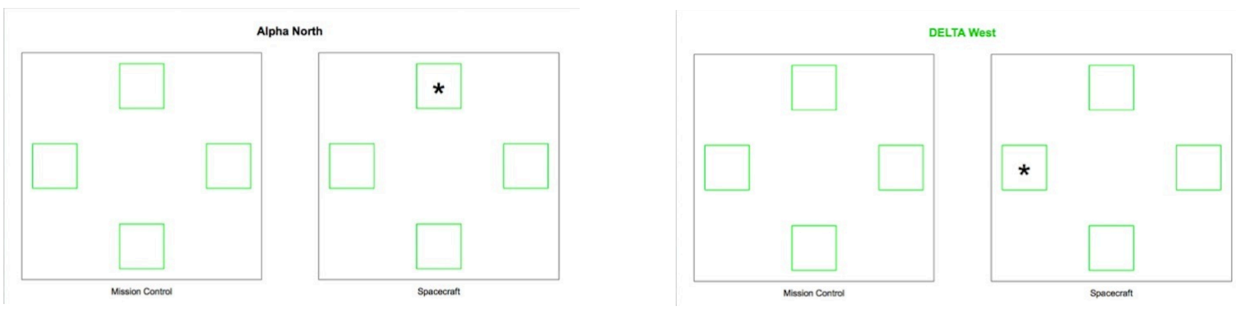


Figure 25. Sample study trial for a retrospective test trial (left panel) and a prospective test trial (right panel) for the mapping task.

•On test trials, subjects are instead presented with an astronaut name in blue font with no associated location (see Figure 26). On these trials, subjects respond with the astronaut's most recent location by clicking at the appropriate location on one of two maps. On *prospective memory* test trials (indicated by the name studied in green font), subjects respond using the left-hand map labeled *Mission Control*, whereas on *retrospective memory* test trials (indicated by the name studied in black font), subject respond using the right-hand map labeled *Spacecraft*.

•Test trials occurred after a short retention interval (study trial occurred two trials before the test trial, 2-back) or after a long retention interval (study trial occurred eight trials before the test trials, 8-back).

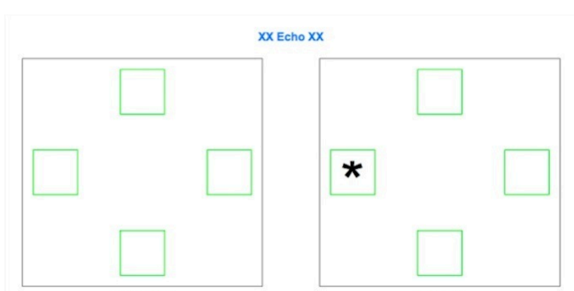


Figure 26. Sample default test trial for the mapping task.

•Six experimental sessions distributed over 480 days:

-Session 1: *Pretest*. Subjects performed two blocks of the mapping task. Both blocks included 37 trials; however, during the second block subjects also performed a cognitively demanding concurrent secondary task (counting backwards by 3's from 100).

-Session 2: *Training Session 1*. Subjects competed two blocks of trials of the mapping task with no concurrent secondary task. Each block consisted of 137 trials, and the same order of stimuli was used for both blocks.

-Session 3: *Training Session 2*. Training Session 2 was identical to Training Session 1.

-Session 4: *Onboard Session 1*. Subjects completed 137 trials of the mapping task with no concurrent secondary task. The stimuli (ordering of astronaut names and locations) were different than during the training sessions, thus involving some degree of transfer.

-Session 5: *Onboard Session 2*. Subjects completed 137 trials of the mapping task, this time with the concurrent secondary task. The stimuli (ordering of astronaut names and locations) were different than during the training sessions and Onboard Session 1.

-Session 6: *Post-test*. The post-test was identical to the pre-test.

•The dependent measures included location accuracy (correctly remembering the associated location for the astronaut, regardless of which map was used for responding) and side accuracy (correctly remembering to use the *Spacecraft* map for retrospective trials and the *Mission Control* map for prospective trials, regardless of whether or not the selected location was correct).

II. Pre-Test and Post-Test Analyses

•Comparison between pre-test and post-test, block (block 1, block 2), trial type (retrospective, prospective), and within-session retention interval (WSRI; 2-back, 8-back). Subject type (crew, crew-like, undergraduate) also included as a factor.

•2 dependent measures: side accuracy and location accuracy.

III. Pre-Test & Post-Test Results: Location Accuracy

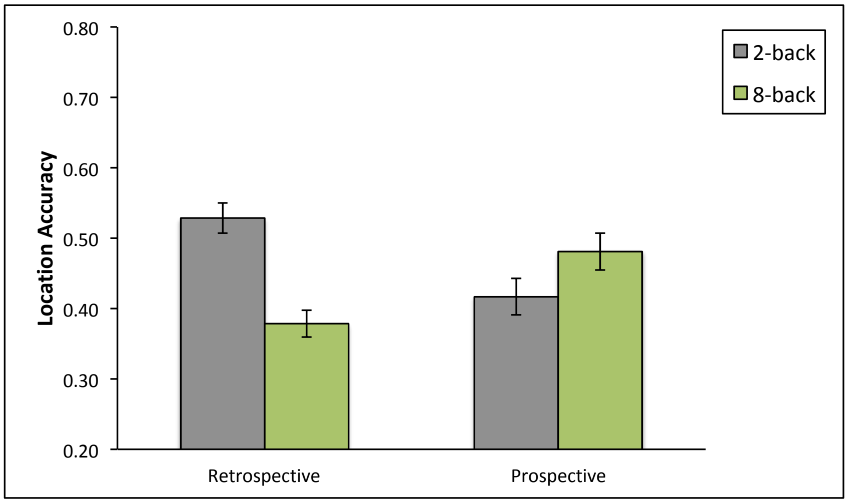


Figure 27. Location accuracy as a function of within-session retention interval and trial type.

•On average across test (pre-test, post-test) and trial type, location accuracy was higher for 2-back (.472) than for 8-back trials (.429), $F(1, 42) = 4.631$, $MSE = .116$, $p = .037$. This result simply reflects forgetting that occurs with increasing time.

•For retrospective memory trials, location accuracy was higher for 2-back than for 8-back, indicating forgetting; for prospective memory trials, the opposite pattern was observed, $F(1, 42) = 24.745$, $MSE = .066$, $p < .001$. This result indicates that when subjects must change their response location from the *Spacecraft* map to the *Mission Control* map, associated information (location) is preserved in memory and less susceptible to forgetting. In previous studies this effect has been called the *protective function of distinctive responding*.

•There were overall differences between subject types (crew = .555, crew-like = .500, undergraduate = .391), $F(2, 42) = 14.216$, $MSE = .131$, $p < .001$.
Crew-Crew-like, $p = .194$
Crew-Undergraduate, $p < .001$
Crew-like-Undergraduate, $p = .005$

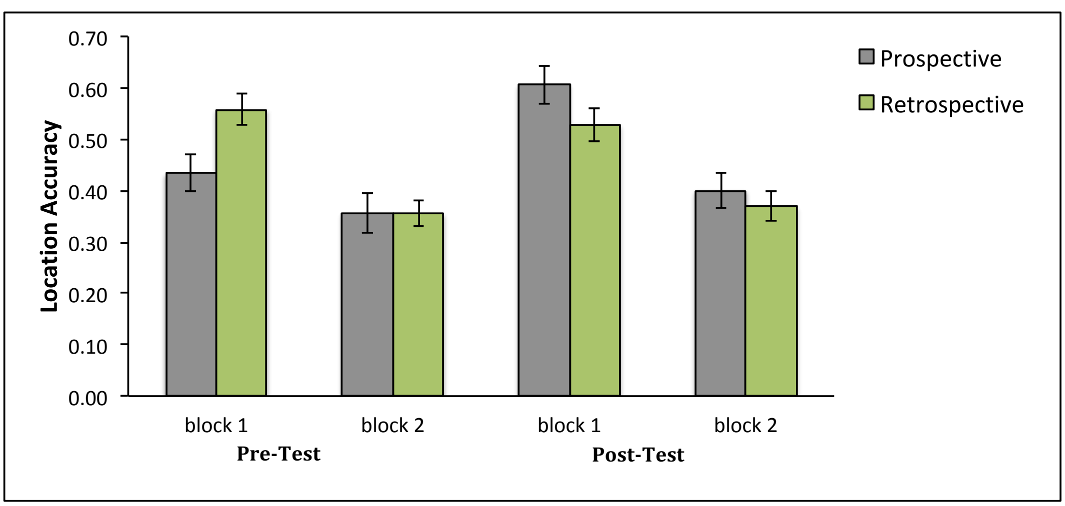


Figure 28. Location accuracy as a function of test, block, and trial type.

•During the pre-test, location accuracy was higher for retrospective memory trials than for prospective memory trials during the first block; there was no difference in location accuracy between prospective memory and retrospective memory trials during the second block (which included the concurrent secondary task). During the post-test, location accuracy was higher for prospective memory trials than for retrospective memory trials for both blocks, $F(1, 42) = 4.710$, $MS = .066$, $p = .036$.

•Averaging across trial type and block, location accuracy increased from pre-test (.426) to post-test (.476), $F(1, 42) = 5.481$, $MS = .097$, $p = .024$, reflecting practice effects. However, this increase in accuracy occurred only for prospective memory trials (pre-test = .394; post-test =.503); on retrospective memory trials location accuracy decreased from pre-test to post-test (pre-test = .457; post-test = .449), $F(1, 42) = 5.537$, $MS = .073$, $p = .023$.

