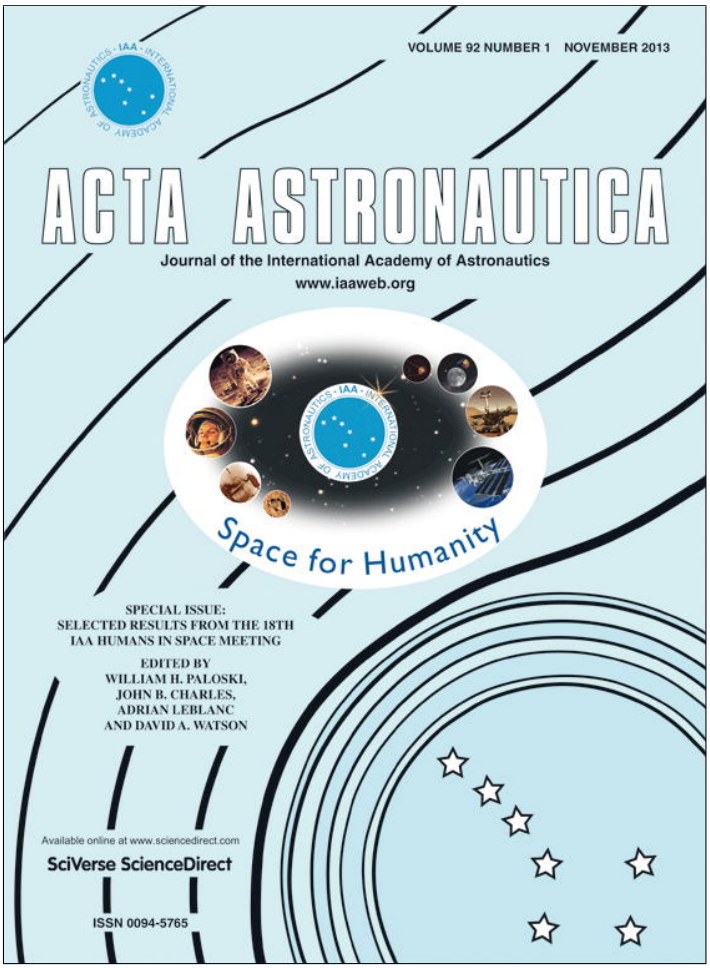


Acta Astronautica 92 (2013) 38–47

Contents lists available at ScienceDirect

Acta Astronautica

journal homepage: www.elsevier.com/locate/actaastro

Predicting space telerobotic operator training performance from human spatial ability assessment

Andrew M. Liu*, Charles M. Oman, Raquel Galvan, Alan Natapoff

Man Vehicle Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, 37-219, Cambridge, MA 02139, USA

ARTICLE INFO

Article history:
Received 5 December 2011
Accepted 2 April 2012
Available online 2 May 2012

Keywords:
Teleoperation
Spatial ability
Mental rotation
Perspective taking
Logistic regression
Training

ABSTRACT

Our goal was to determine whether existing tests of spatial ability can predict an astronaut's qualification test performance after robotic training. Because training astronauts to be qualified robotics operators is so long and expensive, NASA is interested in tools that can predict robotics performance before training begins. Currently, the Astronaut Office does not have a validated tool to predict robotics ability as part of its astronaut selection or training process. Commonly used tests of human spatial ability may provide such a tool to predict robotics ability. We tested the spatial ability of 50 active astronauts who had completed at least one robotics training course, then used logistic regression models to analyze the correlation between spatial ability test scores and the astronauts' performance in their evaluation test at the end of the training course. The fit of the logistic function to our data is statistically significant for several spatial tests. However, the prediction performance of the logistic model depends on the criterion threshold assumed. To clarify the critical selection issues, we show how the probability of correct classification vs. misclassification varies as a function of the mental rotation test criterion level. Since the costs of misclassification are low, the logistic models of spatial ability and robotic performance are reliable enough only to be used to customize regular and remedial training. We suggest several changes in tracking performance throughout robotics training that could improve the range and reliability of predictive models.

1. Introduction

1.1. Robotics operations and training

Safe and efficient control of the International Space Station (ISS) or Shuttle robotic arm is heavily dependent on the spatial skills of the operator. Cameras are mounted on the robotic arm and at various locations on the space station exterior to provide visual feedback about the spatial relationship of the arm with the surrounding structure. The camera images are displayed on one of the monitors in the Robotic Workstation from which the astronauts control the robotic arm (Fig. 1). Often, the cameras are not ideally placed to determine clearance distances from a single view, so operators must be able to integrate imagery from multiple camera views to maintain minimum clearance throughout the task. The difference in viewing angles of the workspace can be quite large, leading to conflicting visual feedback such as the arm moving upward in one view but downward in another (e.g., when the views are from cameras mounted on the nadir and zenith sides of the truss and have opposing directions of "up"). The cameras also have pan and tilt capability, thus the spatial relationships between camera views can change during the task. Spatial skills are

* Corresponding author. Tel.: +1 617 253 7758; fax: +1 617 253 8111.
E-mail addresses: amliu@mit.edu (A.M. Liu), coman@mit.edu (C.M. Oman), rgalvan@mit.edu (R. Galvan), natapoff@mit.edu (A. Natapoff).

0094-5765/$ - see front matter © 2012 IAA. Published by Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.actaastro.2012.04.004

Fig. 1. Robotic Workstation (RWS) used on the ISS. The three monitors of the RWS have three different camera views of the workspace and target payloads. The two laptops mounted above the RWS displays show two additional camera views. Photo credit: NASA.

also utilized when operators manually control the robotic arm using the hand controllers. Operators must be able to quickly and correctly mentally transform their hand controller input into the desired arm motion as seen on the different video displays, otherwise they risk moving the arm in a potentially dangerous direction and possibly colliding with the ISS or a space-walking astronaut.