•Averaging across pre-and post-test and trial type, location accuracy decreased from block 1 (.531) to block 2 (.370), presumably due to the addition of the concurrent secondary task, $F(1, 42) = 28.304$, $MS = .128$, $p < .001$.

IV. Pre-Test & Post-Test Results: Side Accuracy

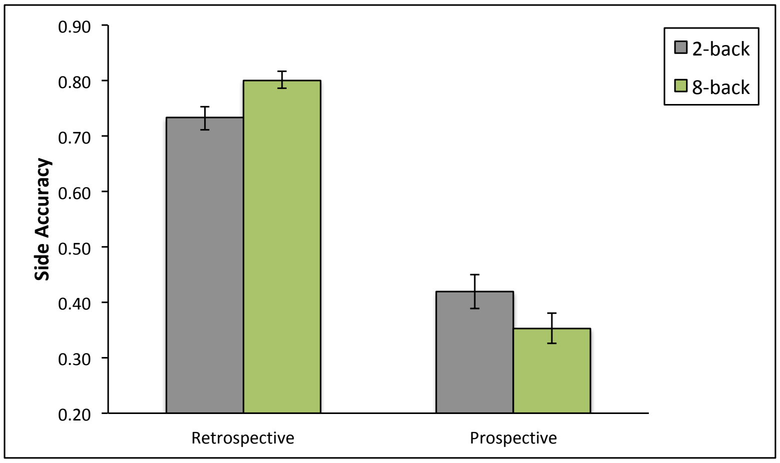


Figure 29. Side accuracy as a function of within-session retention interval and trial type.

•Averaging across WSRI, side accuracy was lower for prospective memory trials (.386) than for retrospective memory trials (.767), F(1, 42) = 117.128, MS = .185, p < .001.

•For retrospective memory trials, side accuracy higher for 8-back than for 2-back; the opposite pattern was observed for prospective memory trials, F(1, 42) = 9.116, MS = .082, p = .004. This pattern is the opposite than that found for location accuracy.

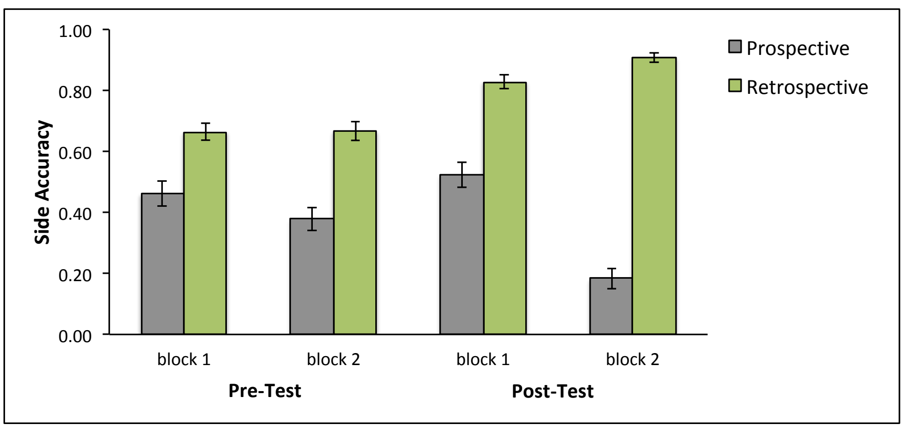


Figure 30. Side accuracy as a function of within-session retention interval and trial type.

•During both the pre-test and post-test, there was an advantage in side accuracy for retrospective memory trials relative to prospective memory trials, which increased from the first block to the second block. The increasing advantage for retrospective memory trials was stronger during the post-test than during the pre-test, F(1, 42) = 11.654, MS = .086, p = .001.

•Averaging across trial type and block, side accuracy increased from pre-test (.542) to post-test (.610), F(1, 42) = 8.497, MS = .063, p = .006.

• Averaging across pre-and post-test and trial type, side accuracy decreased from block 1 (.619) to block 2 (.534), which was more difficult due to the concurrent secondary task, F(1, 42) = 16.054, MS = .061, p < .001.

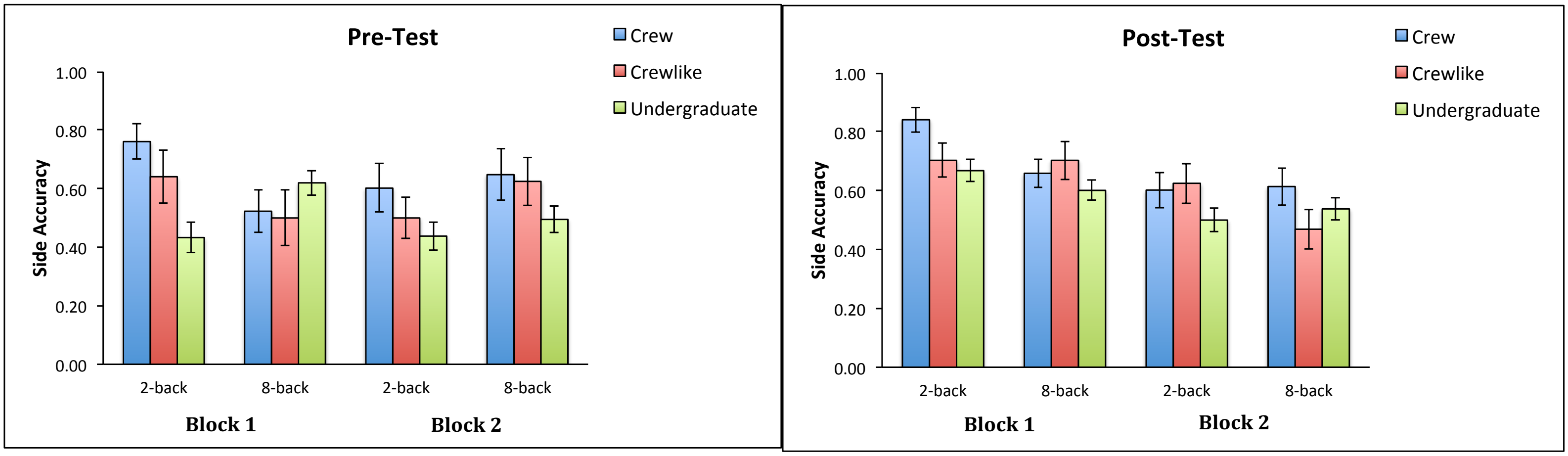


Figure 31. Side accuracy as a function of test, block, within-session retention interval, an subject type.

•During the pre-test, for the first block, side accuracy was greater for 2-back than for 8-back for crew and crew-like subjects, but the opposite pattern was observed for undergraduate subjects. For the second block, side accuracy was greater for 8-back than for 2-back for all subject types.

During the post-test, for the first block, side accuracy was greater for 2-back than for 8-back for crew and undergraduate subjects. For the second block, side accuracy was greater for 2-back than for 8-back for crew-like subjects only, $F(2, 42) = 4.370$, $MS = .083$, $p = .019$

•Averaging across session, block, and WSRI, side accuracy was highest for Crew (.656), lowest for undergraduate subjects (.537), and intermediate for Crew-like subjects (.597).
$F(2, 42) = 11.228$, $MS = .081$, $p < .001$
Crew-Crew-like, $p = .075$
Crew-Undergraduate, $p < .001$
Crew-like-Undergraduate, $p = .047$

V. Summary of Pre- and Post-Test Results

The results replicate the previous finding that retention interval has different impacts depending on trial type (special distinctive location for prospective memory trials, default location for retrospective memory trials) and the dependent measure examined (location accuracy vs. side accuracy). Increasing within-session retention interval decreases location accuracy on retrospective memory trials (forgetting) but has little effect on prospective memory trials (protective function of distinctive responding). However, the opposite pattern was found for side accuracy. In general, for side accuracy, subjects were more likely to select the default response location rather than the distinctive.

Examination of location and side accuracy across blocks during the pre-test as well as during the post-test suggests that the advantage in location accuracy for prospective memory trials as well as the advantage in side accuracy for retrospective memory trials is strengthened with practice and time. For location accuracy, during the first block of pre-test there was an advantage for retrospective memory trials relative to prospective memory trials, but this advantage was eliminated during the second block of pre-test, and reversed (and increased) during the two blocks of the post-test. For side accuracy, the advantage for retrospective memory trials increased across the four blocks of pre-test and post-test.

There were overall differences between crew, crew-like, and undergraduate subjects for both location and side accuracy, with both sets of NASA subjects performing better than undergraduate subjects. In addition to performing better overall, crew and crew-like subjects show a somewhat different performance trajectory for side accuracy than did undergraduate students. For crew, during the first block of pre-test, side accuracy was better for 2-back trials than for 8-back trials, which perhaps suggests that they had already recognized that there was a default location. By the end of the second block of pre-test, side accuracy was similar for 2-back and 8-back trials, which suggests crew subjects had adopted a strategy to better remember the appropriate response location. A similar pattern was found for crew-like subjects; however, during the second block of the post-test they again showed an advantage for 2-back trials. For undergraduate subjects in the first block, there was an advantage for 8-back trials rather than 2-back trials, and over the next block of the pre-test and both blocks of post-test, side accuracy was similar for 2-back and 8-back trials.