Robotics operations during ISS assembly and resupply have been performed nearly flawlessly by the astronauts, reflecting the skill of the operators developed during hundreds of hours of intensive training. Currently, training begins with a "Generic Robotics Training" (GRT), a 30-h course that teaches the basic procedures, techniques and strategies using a virtual environment with simulated displays of the robotic arm and payloads which are controlled with flight-like hand controllers. The older astronauts who entered the Astronaut Corps before the ISS launch began their training with the Shuttle Robotics course which had a similar course structure but different arm operations. Most of GRT is focused on developing appropriate spatial strategies to choose the appropriate camera views and control frames, correctly perceive arm orientation and position, wisely choose movement strategies to avoid collisions and singularities, and to make appropriate inputs to the hand controllers. If the astronaut has difficulty mastering certain aspects of arm operations (e.g., identifying proper arm clearance or determining the correct hand controller input), additional practice sessions or individual tutoring with instructors or colleagues are prescribed. After completing the lesson sequence, astronauts demonstrate their mastery during a final evaluation where they must complete a series of typical robotics operation tasks within a fixed time. Performance is most heavily evaluated on three skill categories ("General Situation Awareness", "Clearance" and "Maneuvers") that are largely composed of spatial tasks. For example, scores for the General Situation Awareness category are based on the selection of appropriate camera views for the task, recognition of unexpected arm movements, and avoiding arm self-collisions. The Clearance category is evaluated on maintaining proper clearance from structure and proper camera selection for clearance monitoring. The Maneuvers category is evaluated on the astronaut operator's ability to make correct hand controller inputs, selecting the correct control frame for the task and planning a safe but efficient arm trajectory. Performance in the evaluation test is assessed by a Robotics Instructor and an Instructor Astronaut, who has had operational experience with the arm. A score between 1 and 5 is awarded for each skill category, reflecting the performance averaged over the series of tasks that were performed. This minimizes the effect of doing poorly on one of the tasks, although certain errors, such as colliding with structure, incur a mandatory score reduction regardless of overall performance. To pass the evaluation, scores in all categories must generally be 4 or higher although some exceptions can be made depending on the category of the low score. Most astronauts pass the final evaluation and continue with specific Shuttle or Station arm training, although a few have even been unable to qualify to continue with their robotics training. Subsequent training builds on these basic spatial skills and strategies and performance evaluations follow the same basic criteria as in GRT, so complete mastery is necessary to get through all of the training in a timely fashion. The evaluation scores are often used to assign astronauts to specific robotic tasks or roles—astronauts with average robotic skills may be assigned as the primary operator for simple routine operations but only as a secondary operator for more complex operations. The best operators will be assigned as primary operators for the complex tasks or assume primary duties in the case of emergencies.

Given the huge time investment, NASA is interested in improving the efficiency of training by being able to

identify potential problems in developing the necessary spatial skills, then customizing the course of training based on the weaknesses identified. One possible approach to predict training performance would be to assess the spatial ability of the astronaut using simple tests before commencing training. This approach has been adopted in a wide variety of jobs, including the military, but more often for selection rather than for customizing training. However, determining the appropriate set of tests of spatial skill is quite difficult. Human spatial ability is generally not considered to be a unitary construct, but rather composed of separate competences that can be inferred from a mathematical factor analysis of statistical data taken from a battery of appropriate performance tests. Based on the interpretation of the mathematical principal components, 2–7 underlying spatial factors have been defined (see [1–5]). Although the interpretation of the factor analyses and the choice of tests that measures each factor most reliably is debated, there is general agreement that two specific factors that are highly relevant to teleoperation are important. They are *spatial visualization*, the ability to manipulate a spatial mental image into other configurations, and *spatial orientation*, the ability to imagine how a complex object looks after it is rotated. Although equivalent visual image rotations relative to the viewer can be achieved either by mental rotation of the object, or by a corresponding change of the viewer's position in the environment, physiological research has shown that these two transformations, object processing and self-orientation in the environment, are subserved by separate dorsal and ventral streams of visual information. As a result, the "spatial orientation" factor has recently been sub-divided into mental rotation ability (i.e., the ability to mentally rotate objects), and perspective taking ability (i.e., the ability to visualize an array of objects as seen from a novel perspective in the environment) [6]. Spatial orientation ability is clearly heavily utilized in many of the robotics tasks described above and the scoring criteria of the GRT final evaluation. For example, the operator must be able to assume each of the camera perspectives (i.e., perform perspective taking) and imagine how the arm and structure will appear in order to determine if that view is appropriate and useful. While moving a payload on the arm, operators must imagine how it will be rotated into the appropriate position in order to apply the necessary hand controller inputs. It is also possible that operators could rely on either or both strategies to perform the required task, depending on their individual abilities.

The effects of spatial ability on teleoperation performance have been studied in several contexts. For example, Menchaca-Brandan et al. [7] studied performance during simulated grappling and docking tasks with a Space Shuttle-like robotic arm. She correlated performance with the Cube Comparisons Test [8] to assess mental rotation, and with the Purdue Spatial Visualization Test [9] and Perspective Taking Ability Test [10], for perspective taking. The subjects used an interface resembling the Space Station Robotics Workstation with two joysticks and three displays each showing a different camera view. The camera views were manipulated to change the control-axis vs. display-axis disparity and the angular separation between views. This changed the difficulty of integrating separate views. Subjects with higher perspective taking ability exhibited faster and more efficient performance especially for the docking tasks. Another study by Tracey and Lathan [11] studied pick-and-place task performance using a mobile robot and found that subjects with higher spatial ability scores (a composite of the Paper Folding Test and Stumpf's Cube Perspectives Test) completed the tasks faster, with approximately the same number of errors. Since their subjects had only a single view of the task space, the study did not test viewpoint integration from multiple camera views. When using a mobile robot with a single camera, Lathan and Tracey [12] found a correlation between spatial ability (a composite score from the Complex Figures, Stumpf Spatial Memory, Block Rotation, and Stumpf Cube Perspectives Tests) and 2-D maze navigation performance. Eyal and Tendick [13] studied the performance of novice surgeons who were learning to properly position an angled laparoscope. This medical teleoperation task is made difficult by the disparity between the camera display reference frame and the laparoscope control reference frame. They measured spatial ability using the Card Rotation, Paper Folding, and Perspective-Taking tests and found significant correlations between those test scores and performance.

In this study, we tested the general hypothesis that metrics of spatial ability can predict the performance of certain types of spatial tasks used during telerobotic operations. We hypothesized that mental rotation and perspective taking test scores would correlate positively with astronaut performance in the General SA and Clearance categories of their final evaluation test. If these correlations could be established, the tests might be useful for identifying astronauts who would be likely to have trouble with these aspects of robotic training. Their robotics training could then be improved by customization to their personal set of spatial abilities.