Performance on the Mapping Task: Training Sessions

<u>I. Method</u>

-For Training Session 1, subjects competed two blocks of trials of the mapping task with no concurrent secondary task. Each block consisted of 137 trials, and the same order of stimuli was used for both blocks. Training Session 2 was identical to Training Session 1.

<u>II. Training Sessions Analyses</u>

•Comparison between training sessions (session 1, session 2), block within each training session (block 1, block 2), trial type (retrospective, prospective), and within-session retention interval (2-back, 8-back). Subject type (crew, crew-like, undergraduate) also included as a factor.

•2 dependent measures: side accuracy and location accuracy.

<u>III. Training Sessions Results: Location Accuracy</u>

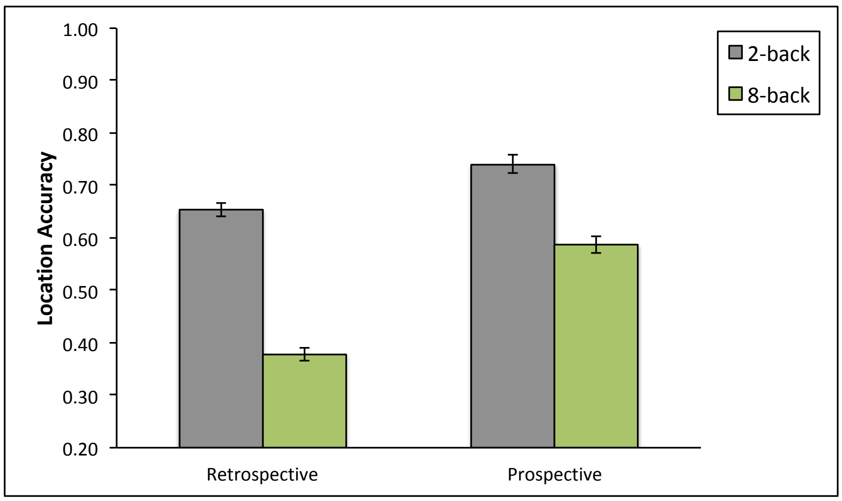


Figure 32. Location accuracy as a function of within-session retention interval and trial type.

•Averaging across WSRI, location accuracy was higher for prospective memory trials (.664) than for retrospective memory trials (.517), $F(1, 41) = 35.878$, $MS = .082$, $p < .001$.

•Averaging across trial type, location accuracy was higher for 2-back (.698) than for 8-back (.483), which reflects forgetting over the retention interval, $F(1, 41) = 122.065$, $MS = .058$, $p < .001$.

•The effect of WSRI was larger for retrospective memory trials than for prospective memory trials, which illustrates the protective function of distinctive responding, $F(1, 41) = 22.211$, $MS = .024$, $p < .001$. That is, if subjects can recall that a trial involves a distinctive response as on the prospective memory trials, then they are more likely to recall information related to that response, such as the last location of the astronaut.

•There were overall differences between subject types (crew = .691, crew-like = .632, undergraduate = .535), $F(2, 41) = 8.402$, $MS = .184$, $p < .001$.

Crew-Crew-like, p = .241
Crew-Undergraduate, p < .001
Crew-like- Undergraduate, p = .025

IV. Training Sessions Results: Side Accuracy

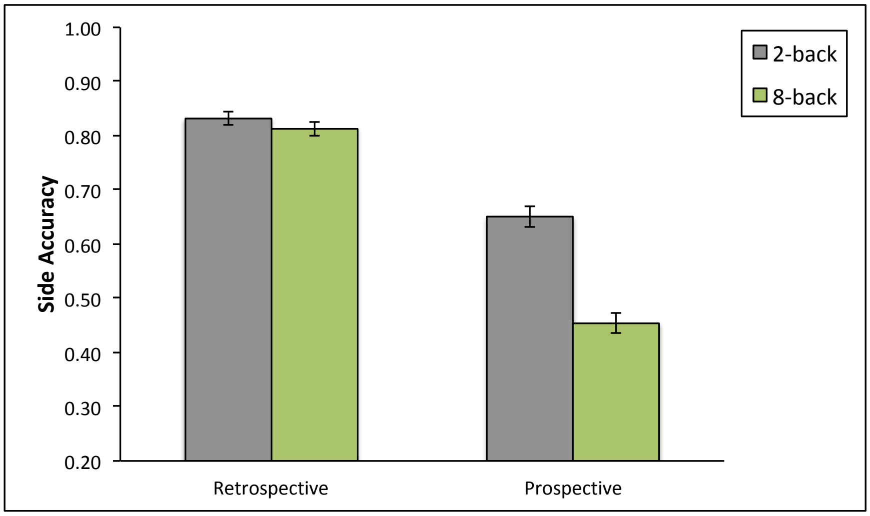


Figure 33. Side accuracy as a function of within-session retention interval and trial type.

•Averaging across WSRI, side accuracy lower for prospective memory trials (.553) than for retrospective memory trials (.823), $F(1, 41) = 73.502$, MS = .146, $p < .001$.

•Averaging across trial type, side accuracy higher for 2-back (.741) than for 8-back (.634), $F(1, 41) = 55.078$, MS = .035, $p < .001$.

•Larger effect of WSRI for prospective memory trials than for retrospective memory trials, $F(1, 41) = 28.515$, MS = .046, $p < .001$.

•Overall differences between subject types (Crew = .768, Crew-like = .718, Undergraduate = .634), $F(2, 41) = 4.919$, MS = .192, p = .012.
Crew-Crew-like, p = .329
Crew-Undergraduate, p = .005
Crew-like-Undergraduate, p = .094

Summary of Training Sessions Results

As for the analysis of pre-test and post-test, retention interval had different impacts depending on trial type (retrospective memory vs. prospective memory) and the dependent measure examined (location accuracy, side accuracy). Increasing within-session retention interval decreases location accuracy on retrospective memory trials (forgetting) but less so on prospective memory trials (protective function of distinctive responding). However, the opposite pattern is found for side accuracy. In general, for side accuracy, subjects were more likely to select the default response location rather than the distinctive. Also as during the pre-test, crew and crew-like subjects outperformed undergraduate subjects on both location and side accuracy.

Performance on the Mapping Task: Onboard Sessions

I. Method

- For Onboard Session 1, subjects completed 137 trials of the mapping task with no concurrent secondary task. The stimuli (ordering of astronaut names and locations) were different than during the training sessions, thus involving some degree of transfer.

- For Onboard Session 2, subjects again completed 137 trials of the mapping task, this time with the concurrent secondary task. The stimuli (ordering of astronaut names and locations) were different than during the training sessions and Onboard Session 1. Thus, Onboard Session 2 was more cognitively demanding and also required a greater degree of transfer than Onboard Session 1.

II. Onboard Sessions Analyses

- Comparison between onboard sessions (session 1/no secondary task, session 2/secondary task), trial type (retrospective, prospective), and within-session retention interval (2-back, 8-back). Subject type (crew, crew-like, undergraduate) also included as a factor.

- 2 dependent measures: side accuracy and location accuracy.

III. Onboard Sessions Results: Location Accuracy

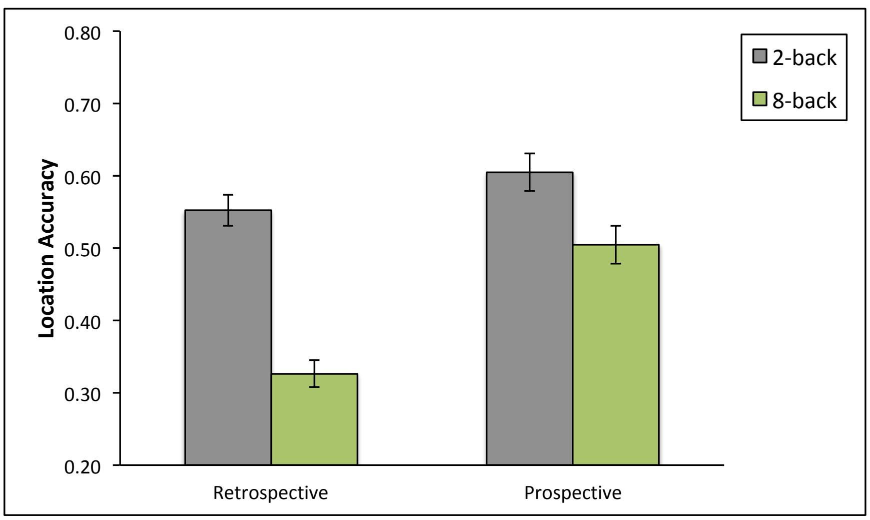


Figure 34. Location accuracy as a function of within-session retention interval and trial type.

- Averaging across WSRI, location accuracy higher for prospective memory trials (.555) than for retrospective memory trials (.440), $F(1, 43) = 29.885$, $MS = .043$, $p < .001$.

- Averaging across trial type, location accuracy higher for 2-back (.578) than for 8-back (.416), $F(1, 43) = 68.371$, $MS = .031$, $p < .001$.

- Larger effect of WSRI for retrospective memory trials than for prospective memory trials, $F(1, 43) = 22.514$, $MS = .017$, $p < .001$ (protective function of distinctive responding).