2. Materials and methods

The study was approved by the MIT Committee on the Use of Humans as Experimental Subjects (COUHES) and by the NASA-Johnson Space Center Institutional Review Board. De-identified Generic robotics and Shuttle robotics final qualification evaluation scores (both individual category and overall scores) were obtained for 115 active astronauts who had completed at least one training course. Although many astronauts have completed multiple courses, the evaluation scores from only the initial robotics training courses were analyzed. The set of skill categories in the evaluation scoring rubric has changed slightly over time. In the original scoring rubrics for Generic and Shuttle robotics training, the General Situation Awareness category was the only category that primarily involved spatial skills. Subsequently, a Clearance category was added to the rubric to separately evaluate an operator's ability to maintain minimum distances to structure during telerobotic operations. The Maneuvers category was not analyzed, because the scoring

was based on both spatial skills and skills involving bimanual control ability which was not measured. The gender of the astronauts was also included.

We tested the spatial ability of 50 out of the 115 astronauts in groups of 1–5 astronauts during a single 1-h session. The astronauts completed four spatial ability tests that were selected to measure mental rotation and perspective taking abilities in both 2D and 3D. Mental rotation skills are utilized when identifying arm configuration, pre-visualizing maneuvers with payloads attached to the end-effector or recognizing or differentiating structural elements from multiple camera views. Perspective taking skills are needed to interpret multiple camera views and integrate them into a unified representation of the workspace. The spatial tests are described below:

(1) Card Rotation Test ("Card Test")—This is a paper-and-pencil test of 2D mental rotation ability where subjects view a random shape and judge which of eight alternative test figures are planar rotations of that target figure. They had 3 min to complete each of the two parts of the test, with 10 target figures per part. Their score is the number correct minus the number incorrect [8].
(2) Vandenberg Mental Rotation Test ("Vandenberg Test")—This is a paper-and-pencil test of 3D mental rotation ability. Subjects view a three-dimensional target figure and choose which two of the four alternative test figures are rotations of it. They had 3 min to complete each of the two parts, with 10 target figures per part. Their score is the number correct minus the number incorrect [14].
(3) Purdue Spatial Visualization Test: Visualization of Views ("Purdue Test")—This is a paper-and-pencil test of 3D perspective taking ability. The subjects view a three-dimensional target figure surrounded by a "glass cube" and must select which of five alternative test figures represents the view from the designated corner of the glass cube. They had 6 min to complete 30 problems. Their score is the number correct minus one-fourth the number incorrect [9].
(4) Perspective Taking Ability Test ("PTA Test")—This is a self-paced computer-based test of 2D perspective taking ability. Subjects see a top-down plan view of an observer surrounded by an array of seven labeled items (e.g., a school, hotel, airport, etc.). They are told which object they are "facing", then after 5 s, they must indicate the direction of a flashing target object, relative to that specified orientation in the plan view. There are 58 trials and the score is based on the angular error and response time [10].

Statistical analysis was performed using Systat v.13 (Systat Software Inc., San Jose, CA).

3. Results

3.1. Astronaut spatial ability scores

For all four spatial ability tests, the average scores for the astronaut population were generally higher than the average scores from the subject population of our previous experiments at MIT (Table 1). The difference between the scores was significant only for the Purdue Spatial Visualization Test ($t=2.05$, df$=82.25$, $p=0.044$). For both the astronaut and MIT subject populations, we also found that the average scores for males were higher than for females, although we had about twice as many male subjects as female subjects. The differences were statistically significant for all four tests in both groups (t-test, $p=0.035$ or less). The strongest differences were found in the Vandenberg and Purdue tests, followed by the PTA, then the Card Rotation test. The astronaut test scores from all four spatial ability tests were also significantly correlated with each other (Pearson correlation, $\lambda^2=78.35$, df$=6$, $p<0.005$).

3.2. Astronaut robotics evaluation scores

The final evaluation scores for 67 astronauts (54 males) who had taken Generic Robotics Training as their initial training course were analyzed. Table 2 shows the average scores for the General Situation Awareness and Clearance categories with the astronauts grouped by scoring rubric version and gender. For both scoring rubric version and score categories, the average scores were higher for male astronauts, but only the difference between Clearance category scores (Expanded rubric) approached significance ($\lambda^2=0.057$, df$=1$, $p=0.057$, Kruskal–Wallis non-parametric test). Similarly, we did not find any significant effect of gender in the General Situation Awareness scores (Original rubric) for the 27 astronauts (20 males) who first completed the Shuttle Robotics training even though the average scores for males were generally higher.

Table 1
Average spatial ability test scores. Astronaut subjects have generally higher scores than the population tested in previous experiments in our laboratory. Male subjects had significantly higher scores than female subjects in both the astronaut and MIT subject groups.

	Card Test	Vandenberg Test	Purdue Test	PTA Test
Astronauts (combined)	122.0 ± 23.8 ($n=50$)	18.3 ± 9.1 ($n=50$)	17.7 ± 7.2 ($n=50$)	20.4 ± 3.8 ($n=50$)
Male	125.9 ± 24.1 ($n=37$)	21.5 ± 8.2 ($n=37$)	20.0 ± 6.5 ($n=37$)	21.2 ± 3.9 ($n=37$)
Female	110.8 ± 19.8 ($n=13$)	9.5 ± 4.4 ($n=13$)	11.0 ± 4.2 ($n=13$)	18.0 ± 2.3 ($n=13$)
MIT subjects (combined)	114.3 ± 33.2 ($n=70$)	16.4 ± 10.1 ($n=113$)	15.3 ± 6.8 ($n=143$)	20.1 ± 4.8 ($n=258$)
Male	116.1 ± 30.0 ($n=43$)	18.8 ± 10.3 ($n=77$)	16.8 ± 7.0 ($n=95$)	21.2 ± 4.6 ($n=152$)
Female	115.0 ± 36.2 ($n=25$)	11.0 ± 7.7 ($n=34$)	12.2 ± 5.4 ($n=47$)	18.5 ± 4.8 ($n=105$)

Table 2
Robotics Evaluation Performance. Male astronauts had a better average performance score when compared to the female astronauts, although only the difference for General Situation Awareness in Generic robotics training was significant. The original version of the scoring rubric did not include a separate Clearance category.

	Generic Robotics Training			Shuttle Robotics Training	
	Original rubric	Expanded rubric		Original rubric	Expanded rubric
	General Situation Awareness	General Situation Awareness	Clearance	General Situation Awareness	General Situation Awareness
Overall	4.23 ± 0.63 ($n=24$)	4.54 ± 0.66 ($n=43$)	4.37 ± 0.57 ($n=43$)	4.05 ± 1.01 ($n=19$)	4.31 ± 0.75 ($n=8$)
Male	4.28 ± 0.53 ($n=20$)	4.57 ± 0.55 ($n=34$)	4.47 ± 0.49 ($n=34$)	4.27 ± 0.73 ($n=13$)	4.50 ± 0.58 ($n=7$)
Female	4.00 ± 1.1 ($n=4$)	4.39 ± 0.99 ($n=9$)	4.00 ± 0.71 ($n=9$)	3.58 ± 1.42 ($n=6$)	3.00 ($n=1$)

3.3. Spatial ability tests of as predictors of robotics training performance

We had both spatial ability and robotics training data for 46 astronauts: 36 astronauts who completed Generic training as their first course and 10 astronauts who completed Shuttle training first. Most astronauts scored between 4 and 5 in the final evaluation skill categories since they are trained to a criterion level of performance in the preceding lessons. This range compression of the dependent variable makes a conventional linear regression model inapplicable.

Instead, we performed a logistic regression analysis of each of the spatial ability test scores for the final evaluation performance. The performance scores were categorized as "high" if scores were greater than 4 and "average" for scores less than 4. The performance data was grouped by the type of initial robotics training (Generic or Shuttle), the evaluation category (General Situation Awareness or Clearance) and the scoring rubric (Original or Expanded). Spatial ability scores were standardized to mean 0 and standard deviation 1. A logistic regression model that was an ideal predictor of robotics training performance would have a sharp transition in the probability of achieving a high score at the threshold spatial test score, e.g., a step function from 0.0 to 1.0 probability at the criterion score. However, the models typically have a less steep probability transition indicating some errors or variability in classification performance. A suitable classification threshold can be selected based on the receiver operating characteristic (ROC) curve, which shows the tradeoff between the probability of a false positive result, (1—specificity) (i.e., classifying an average scorer as a high scorer) and the probability of correctly identifying the high scorers (sensitivity), and the costs associated with those decisions. The classification performance of different logistic models can be compared by calculating the area under the ROC curve (AUC) [15]. The AUC value for a perfect classifier would equal 1.0 whereas an AUC value of 0.5 would indicate random classification performance or no discriminative power. In clinical applications, AUC values less than 0.75 are generally not useful while values above 0.97 have high clinical value [16].

No significant model fits were found for Generic training data with the original scoring rubric ($n=8$), although the fit of the Purdue test with the General Situation Awareness category was nearly significant ($\lambda^2=3.58$, $p=0.06$). The small sample size could be a factor in the lack of significant fit. Peduzzi et al. [17] have suggested that the number of events (i.e., high scoring astronauts) per predictive variable (i.e., one of the spatial ability tests) should be >10 to avoid potential problems with model validity. In this astronaut group, five of the eight astronauts achieved the highest score.

From the Generic robotics training data using the expanded scoring rubric ($n=28$), significant logistic model fits were found for the Card ($p=0.03$) and the PTA ($\lambda^2=4.54$, $p=0.03$) tests with the General Situation Awareness category and the Vandenberg test ($\lambda^2=4.37$, $p=0.037$) with the Clearance category. We have overlaid the logistic curve on a scatterplot of the astronauts' performance on the General Situation Awareness (Fig. 2a and b) or Clearance (Fig. 2c) categories vs. spatial ability test score. The circles on the top border of the plots represent astronauts with high scores, while the circles on the bottom border are the average scorers. The raw data shows that in many cases, astronauts with low spatial ability test scores (e.g., more than one standard deviation below the mean) can still achieve high performance, while in fewer cases, astronauts with high spatial test scores (e.g., more than one standard deviation above the mean) only manage average task performance. Even for spatial test scores more than 3 standard deviations below the mean, there is a small and decreasing probability that they will still achieve a high score. The gentle slope of the curve indicates the large overlap in the distribution of scores of the two groups. The ROC curves (Fig. 3) show that a threshold test score with 80% correct classification will also lead to a 30–40% false positive rate. In two of the three cases, the AUC values, shown below each plot, are slightly less than the desired 0.75 which suggests that discriminatory power of the models may be limited.

For Shuttle training with the original scoring criteria ($n=7$), significant model fits were found for the Card ($\lambda^2=3.78$, $p=0.05$), the Vandenberg ($\lambda^2=9.56$, $p=0.002$), and the Purdue ($\lambda^2=3.74$, $p=0.05$) tests with the General Situation Awareness category. The distributions of the high and average scorers have very little overlap which leads to a sharp transition in the logistic curves (Fig. 4). The ROC curves and associated AUC values, which are all above 0.83, reflect this clear differentiation between