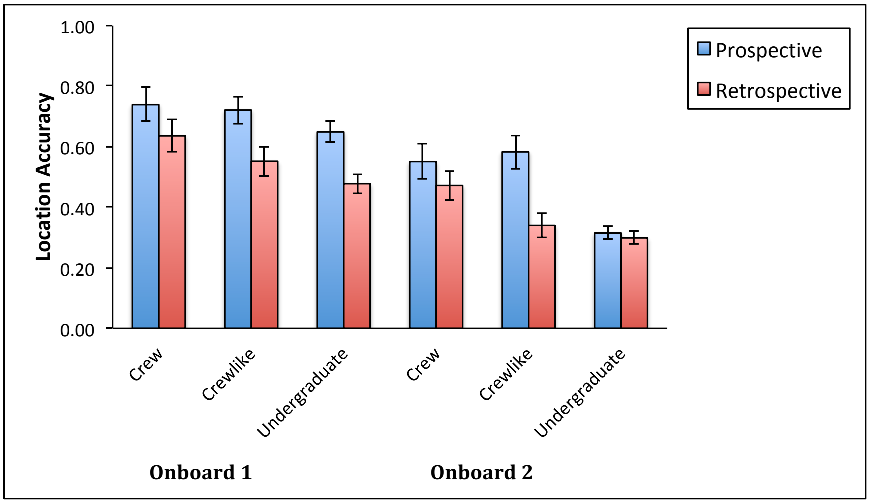


Figure 35. Location accuracy as a function of onboard session, trial type, and subject type.

•Location accuracy higher during first onboard session (.608, no secondary task) than second onboard session (.386, secondary task), $F(1, 43) = 59.747$, $MS = .052$, $p < .001$.

•For the first onboard session, location accuracy was higher on prospective memory trials than on retrospective memory trials for all subject types. During second onboard session, the same pattern was evident, but weaker for crew and undergraduate subjects and stronger for crew-like subjects, $F(2, 43) = 4.346$, $MS = .022$, $p = .019$.

•Overall, location accuracy higher for crew (.599) and crew-like (.551) subjects, and lower for undergraduate subjects (.435), $F(2, 43) = 10.059$, $MS = .096$, $p < .001$.
Crew-Crew-like, $p = .332$
Crew-Undergraduate, $p < .001$
Crew-like-Undergraduate, $p = .009$

<u>IV. Onboard Sessions Results: Side Accuracy</u>

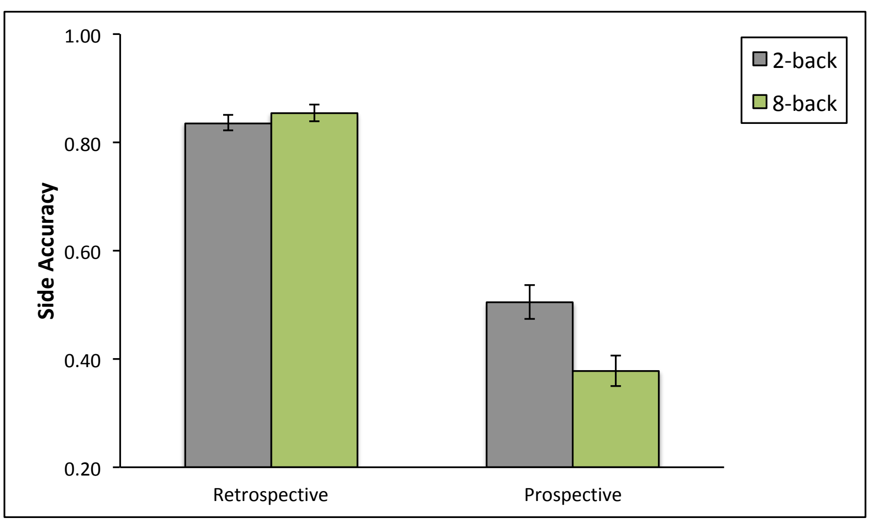


Figure 36. Side accuracy as a function of within-session retention interval and trial type.

•Averaging across WSRI, side accuracy lower for prospective memory trials (.442) than for retrospective memory trials (.845), $F(1, 43) = 111.572$, $MS = .092$, $p < .001$.

•Averaging across trial type, side accuracy higher for 2-back (.671) than for 8-back (.616), $F(1, 43) = 23.841$, $MS = .014$, $p < .001$.

•For retrospective memory trials, side accuracy equivalent for 2-back and for 8-back; for prospective memory trials, side accuracy higher for 2-back than for 8-back, $F(1, 43) = 38.121$, $MS = .015$, $p < .001$.

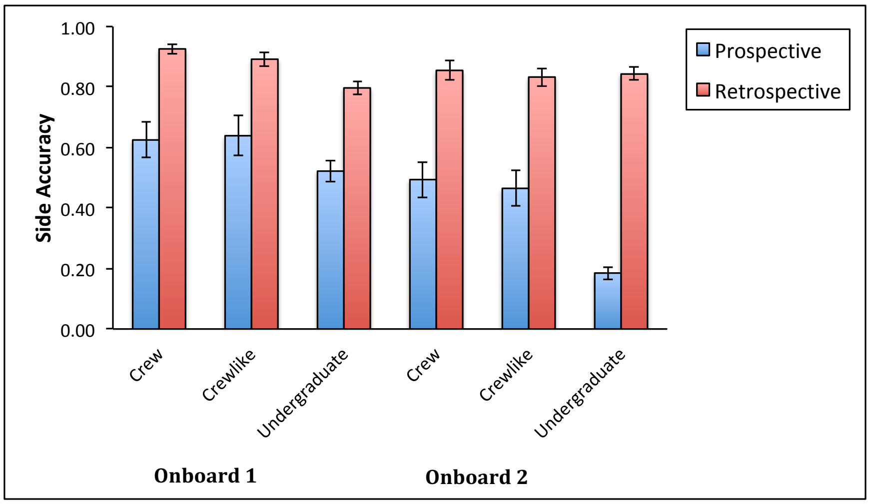


Figure 37. Side accuracy as a function of within-session retention interval and trial type.

•Averaging across subject type and trial type, side accuracy was higher for onboard session 1 (.708) than for onboard session 2 (.579), $F(1, 43) = 43.096$, $MS = .025$, $p < .001$.

•During the first onboard session, side accuracy was higher on retrospective memory trials than on prospective memory trials for all subject types. During the second onboard session, the same pattern was evident, but was especially strong for undergraduate subjects, $F(2, 43) = 5.702$, $MS = .045$, $p = .006$.

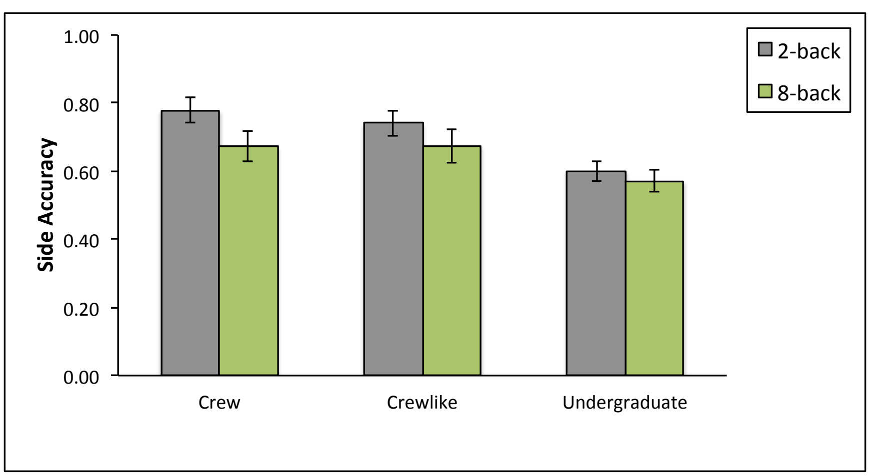


Figure 38. Side accuracy as a function of within-session retention interval and subject type.

•Overall, side accuracy highest for Crew (.725), lowest for undergraduate (.587), and intermediate for Crew-like subjects (.707).
$F(2, 43) = 10.959$, $MS = .071$, $p < .001$
Crew-Crew-like, $p = .678$
Crew-Undergraduate, $p < .001$
Crew-like-Undergraduate, $p = .002$

•Effect of WSRI (decrease in side accuracy from 2-back to 8-back) was strongest for crew and weakest for undergraduate students, $F(2, 43) = 3.368$, $MS = .014$, $p = .044$.

V. Summary of Onboard Sessions Results

As found for the pre-test/post-test as well as for training, within-session retention interval had different impacts depending on trial type (prospective memory vs. retrospective memory) and the dependent measure examined (location accuracy, side accuracy). Increasing WSRI decreases location accuracy on retrospective memory trials (forgetting) but less so on prospective memory trials (protective function of distinctive responding). Addition of a concurrent secondary task during the second onboard session decreased location accuracy, but did not eliminate the protective function of distinctive responding. The opposite pattern was found for side accuracy, for which increasing WSRI decreased accuracy for prospective memory trials rather than retrospective memory trials. The addition of a concurrent secondary task during the second onboard session also decreased side accuracy overall.

Crew and crew-like subjects outperformed undergraduate subjects on both accuracy measures. For side accuracy, all subject types showed greater accuracy for retrospective memory trials than for prospective memory trials. However, the addition of a concurrent secondary task did not change this pattern for crew and crew-like subjects, but increased the difference for undergraduate subjects. This finding suggests that crew and crew-like subjects are better able to handle the increase in cognitive demands required by secondary tasks.

Appendix B

Background Information Concerning the Study of the Effects of Long-Duration Space Flight on Training Retention and Transfer

NASA's future deep space missions to the lunar environment, lunar surface, and to Mars involve both a long-duration and long communication delays. These factors drive a requirement for durable training, namely, training that can survive long delays between time of training and actual use of the learned knowledge and skills. In addition, the training needs to be flexible, allowing for transfer to new, unexpected situations beyond those included in the training regimens, because it will not be possible to accurately anticipate all the possible situations in which crewmembers might find themselves. Furthermore, onboard just-in-time (JIT) training and performance support tools will be required.

On such long-duration space missions with communication delays, the crew will need to be semi-autonomous, if not completely autonomous. Current crew training practices are based on the assumption that ground support is available continuously and in real-time. That will not be the case in future long-duration space missions. Crew training must, therefore, evolve. However, we have no direct experience with long-duration space missions and serious gaps exist in our knowledge upon which new crew training must be based. We must use International Space Station (ISS) missions to develop the operational understanding of training retention and transfer, as well as JIT training requirements under the conditions of a long-duration space mission. Specifically, we must, for the first time ever, collect data on training effectiveness while on orbit.

The Training of our crew members for long duration, exploration-class missions will have to maximize long-term retention and transfer of the trained skills. The expected duration of the missions, our inability to predict all the possible tasks the crew will be called upon to perform, and the low training-to-mission time ratio require that the training be maximally effective such that the skills acquired during training will be retained and will be transferrable across a wide range of specific tasks that are different from the particular tasks used during training. However, to be able to design training that can achieve these ambitious goals, we must first understand the ways in which long-duration spaceflight affects training retention and transfer.

Current theories of training retention and transfer are largely based on experimental studies conducted at university laboratories using undergraduate students as participants. Furthermore, all such studies have been conducted on Earth. We do not know how well the results of these studies predict the performance of crew members. More specifically, we do not know how well the results of these studies predict the performance of crew members in space and especially during long-duration missions. To address this gap in our knowledge, the study sought to test the null hypothesis that performance of university undergraduate students on Earth on training retention and transfer tests do in fact predict accurately the performance of crew members during long-duration spaceflights. To test this hypothesis, the study employs a single 16-month long experimental protocol with 4 different participant groups: undergraduate university students, crew members on the ground, crew-like subjects on the ground, and crew members in space.

The study was originally approved for flight and assigned crew members were being trained to participate in the study. The first participant in space was to be the 1-year mission astronaut, to be followed by standard duration crew members. About a month into the 1-year mission, and shortly before the astronaut was to start the experiment, the study was dropped from the crew timeline. At the time, it was promised that the study will be reinstated once crew time became available.

Without the flight portion of the study, the fundamemntal knowledge of crew performance on trained and transfer tasks in space is still missing.

As shown in the report above, the Ground Phase of the study shows significant differences in magnitude and direction between the performance of university undergraduate students and that of crew members. These differences were particularly clear under increased cognitive demands. These findings highlight the need for the flight portion of the study as the space-related cognitive demands are much greater than those on Earth. The most critical part of the awarded Directed Research Project was collecting data from crew members on board the ISS, which would have helped make progress towards Training gaps closure. Unfortunately, the ISS data collection phase of this DRP was canceled and only the Ground Phase was performed. The results of the Ground Phase alone do not by themselves provide input towards gap closure. On the contrary! The current results suggest that the gaps are in fact wider than originally assumed.

The description below provides background information on the study as a whole, and on the specific tasks used in the ground phase of the study discussed in the main body of this report.

Specific Aims

Three specific aims were included in the originally proposed study:

> *Aim A***.** Test the retention and transfer of specific motor, perceptual and cognitive processes learned pre-launch to assess the need for (and possible schedule of) onboard refresher and just-in-time (JIT) training.
>
> *Aim B.* Compare the process of knowledge/skill decay on orbit with that of closely matched participants on Earth and with that of undergraduate university students to determine the applicability of existing Earth-based theories of learning to space operations.
>
> *Aim C*. Collect naturalistic data from onboard crew and ground-control personnel on training-related crew performance including performance errors, requests for ground support, need to review previously learned material, and training success stories.

Note: The portion of the study reported above addresses only Aim A, and part of Aim B. Aim C was removed from the study by HRP/ISSMP (now HRP/ROI.

Overview

To date, no systematic data collection has been conducted in spaceflight to document the effectiveness of pre-flight crew training. Crewmembers have been largely successful in their performance, but that success could have primarily been the result of excellent innate capabilities, extreme motivation, and "as needed" support from mission control. Many studies have documented the processes of skill decay and the forgetting of acquired knowledge. However, all these studies have been conducted on Earth.

It is an understatement to say that space is a very different environment than the one people are accustomed to on Earth. Yet, almost all current crew training is done on Earth. Zero-G is only one aspect of the difference that cannot be properly simulated in Earth-based training, but it is a feature of space operations that may have significant impact on the effectiveness of Earth-based training, and on the ability of crewmembers to retain their knowledge and to acquire new skills in space.

In addition to zero-G, the phenomenon of space adaptation, the stresses of confinement, noise, reduced-quality sleep, the effects of high CO2 levels on CNS and cognition ("space fog"), and the ever-present threat to basic survival are all factors that affect people's behavior and cognitive capabilities. Little to no data are available on how people learn in space or how retention and retrieval of Earth-based training are affected by being in space over a long period of time.

Currently available training principles, methodologies, and tools are based on studies and training experience on Earth (see, e.g., Barshi, 2015; Healy & Bourne, 1995, 1998, 2012a, 2012b; McDaniel, 2007; Schmidt & Bjork, 1992). It is not clear to what extent Earth-based knowledge truly extends to long-duration space operations. Not only does most Earth-based research fail to incorporate the impacts of space operations on skill acquisition, retention, and retrieval, but also much of this research is conducted in universities with undergraduate college students as research participants.

The common argument made in support of using college students as research participants is that basic cognitive processes are similar across the population, and crewmembers are all college graduates—thus, results obtained in studies with college students are indeed representative of crewmembers' cognitive processes. Recent results reported by Hah, Willems, Ayaz, Bunce, Izzetoglu, and Deshmukh (2011) bring this argument into question.

Hah and his colleagues compared performance on memory tasks and brain activities in college students and air traffic controllers. As in the case of astronauts and space operations, much of the research underlying design choices for air traffic control systems and procedures is based on university studies done with college undergraduate students. Hah used cognitive tests and functional near-infrared spectroscopy, identifying significant differences in both memory performance and in the nature of the prefrontal cortex brain activity between the two groups of participants (i.e., experienced air traffic controllers and undergraduate college students). These differences can be attributed to age and brain maturation, as well as to experience with developing and employing different cognitive strategies, both in number and in structure. Similar differences are likely to be present between undergraduate college students and astronauts, and to characterize such differences, we must study both of these subject groups.

To develop the operational understanding of the relevancy of Earth-based research results for the design of crew training, we must study crewmembers' experience, and we must utilize ISS missions to study the retention, transfer, and generalizability of skills in long-duration space operations. It is only with such understanding that we can design and implement an effective and efficient training program to support our crews on future long-duration, Exploration Class and Deep Space missions.

Background and Significance

There are three basic concerns in optimizing training (see Healy & Bourne, 2012a; Healy, Wohldmann, Kole, Schneider, Shea, & Bourne, 2013). The first concern is *efficiency*, so that the time required to complete training is minimized (Healy, et al., 2013). The second concern is *durability*, so that the trained knowledge and skills are available following delay intervals without practice (Healy, Fendrich, Crutcher, Wittman, Gesi, Ericsson, & Bourne., 1992, and Healy, Clawson, McNamara, Marmie, Schneider, Rickard, Crutcher, King, Ericsson, & Bourne, 1993). The third concern is *generalizability* or transferability (e.g., Healy & Wohldmann, 2012), so that the trained knowledge and skills may be applied to new tasks and new situations. Some studies have shown that efficient training is not necessarily durable (e.g., Schneider, Healy, & Bourne, 2002), and other studies have shown that durable training is not necessarily generalizable (e.g., Lohse & Healy, 2012). Thus, each aspect of training needs to be considered both separately and in combination.

In addition to these basic concerns, there are many practical concerns. For example, research has shown that typical operational procedures and training programs often assume an idealized operational environment free of interruptions and distractions, despite evidence that shows that effective and efficient training must account for the nature of the real-world operation (Barshi & Loukopoulos, 2012; Loukopolous, Dismukes, & Barshi, 2009). Further, when it comes to the design of JIT training programs and the optimal interval of referesher training, no useful research results exist. Such concerns must be adequately addressed in the design of crew training for long-duration space missions (for further discussion of these issues, see Barshi, 2006; 2011).

Training involves three fundamental underlying cognitive processes: acquisition (learning), retention (memory), and transfer (generalization). There are basic laws that apply at the level of these fundamental processes.

a. ***Acquisition: Power law of practice.*** There are two major measures of performance during the acquisition of knowledge and skills: accuracy and speed of responding. With respect to response speed, Newell and Rosenbloom (1981) have argued that the *Power Law of Practice* describes the acquisition process for most skills. This law formalizes the relationship between trials of practice and time to make a correct response as a power function, $R = aN^{-b}$, where R is response time on trial N, a is response time on trial 1, and b is the rate of change. It follows that the relationship between response time and trial number is linear in log-log coordinates (i.e., $\log R = \log a - b \log N$). In some cases, where more than one strategy can be used in the task, separate power functions apply to the different strategies (Delaney, Reder, Staszewski, & Ritter, 1998; Rickard, 1997). This law affords a way of predicting response time in a variety of tasks as a function of degree of practice. With respect to response accuracy, a similar function seems to apply (e.g., Bourne, Healy, Parker, & Rickard, 1999) although a power function has not been proposed for such data.