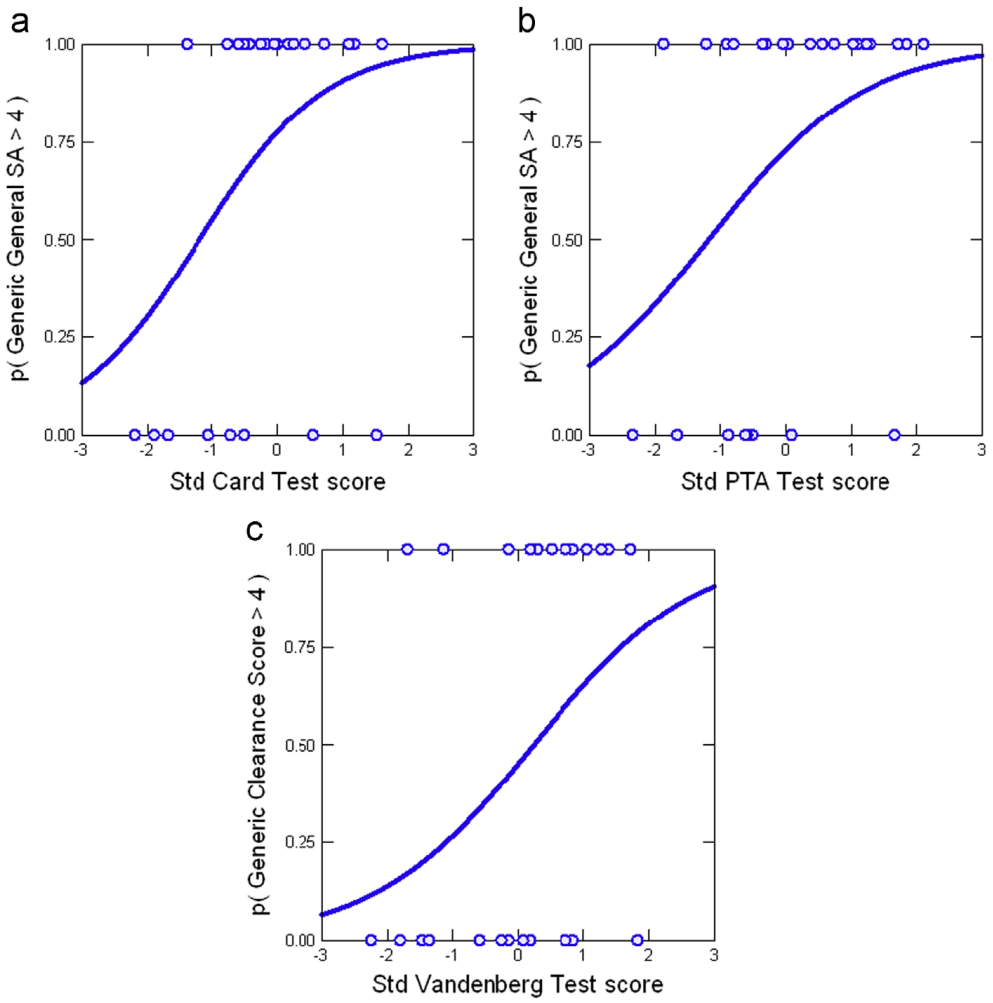


Fig. 2. Logistic curves of the probability of achieving a high evaluation score (Expanded rubric) after Generic training as predicted by spatial ability test score. The circles represent the raw performance data with high scorers at the top and average scorers at the bottom. Top left: Card test, General Situation Awareness category, top right: PTA test, General Situation Awareness category and bottom: Vandenberg test, Clearance category.

groups (Fig. 5). In the case of the Vandenberg Test, classification of the astronauts was perfect with a threshold score just above the mean and AUC=1.0. Again, the significance of these results must be tempered by noting the small number of events: in this astronaut grouping, we had only 3 high scoring astronauts and 4 that had average scores. The Shuttle Robotics data using the expanded scoring criteria category was not analyzed because of the small sample size (n=3).

General SA and Clearance categories both describe spatial tasks that presumably could utilize both mental rotation and perspective taking skills and, thus, a combination of predictor scores might model the data more closely. We therefore repeated the logistic regression using linear combinations of the mental rotation and perspective taking tests as the independent variables and the GRT General SA and Clearance scores as the dependent variables. None of these models produced a better fit to our data than a model with a single independent variable. This is not a surprising result since the spatial ability test scores were highly correlated with one another.

4. Discussion

We have indirectly shown that a logistic regression model can characterize the relationship between spatial ability test performance and performance in two final evaluation categories. However, the predictive models either have too high a misclassification rate or have uncertain reliability because of the small sample size to be used for astronaut selection or career decisions where costs are high. Yet, the predictors could still be a useful tool to determine when to schedule astronauts for robotics training and how much time to allot to training. For example, it may be useful to admit astronauts with lower spatial ability scores to robotics training earlier to allow them the flexibility to repeat a lesson or take additional self-study time to acquire the necessary skills. Astronauts with better spatial skills could be scheduled more flexibly into shorter time windows since they are more likely to do well. By incorporating information about the number of hours spent in training and self-study, or performance in the initial lessons of training, models that predict the likely difficult lessons or topics for