In some cases, speed and accuracy might not be positively correlated (e.g., Pachella, 1974). People sometimes trade speed for accuracy or vice versa (i.e., they exhibit a speed-accuracy tradeoff). Likewise, the speed of executing the different steps of a complex task may not be positively correlated, with people slowing down on one step in order to be faster on another step (Healy, Kole, Buck-Gengler, & Bourne, 2004; Kole, Healy, & Bourne, 2008). In these cases, the power law of practice might not be a good description for all measures. Furthermore, for optimal training, instructors need to be aware of the various steps in any task as well as whether speed or accuracy is more important in each step, so that the more important aspect can be emphasized in training.

b. ***Retention: Power law of forgetting.*** With passage of time and lack of opportunity to refresh acquired knowledge or skills, performance declines, reflecting forgetting of what was learned. This decline in performance, exhibited in increased response time (or decreased accuracy), has been known since the time of Ebbinghaus (1885/1913), who used a measure of savings (i.e., the amount of relearning required to reach a criterion level of performance achieved during original learning) to measure retention. Subsequently this relationship between response time and retention interval was described as a power function (Wickelgren, 1974), $R = d + fT^{-g}$, where R is response time, T is the retention interval, d is the criterion of original learning, f is a scaling parameter, and g is the rate of forgetting. This *Power Law of Forgetting* (Wixted & Carpenter, 2007; see also Rubin & Wenzel, 1996) can be thought of as the inverse of the *Power Law of Practice.*

c. *Transfer: Laws relating to similarity*. Training on a particular task has implications for performance on other, related tasks. The effect of training on one task can be either positive (facilitation) or negative (interference) on performance of another task. When the acquisition of one task affects performance on another, that effect is called *transfer*. The major variable determining the extent and direction of transfer is similarity between the two tasks. Osgood (1949) has conceptualized this relationship as a transfer surface, which relates transfer magnitude both to response similarity and to stimulus similarity between the training and the transfer tasks. When the stimuli in the two tasks are varied in similarity and the responses are held constant, positive transfer is obtained, with its magnitude increasing as the similarity between the stimuli increases. On the other hand, when the stimuli are held constant and the responses are varied in similarity, negative transfer is obtained, with its magnitude decreasing as the similarity between the responses increases. Finally, when both the stimuli and responses are simultaneously varied in similarity, negative transfer is obtained, with its magnitude increasing as the similarity between stimuli increases. Shepard (1987) has given a quantitative expression to such similarity functions, which he refers to as a *universal law of generalization*.

In this study, we test performance on all three fundamental training processes described above. Such a study has never before been conducted in space. This is the first time we have an opportunity to collect data on actual crew performance during a long-duration space mission, and infer acquisition, retention, and transfer functions.

Current Training Practices for Space Missions

Astronaut training is largely based on extended practice and rote learning (e.g., http://www.spaceflight.nasa.gov/shuttle/support/training/ascan/2004/index.html; personal communications). Several implicit assumptions underlie this practice, including trust that the selection process guarantees highly capable individuals, and that the nature of the mission guarantees extremely high motivation. Long lead time (2-3 years) for training, short (1-2 weeks) and highly scripted Shuttle missions, and continuous communication with ground support have allowed for mission success. Approaches and methods in astronaut training have hardly changed since the early days of the space program.

During the Space Shuttle era, crewmembers were in training for several years prior to being assigned for a mission, then spent 18-24 months in training for their specific mission, at the end of which they spent 10-17 days in orbit. Such extensive training allowed for repeated practice and rehearsal of a tightly scripted mission, almost to the point of automaticity. Much of the ISS training for mission tasks and space science is similar, except that the mission is longer (usually, 3 or 6 months), and the schedule isn't as tightly scripted in advance - as was the case for Shuttle missions. In either case, our experience, current knowledge and practice, and our training assumptions are all based on extensive pre-mission training and relatively short mission duration. All that must change as we turn to future long-duration Exploration Class space missions.

Long-duration spaceflights (those flights in excess of six months) are so physically, mentally, and emotionally demanding that simply selecting the individual crewmembers with the "right stuff" is not sufficient (Flynn, 2005). Training and supporting optimal performance, as well as selecting high performers, is a more effective and efficient approach than just selecting the high performers (Holland, Hysong, & Galarza, 2007). However, training team skills and supporting optimal performance entails more than educating astronauts about the technical aspects of the job. It also requires equipping astronauts with the resources to maintain their psychological and physical health in an isolated, confined, and extreme environment over an extended period of time. While crewmembers often engage in expeditionary training activities such as National Outdoor Leadership School (NOLS) to promote team cohesion, there is very little scientific evidence regarding what type of training offers the best means of promoting team performance for long-duration missions, and the best method for implementing this training, including how often to conduct training and the sequence of training.

It is also likely that vehicle and interface technology will continue to evolve and change all the way until shortly before launch of the first new Exploration mission. For example, powerful smaller computing device capabilities, advanced development of decision support systems, and artificial intelligence could change the nature of the space mission technologies. Thus, pre-mission training may have only a limited application. Furthermore, given the expected gap between the end of the Shuttle era (2011) and the possible end of the Station era (2028) on the one hand, and the beginning of a new Exploration Class mission on the other hand, it's possible that crewmembers assigned for the first new Exploration mission will have had little to no prior spaceflight experience. Also, given the expected length of the mission, it will not be possible to maintain the same ratio of training time to mission time as we have had in previous short-duration missions. Finally, given the expected length of the mission, pre-mission training cannot be expected to be retained throughout mission duration without some sort of refresher training. Thus, onboard training is unavoidable, and given the length of time the crew will be expected to be in transit, training is one of the few activities they'll be able to engage in during that otherwise largely passive and monotonous phase of the mission. Thus, having data on retention during ISS missions and data that could help determine schedules for refresher training will be critical to our ability to design effective crew training for long-duration Exploration missions.

The Training Continuum

Training opportunities can be viewed along a timeline continuum. This timeline starts before candidates are even selected, and extends all the way through the completion of a mission and the various post-mission debriefs. Key points along this training continuum include the following:

- Earth-based pre-selection training and skills
- Earth-based pre-assignment training
- Earth-based pre-mission training
- On-board initial training
- On-board refresher training
- On-board JIT training (including crew-generated)
- On-board performance and decision support tools.

Note, depending on a projects' schedule, it is possible that crewmembers assigned to the first new mission, for example – a cis-lunar mission, will have the opportunity to spend time in training onboard the ISS prior to their cis-lunar mission. In such a case, we might have a "space-based pre-mission training" point added to the list above. Although such training is likely to be very beneficial for the crew, it doesn't obviate the need to provide that crew with onboard training and performance support on their long-duration mission.

Although we have extensive experience with Earth-based training, as discussed earlier, not all of it is applicable to future missions. What's more, we have very little experience with onboard training, and practically no experience with performance and decision support tools (Wenzel et al., 2019). For the Space Shuttle and for ISS, Mission Control Center (MCC) has been providing the necessary performance and decision support, sometimes in the form of talking crewmembers through each and every step of a procedure. Such support is relatively easy when two-way communication is available continuously and in real time. The communication delays anticipated for future long-duration missions significantly reduce the practicality of such support (Fisher & Mosier, 2014). In most situations, and certainly in all time-critical situations requiring immediate response, the crew will have to act autonomously. Documenting and understanding the ways in which crewmembers rely on MCC for performance support will enable us to design the kind of onboard support the crew will need, when access to MCC will be restricted (Wu & Vera, 2019; Holden et al., 2019).

Given the increased importance of on-board training for mission success, training retention takes on another degree of importance with regard to successfully implementing on-board training in terms of both timing and sequence. This is particularly true with regard to training related to behavioral health and performance, including training countermeasures focused on coping with stress, maintaining and improving team performance, preventing conflicts within the crew, and maintaining resilience and adaptability. While a number of training countermeasures have been developed and validated to maintain and improve crew well-being and performance, the timing for when these countermeasures should be implemented and the sequence in which they should be implemented for maximum benefit is unknown. Studies investigating training retention over extended durations in spaceflight will be critical in informing the implementation of these training countermeasures for future exploration missions.

The medical providers on a long-duration space flight must be prepared to manage diverse clinical quandaries from the Space Medicine Exploration Medical Condition List (EMCL). There are many variables that may affect crew health during an exploration mission, and there is no way to accurately predict or prepare for every eventuality. Medical conditions that could be encountered during an Exploration-Class mission span a wide spectrum, with diagnostic and treatment requirements ranging from relatively standard treatment of common EMCL conditions with delayed definitive care, to Advanced Trauma Life Support and point-of-injury stabilization, to improvisation and high-level medical technique. As vocalized by the majority of surveyed training experts, Exploration-Class missions will likely demand physician-level capabilities to effectively address the variety of potential medical contingencies, and medical providers must be fully prepared for the challenges they will face during an exploration mission.

In terms of JIT training, Byrne, Schmid, and Barshi (2010) have conducted investigations using a human patient simulator to look at the feasibility of using novel display technologies (e.g., hand-held display device, head mounted display device) for JIT medical procedures. Results show that these implementations are feasible, but to provide a complete JIT solution for long-duration missions, we need to know what types of information need to be displayed, and under what circumstances.

Our limited knowledge of onboard training is unfortunately illustrated in the experience of the refresher training for the Dust and Aerosol Measurement Feasibility Test (DAFT) experiment, which was uplinked to ISS on February 3rd, 2005; it consisted of only three PowerPoint slides, each with a few bullet points, one slide with two pictures of relevant equipment, and one slide of the general process flow. Multiple challenges arose, leading to the premature termination of the protocol (see, e.g., Evans, Robinson, Tate-Brown, Thumm, & Crespo-Richey, 2009; Urban, Griffin, Ruff, Cleary, Yang, Mulholland, & Yuan, 2005). Future crews may not even be able to rely on training uplinked from Earth, and instead may have to be able to generate their own training when needs arise. Thus, understanding retention during ISS missions and the needs of refresher and JIT training will be critical to our ability to design effective crew training for long-duration exploration missions. This research does not aim to be an exhaustive assessment of training retention and JIT needs; rather, it aims to gain some initial insight into training issues that will help characterize the Training risk. All conclusions will be made within the context of the data collected.

Research Design and Method

The portion of the study reported above addresses, in part, Aim A and Aim B, repeated below. The complete study would have provided converging lines of evidence to support the design of effective crew training for long-duration exploration missions. The complete study would have informed risks in the HFBP and ExMC Elements of HRP.

> *Aim A.* Test the retention and transfer of specific motor, perceptual and cognitive processes learned pre-launch to assess the need for (and possible schedule of) onboard refresher and just-in-time (JIT) training.

Aim B. Compare the process of knowledge/skill decay on orbit with that of closely matched participants on Earth and with that of undergraduate university students to determine the applicability of existing Earth-based theories of learning to space operations.

Null Hypotheses for the Ground Phase of the study

- Training principles established in Earth-based, university-based studies hold for astronauts in space operations.

- The Power Law of Forgetting and the Universal Law of Generalization that have been established in Earth-based, university-based studies hold for astronauts in a long-duration study.

Note, since the null hypotheses above make predictions about the performance of astronauts in space operations, and since the portion of the study reported here does not include data from the performance of crew members in space, these two hypotheses are not discussed further in this report.

Method

To address Aim A, this study entails administering a pre-launch training program to subjects based on the training and mission of a standard-duration ISS crewmember. There are six sessions in the 16-months long study, beginning with a baseline data collection (BDC), followed by two training sessions, two test sessions, and a final post-flight session (Figure 1). This 16-months timeline is the schedule of the study when conducted with standard duration (approx. 6 months long) crew members on board the ISS. It is not the timeline designed for a year-long mission.

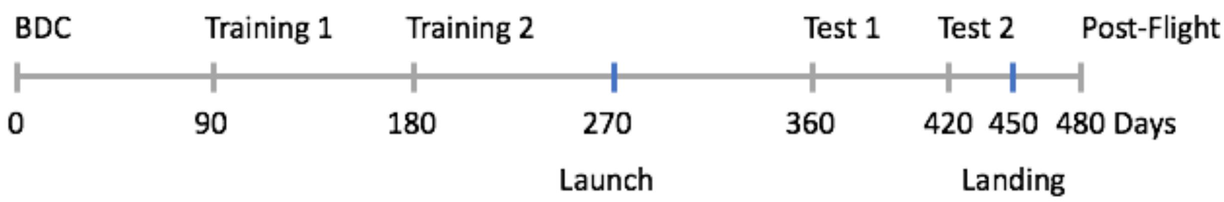


Figure 1. 16-Months Training Retention Study

Training consists of two tasks:

- A data entry task
- A mapping task

Following launch, retention tests are administered at two intervals: 6 months after completion of the two training sessions and 8 months after training. Some test items are specifically designed to determine transfer and generalizability of the material learned. Speed and accuracy of responses to the test items are be recorded.

To address Aim B, the study utilizes ground-matched participants, who complete the same protocol under the same schedule as that of the ISS participants.

Participants

Participants consist of astronauts, "crew-like" subjects, and university undergraduate students. "Crew-like" subjects are drawn from a population that is matched to crew members in age, gender, health and level of education.

Data Entry Task

In the data entry task, subjects typically see four-digit numbers, read them silently, and type them into the computer. (For training principles derived from the data entry task, see Healy, Kole, Wohldmann, Buck-Gengler, & Bourne, 2011). The task is a sequential task with both cognitive and motoric requirements that can be examined separately through the different components of response time. For example, response execution time (which is the average time to type the second, third, and fourth digits after typing the first) has been shown to reflect the motoric aspects of the task (e.g., Chapman, Healy, & Kole, 2016; Fendrich, Healy, & Bourne, 1991).

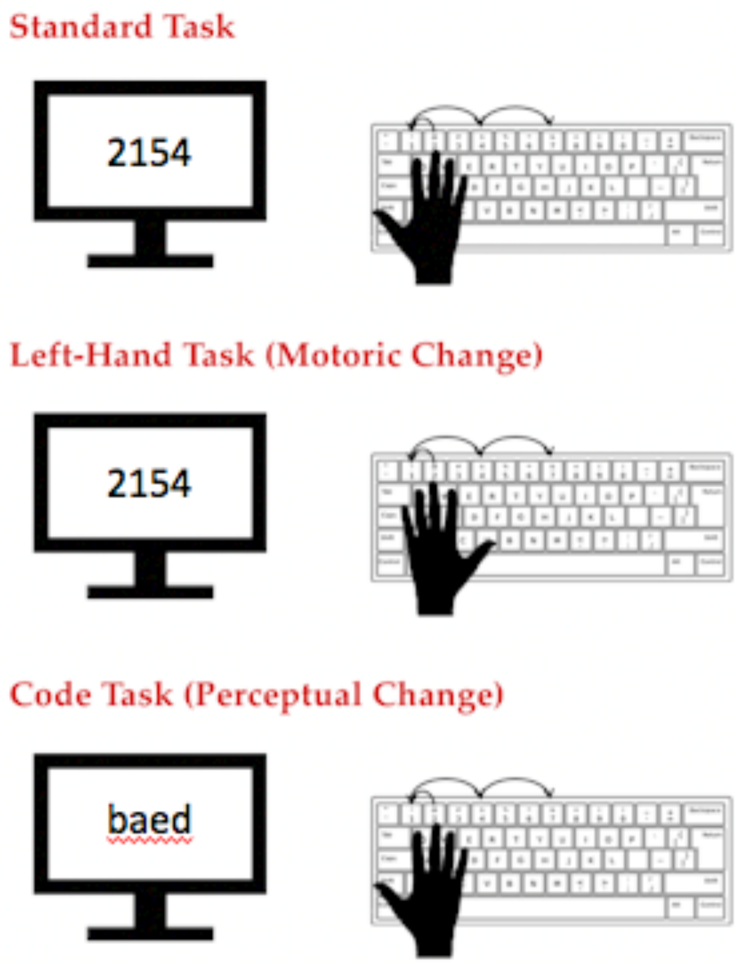


Figure 2. Data Entry Task: Standard, Left-Hand, Code Task

At Training, subjects are given the standard version of the data entry task, in which they are shown four-digit numbers presented as numerals and entered them using the keyboard with their right/dominant hand (Figure 2). There are 2 Training sessions, scheduled 3 months apart with 3 blocks, 100 trials per block, each session. The same 100 numbers in all 6 blocks.

At Test 1, 6 months after completion of both training sessions, subjects are given the standard task along with a left-hand/non-dominant-hand variant (Figure 2). The left-hand variant involves different motoric processes because the numbers were entered with the left hand rather than with the right hand, but the perceptual aspects of the task did not change. For the standard task there are 50 old numbers (used during training) and 50 new numbers and for the left-hand/non-dominant hand variant there are also 50 old and 50 new numbers.

At Test 2, 8 months after training, subjects are given the standard task along with a code variant (Figure 2). The code variant involves different perceptual processes because participants see letters and enter digits, but the motoric aspects of the task did not change. For the standard task there are 50 old numbers (used during training) and 50 new numbers and for the code variant there are also 50 old and 50 new numbers.
The numbers entered during Tests 1 and 2 are either the same as during training (old) or numbers entered for the first time during the test (new). Repetition priming (old faster than new) at test for the standard task reflects specificity of training. Repetition priming for the left-hand task implies motoric transfer and for the code task implies perceptual transfer.