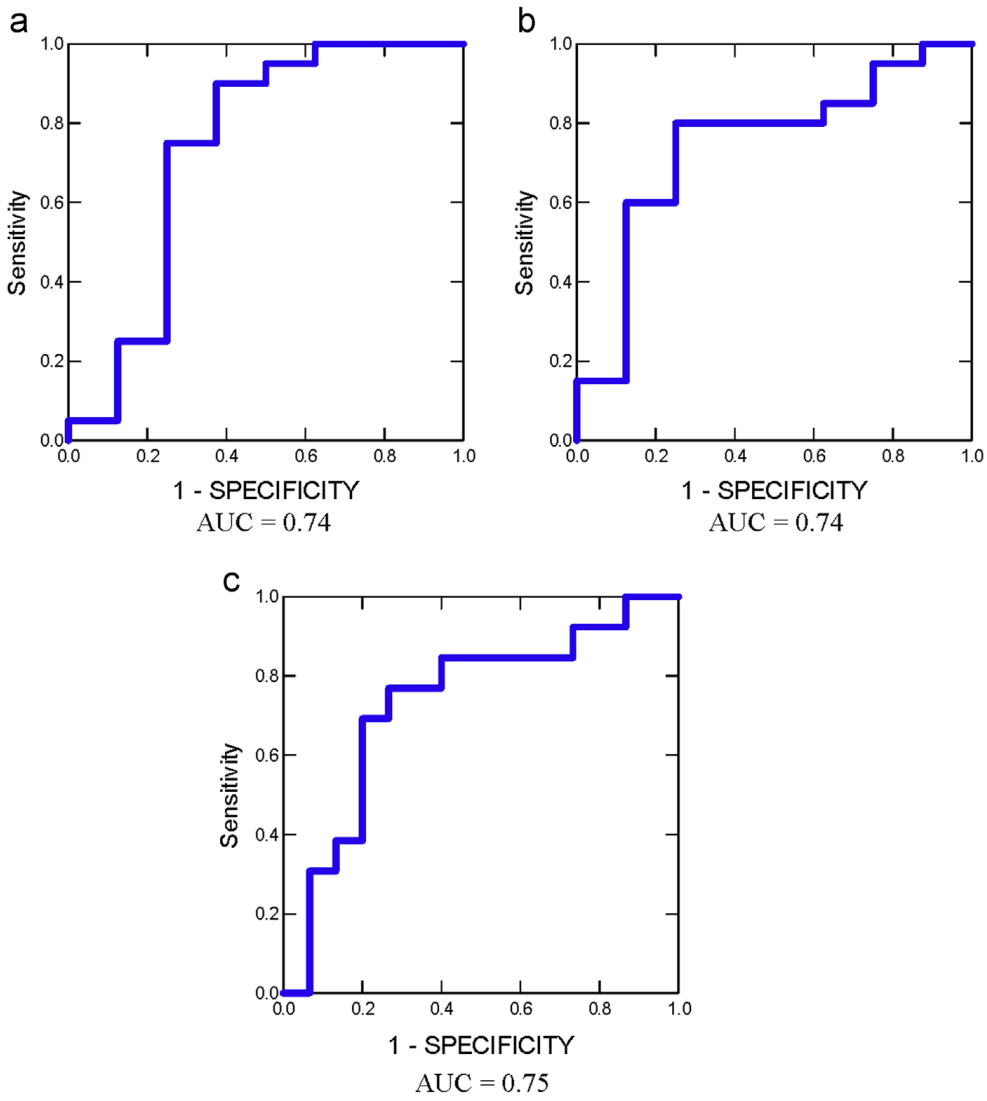


Fig. 3. ROC curves for the three logistic models predicting performance in Generic training, General Situation Awareness category, Expanded rubric (a,b) and Clearance category (c). Sensitivity indicates the probability of a correct prediction. (1—specificity) indicates the probability of incorrectly predicting high performance.

astronauts with below-threshold spatial ability test scores could be developed to support customization of training lessons. Unfortunately, individual category scores from the early lessons (e.g., Visualization Skills or Kinematic Skills) of Generic training have not been saved for most of the astronauts. Other possible metrics, such as the number of remedial lessons required or the total number of hours spent in training to achieve criterion performance (which could characterize the arc of progress of the successful candidates and distinguish the success of the training that contributed to their rank) have also not been kept.

There are a number of other aspects of the robotics evaluation process for which information has been lost potentially adding undesirable and extraneous variability to our measures. During a final evaluation, astronauts complete a series of tasks under different operating conditions (e.g., berthing with internal command frames, fly-to's with external command frames, etc.) but their performance on these different tasks is averaged into a single category score. Therefore, differences related to individual conditions, such as which command frame was being used or which camera viewpoints provided feedback, cannot be extracted from the final score. If instead, scores on meaningful "classes" of tasks were recorded, it would be possible to perform a more detailed analysis of training performance. Also, despite the defined criteria, an instructor's evaluation score is ultimately subjective. Individual differences in the subjective scoring based on personal expectations of the trainer contribute additional variability for which we cannot correct with the information currently available. Further, since the cadre of Robotics Instructors and Instructor Astronauts changes over time, the general scoring heuristics have likely not remained constant. Even for the core groups of senior instructors who have worked for many years, scoring standards have shifted as they became more attuned to average performance levels within the astronauts. (P. Williamson, personal communication). The current database does specify which trainers and Instructor Astronauts scored a particular

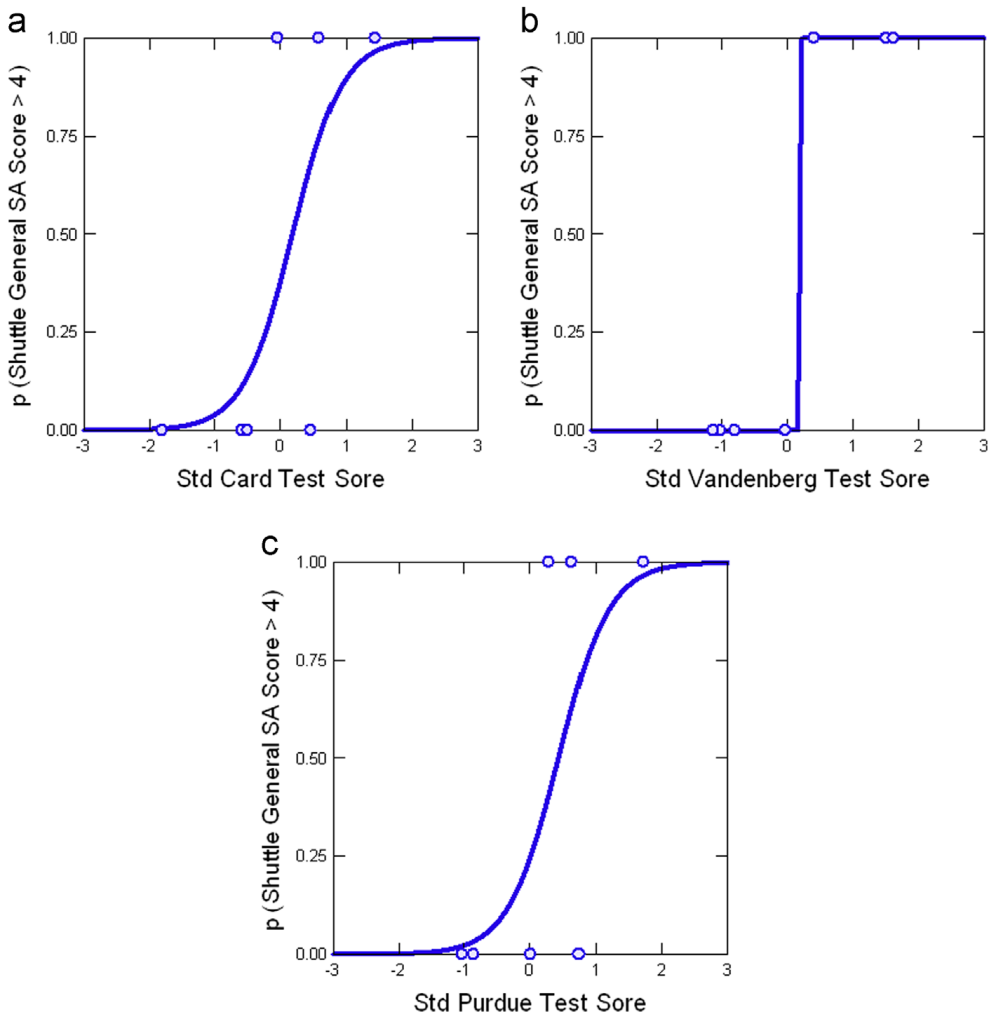


Fig. 4. Logistic curves of the probability of achieving a high General Situation Awareness score (Original rubric) after Shuttle robotics training as predicted by spatial ability test score. The circles represent the raw performance data with high scorers at the top and average scorers at the bottom. Top left: Card Test, top right: Vandenberg Test and bottom: Purdue Test.

astronaut's evaluation. If the spatial ability of the instructors is also assessed, then that data could be used to make more refined estimates of the effect of underlying abilities in any future analysis. The fact that several logistic regression models give statistically significant fits to the data despite these many sources of variability suggests that if more detailed training data can be collected from future astronauts, better predictions about their progress through training can be made, and the entire training process can be made more efficient.