Mapping Task

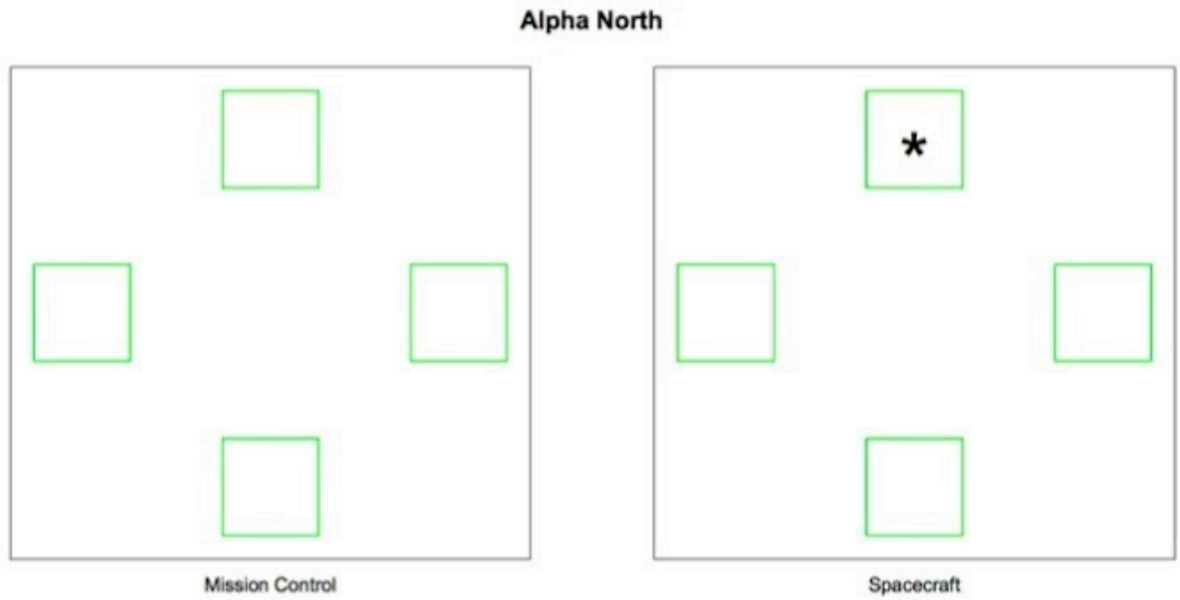


Figure 3. Mapping Task Default Format

In the mapping task, subjects make associations between eight astronauts identified alphabetically (Alpha, Bravo, Charlie, etc.) and four locations in an explored space (North, South, East, West). All study trials consisted of presenting a name-location association, to which subjects responded by clicking at the appropriate location on a map on the right-hand side of a display screen labeled Spacecraft (Figure 3).

Most of the messages are in the default format (with Black type and mixed font). A small number, however, are distinctive by being in an atypical format (with blue type and all capital letters).

Test trials, intermingled with study trials, require responding with the last known location and consisted of the name alone in Blue.

- On default test trials the response is made in a default manner by clicking on the right-hand Spacecraft map. This is a retrospective memory task.

- On distinctive trials the response is made instead on a left-hand Mission Control map. This is a prospective memory task.
- Two retention intervals are included, involving either one interpolated trial (2-back) or seven interpolated trials (8-back) between the trial stating the location of an astronaut and the trial inquiring about the previously stated location of that astronaut.

At Training subjects are given 2 sets of trials with 124 trials per set (including study trials). There are 32 Default (retrospective) test trials with 16 2-back and 16 8-back, and 16 Distinctive (prospective) test trials with 8 2-back and 8 8-back. There are 2 Training sessions, scheduled 3 months apart.

At Test 1, 6 months after training, subjects are given different stimuli.

At Test 2, 8 months after training, subjects are given different stimuli and are given a counting task in which they count out loud backwards by 3's from 100 concurrently while conducting the test.

Appendix C

Brief responses to comments (repeated below) provided by the HFBP Element Scientist in an email dated 8/22/19.

1. This will represent the final report for this Grant. Therefore, the Deliverables should also include a comprehensive report that not only provides the "Ground Phase Report", but also ensures the Ground Phase Report is incorporated into, and placed in the context of the overall aims of this Grant to ensure it addresses the relevance of this research to the Gaps in knowledge in the identified risk(s) originally identified as the basis for this Grant. In addition, for the Ground Phase Report section of the overall report, the aims of the research should be addressed along with how the research methodology and results to help make evident the contribution your research is making to the respective gap closures. While this may have been implied, the current wording only references "The deliverable for this supplement *will be an analysis comparing the retention and transfer for the three ground-phase subject groups* over a long duration, mimicking a deep space mission. This comparison will help close the gap on how to determine retention in operational spaceflight and will identify information and guidance likely to be needed in the form of refresher training, Just-in-Time materials JITT, or performance support tools." [Emphasis added]

The work described in this report is part of a research effort that was specifically requested directly by the HRP Program Scientist, independently of any particular program gaps. Following reviews of the proposal for this research by members of an expert external review panel, the research effort received Authorization to Proceed as a Directed Flight Study Task, rather than a "grant." The original request of the Program Scientist was to study crew members on board the ISS during a 1-year long mission. The ground phase of the study was proposed as a "control group" to the on-orbit crew, to provide us with new fundamental knowledge of astronaut training retention, and as a way to test the relevancy of existing university-based research to space operations. Since the Training Gaps are concerned with the performance of crew members in space during long-duration space operations, results from the ground phase alone do not contribute directly to gap closure. Furthermore, because the schedule of the ground phase of the study was aligned with on-board standard duration missions rather than a year-long mission, the results do not contribute directly to our understanding of crew performance under the demanding conditions of substantially longer duration, Exploration-Class or Deep Space missions. However, given the clear differences in performance in general, and in the shape of performance functions in particular, between university undergraduate students and crew members, especially under conditions of increased cognitive load, it is safe to say that results obtained in studies with undergraduate university students do not fully predict the performance of crew members. Because the space environment imposes on the astronaut crew significant physiological, psycho-social, and cognitive loads that cannot be replicated on the ground, it is still to be determined the extent to which the performance of crew members on the ground predict the performance of crew members in space. Moreover, it is still to be determined the extent to which the performance of standard duration crew members on-board the ISS predict the performance of crew members in space on long duration Exploration-Class or Deep Space missions. Without that knowledge, gaps cannot be closed. In fact, the results of the Ground Phase alone do not by themselves provide input towards gap closure, and the current results suggest that the gaps are actually wider than originally assumed.

2. Since the Training Risk is currently nested within the overall HSRB HSIA risk, we also request that you specifically address the context of your research and results with regard to their specific contribution to the HSIA risk.

The Training Risk is a critical component necessary to reducing the overall HSRB HSIA risk.

> **HSIA Risk Statement**: Given decreasing real-time ground support for execution of complex operations during future explorations missions, there is a possibility of adverse performance outcomes including that crew are unable to adequately respond to unanticipated critical malfunctions or detect safety critical procedural errors.

Training for future missions depends on the concept of operations, as well as on the work environment, including the habitat and human-computer capabilities provided for the crew (such as autonomous systems, procedures and interfaces, and robotics support). Research on these work environment capabilities resides within the other risks embedded in the HSIA risk. These risks combine synergistically to ensure crew readiness, along with other HFBP risks, and to reduce the overall HSIA risk (see Figure 1).

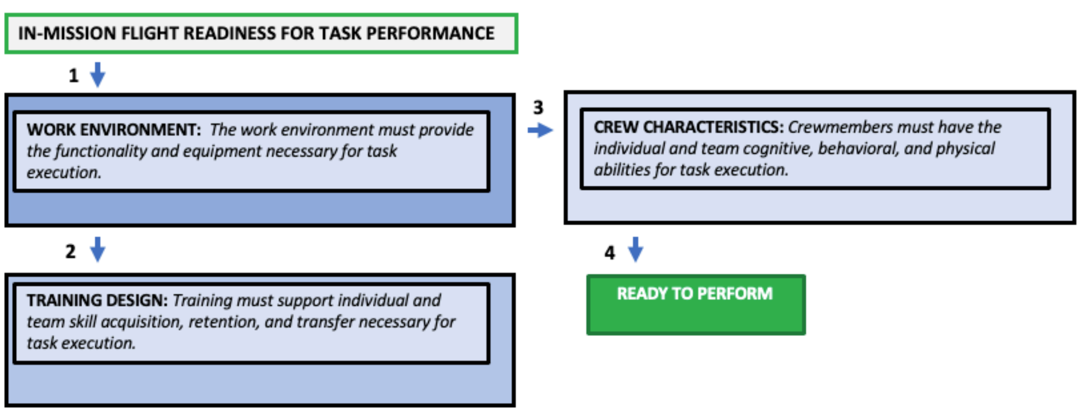


Figure 1. Readiness to perform complex operations, including responding to unanticipated critical malfunctions, depends on several key factors including the work environment, training, and crew characteristics (see e.g., Baldwin and Ford, 1988).

This report provides the results of a portion of a research effort that begins to provide a foundational understanding of the astronauts as a unique learner population. Future Training Risk research will build upon this knowledge to provide answers to questions on models of training for expeditionary skills and advanced training models. The results of this report along with the results of future research tasks will be used to develop agency-level training standards for crew training. These standards will ensure that NASA provides future exploration mission crews with training that produces the durability and transferability of knowledge and skills necessary to perform complex operations. More importantly, the training will ensure that the crews are able to "adequately respond to unanticipated critical malfunctions or detect safety critical procedural errors."

3. Finally, please provide an explanation for how these results will contribute to the Training Evidence Book, what section in the HIDH and standards in NASA STD 3001 you are proposing to augment.

A discussion of the implications of the results of this portion of the study will be included in the next update to the Training Chapter of the Evidence Book. These results suggest some clear limits on retention and transfer of knowledge and skill, and they help explain some of the anecdotal evidence of skill decay and forgetting experienced in space operations. Unfortunately, currently, there are no Training Standards in NASA STD 3001, and thus no chapter on Training in the HIDH. The Training Risk Team is in the process of proposing such standards.