Other psychometric factors that have been proposed as spatial ability metrics including *closure speed* (ability to rapidly access representations from long-term memory), *flexibility of closure* (ability to maintain the representation of an object in visual working memory while trying to distinguish it in a complex pattern), *perceptual speed* (ability to rapidly compare or find symbols or figures), and *visual memory* (ability to remember the spatial distribution of objects or figures). There are many aspects of the teleoperation task that relate to these abilities. Psychometric tests of both visual and verbal working memory span have been developed, e.g., Shah and Miyake [18]. Most traditional spatial ability tests are static and do not involve motion assessment. Recently, another set of dynamic spatial-ability factors has been suggested [19,20] involving ability to perceive and extrapolate visual motion, predict trajectories and estimate arrival time. These psychometric factors will be considered in future experiments.

5. Conclusions

In this study, we show via logistic regression of four different types of spatial ability test scores against final evaluation spatial skill category scores, that we can identify top performers at a statistically significant level. None of the spatial tests had a statistically significant fit to all performance measures but the Card Test and Purdue Test had significant fits to the General Situation Awareness (Generic and Shuttle robotics Training). The classification performance of these logistic models, measured by an area-under-the-curve (AUC) metric, is reliable enough to help plan training schedules or decide if extra or remedial training is required, but not reliable enough to support career-defining decisions. We have suggested to our colleagues in the Robotics Branch of the NASA

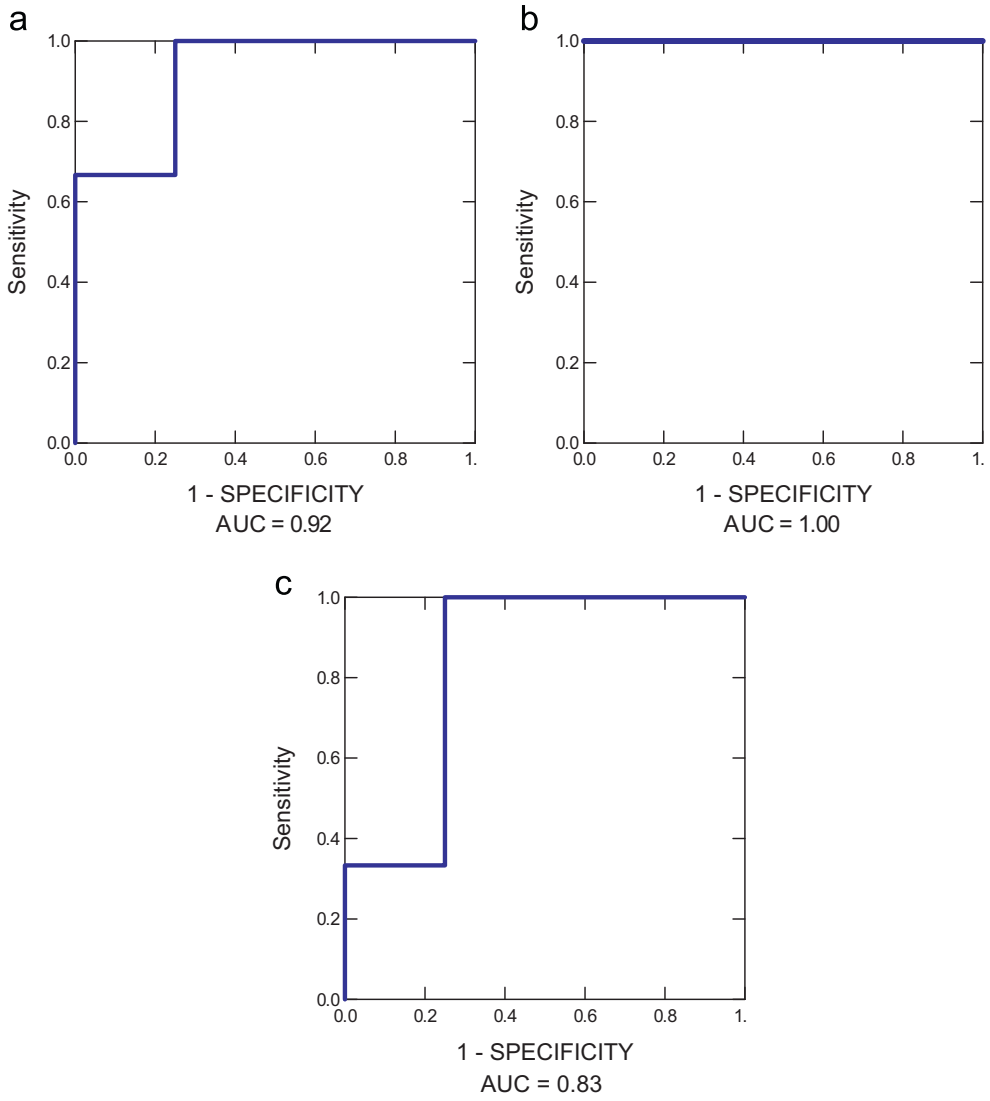


Fig. 5. Receiver Operating Characteristic curves for the 3 logistic models predicting performance in Shuttle training, General Situation Awareness category and Original rubric. The AUC values for the three models exceed the desired 0.75 level. Sensitivity indicates the probability of a correct prediction. (1—specificity) indicates the probability of incorrectly predicting high performance.

Astronaut Office that tracking the performance in early training lessons with a broader range of objective measures (e.g., time needed to complete lessons) will be necessary to improve the predictive capabilities of the models. Other tests of working memory and dynamic spatial ability should be evaluated for their correlation with other task performance or hand-controller categories.

Although Space Station assembly operations are almost complete, astronauts will continue to perform many telerobotic operations, e.g., inspection and maintenance or free-flyer capture of resupply vessels such as the Japanese H2 Transfer Vehicle. Training for the latter task is currently under development, so customized training could be very helpful to improve the efficiency of training. As astronauts participate in missions of increasing duration, more in-flight training for unplanned operations will likely be needed. In these cases, training that is tailored for a specific astronaut operator will be faster and better assimilated, and should reduce the risk of errors during operations.

Acknowledgments

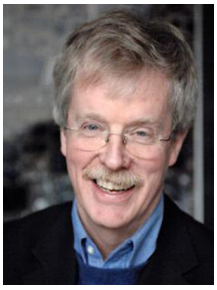
This research was performed under National Space Biomedical Research Institute contract SA001, and supported by NASA Cooperative Agreement NCC9-58. We thank our NASA collaborators James Tinch and Heidi Jennings for their help collecting astronaut data and Cady Coleman, Jennifer Young, Brett Levins, and Paul Williamson for their helpful discussions about robotics training, and the astronauts who participated in this experiment.

References

[1] J.P. Carroll, Human Cognitive Abilities: A Survey of Factor-Analytical Studies, Cambridge University Press, New York, 1993.

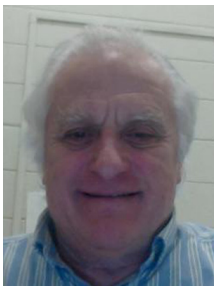
[2] M. Hegarty, D. Waller, Individual differences in spatial ability, in: P. Shah, A. Miyake (Eds.), The Cambridge Handbook on Visuospatial Thinking, Cambridge University Press, New York, 2005, pp. 121–169.
[3] D.F. Lohman, Spatial abilities as traits, processes, and knowledge, in: J. Sternberg (Ed.), Advances in the Psychology of Human Intelligence, Earlbaum, Hillsdale, NY, 1988, pp. 181–248.
[4] M.G. McGee, Human spatial abilities: psychometric studies and environmental, genetic, hormonal and neurological influences, Psychol. Bull. 86 (1979) 889–918.
[5] W.B. Michael, G.P. Guilford, B. Frutcher, W.S. Zimmerman, The description of spatial-visualization abilities, Educ. Psychol. Meas. 17 (1957) 185–199.
[6] M. Kozhevnikov, M. Hegarty, A dissociation between object manipulation spatial ability and spatial orientation ability, Mem. Cognit. 29 (2001) 745–756.
[7] M.A. Menchaca-Brandan, A.M. Liu, C.M. Oman, A. Natapoff, Influence of perspective-taking and mental rotation abilities in space teleoperation, in: Proceedings of the 2007 ACM/IEEE International Conference on Human–Robot Interaction, Washington, DC, 2007.
[8] R.B. Ekstrom, J.W. French, H.H. Hartman, Manual for Kit of Factor Referenced Cognitive Tests, Educational Testing Service, Princeton, NJ, 1976.
[9] R. Guay, Purdue Spatial Visualization Test—Visualization of Views, Purdue Research Foundation, West Lafayette, IN, 1977.
[10] M. Kozhevnikov, M.A. Motes, B. Rasch, O. Blajekova, Perspective-taking vs. mental rotation transformations and how they predict spatial navigation performance, Appl. Cognit. Psychol. 30 (2006) 397–417.
[11] M.R. Tracey, C.E. Lathan, The interaction of spatial ability and motor learning in the transfer of training from a simulator to a real task, in: J.D. Westwood, et al., (Eds.), Medicine Meets Virtual Reality, IOS Press, Amsterdam, 2001, pp. 521–527.
[12] C.E. Lathan, M. Tracey, The effects of operator spatial perception and sensory feedback on human–robot teleoperation performance presence: teleoperators, Virtual Environ. 11 (2002) 368–377.
[13] R. Eyal, F. Tendick, Spatial ability and learning the use of an angled laparoscope in a virtual environment, in: J.D. Westwood, et al., (Eds.), Medicine Meets Virtual Reality, IOS Press, Amsterdam, 2001, pp. 146–152.
[14] S.G. Vandenberg, A.R. Kuse, Mental rotations, a group test of three-dimensional spatial visualization, Percept. Mot. Skills 47 (1978) 599–604.
[15] T. Fawcett, An introduction to ROC graph analysis, Pattern Recognition Lett. 27 (2006) 861–874.
[16] J. Fan, S. Upadhye, A. Worster, Understanding receiver operating characteristic (ROC) curves, Can. J. Emerg. Med. 8 (2006) 19–20.
[17] P. Peduzzi, J. Concato, E. Kemper, T.R. Holford, A.R. Feinstein, A simulation study of the number of events per variable in logistic regression analysis, J. Clin. Epidemiol. 49 (1996) 1373–1379.
[18] P. Shah, A. Miyake, The separability of working memory resources for spatial thinking and language processing: an individual differences approach, J. Exp. Psychol.: General 125 (1996) 4–27.
[19] M.J. Contreras, R. Colom, J.M. Hernandez, J. Santacreu, Is static spatial performance distinguishable from dynamic spatial performance? A latent-variable analysis, J. Gen. Psychol. 130 (2003) 277–288.
[20] J.W. Pellegrino, E.B. Hunt, R. Abate, S. Farr, A computer-based test battery for the assessment of static and dynamic spatial reasoning abilities, Behav. Res. Methods Instrum. Comput. 19 (1987) 231–236.

Andrew M. Liu is a research scientist in the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology. His current research interests are human–machine and human–automation interaction, human spaceflight, spatial cognition, and fatigue. He is an investigator on projects sponsored by the National Space Biomedical Research Institute, Nuclear Regulatory Commission, and Federal Rail Administration. He received a B.S. in Biomedical Engineering from the Johns Hopkins University and a Ph.D. in Bioengineering from the University of California, Berkeley.



Charles M. Oman is a senior research engineer and senior lecturer in the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology and the current director of the Man Vehicle Laboratory. His research addresses the physiological and cognitive limitations of humans in aircraft and spacecraft. He has conducted numerous experiments on Shuttle/Spacelab. He currently leads the Sensorimotor Adaptation Team of the National Space Biomedical Research Institute, is PI on two NSBRI projects and co-I on two others. He holds a B.S. in Aerospace and Mechanical Sciences from Princeton University and a M.S. and Ph.D. in Instrumentation and Control from the MIT.

Raquel Galvan is a graduate research assistant in the Department of Aeronautics and Astronautics and Man Vehicle Laboratory at the Massachusetts Institute of Technology. She is currently conducting research sponsored by the National Space Biomedical Research Institute on the effects of sleep restriction and time shifting on cognitive and space teleoperation performance. She received her B.S. in Aerospace Engineering from the University of Texas, Austin.



Alan Natapoff is a Research Scientist in the Department of Aeronautics and Astronautics at the Massachusetts Institute of Technology. He specializes in the statistical design and analysis. He received his Ph.D. in Physics from the University of California, Berkeley.