Human Factors in Design, Engineering, and Computing, Vol. 238, 2026, 2760–2769
https://doi.org/10.54941/ahfe1008500

Terrain Orientation Test Predicts Navy Flight Training Outcomes in High and Low Stakes Settings

Joseph T. Coyne[1], Ciara Sibley[1], Kaylin Strong[1], and Michael Kukenberger[2,3]

[1]Naval Research Laboratory, Washington, DC 23505, USA
[2]Naval Aerospace Medical Institute, Pensacola, FL 32508, USA
[3]University of New Hampshire, Durham, NH 03824, USA

ABSTRACT

This study evaluates the predictive validity of the U.S. Navy's Terrain Orientation Task (TOT), a spatial ability test for pilot selection introduced in 2023. The new test was developed after a previous one lost its effectiveness, partly due to applicants learning test-taking "tricks." However, a replica of the new TOT appeared online shortly after its official rollout, raising concerns about its continued validity. To investigate this, researchers studied 200 flight students who had already taken the TOT as part of their official Aviation Selection Test Battery (ASTB). In a low-stakes lab setting, the students completed the TOT again, practiced it extensively, and then took an alternate version. The results showed that scores from the official high-stakes test and both low-stakes lab attempts were all significantly correlated with the students' flight training grades. While practice did improve scores, the test remained a valid predictor of training outcomes. The study concludes that the TOT is a robust predictor of pilot performance in both high-and low-stakes settings, even with replicas available online. The authors recommend continued monitoring and leveraging the TOT's ability to generate new items to maintain its long-term integrity.

Keywords: Spatial ability, Individual differences, Pilot selection

INTRODUCTION

The Aviation Selection Test Battery (ASTB), is used by the United States Navy, Marine Corps, and Coast Guard to identify pilots and flight officers. Each year several thousand applicants take the battery which is comprised of measures of crystalized intelligence, (subtests for math, verbal, mechanical, and aviation/nautical knowledge) fluid ability measures (e.g., psychomotor, attention, spatial and multi-tasking abilities), and personality. The emphasis of this paper is on the newest subtest in the ASTB, the terrain orientation task (TOT) (Coyne et al., 2024) which was introduced to the ASTB in 2023. The ASTB is a high stakes test, and motivated applicants can find an abundance of unofficial test preparation material. Implementation of the TOT illustrates this challenge, as a replica of the task (with different items) was available on an unofficial test preparation app within three weeks of its addition to the ASTB.

Received October 8, 2026; Revised November 7, 2026; Accepted November 18, 2026; Available online December 1, 2026

© 2026 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

The TOT is designed to measure spatial ability, and more specifically spatial orientation. Spatial ability is defined as the capacity to generate, retain, retrieve, and transform well-structured visual images (Lohman, 1996). Spatial ability measures have been used in military selection since World War 1 (see Damos, 2011 for a history of aviation selection). Barron and Rose (2013) discuss how different aspects of spatial ability (spatial orientation, visualization, processing speed, and closure) relate to tasks performed by pilots. Their analysis showed that all but closure added incremental validity to the prediction of military flight training performance.

Ackerman (2022) describes spatial orientation as a higher-level process-based measure. Ackerman points out that there are limitations in assessing individual difference with these types of process based measures, as they require some degree of novelty. He argues that familiarity with the stimuli or the process being measured can limit the tasks' ability to measure meaningful differences. Coyne et al. (2025) demonstrate that some measures of spatial ability may be especially vulnerable in high stakes settings like the ASTB. The sharing of non-spatial strategies, and easy access to replicas of the direction orientation task (DOT), a spatial ability subtest in the ASTB that was introduced in 2013, ultimately led to the DOT's loss of both predictive validity (Coyne et al., 2022) and construct validity (Coyne et al., 2025). The loss of construct validity for the DOT was interesting since it only appeared with the high-stakes data, and not the low stakes lab-data. That is ASTB DOT data from a large group of flight students (N = 200) did not correlate with DOT data collected in a lab setting for the same individuals. Several participants in that study indicated that they had forgotten the non-spatial "trick" to solving the task when given it again in the lab. The lab DOT data correlated with spatial ability data from several different tasks including the TOT. The lab data collected when participants were not prepared and likely less familiar with the task loaded on to a spatial ability construct whereas the high-stakes ASTB data did not.

The TOT was developed to overcome some of the practical limitations of the DOT, by allowing for an infinite number of new items, the ability to manipulate difficulty, and making the use of non-spatial strategies more difficult by requiring individuals to visually compare the two images within the TOT to solve each item (Coyne et al., 2024). The essentially unlimited item bank allows for the establishment of multiple equated forms which helps enhance test security. The present study is focused on comparing high (ASTB) and low stakes (lab) TOT data from a large group of flight students. The study is also interested in how repeated exposure to the TOT will impact both scores and the ability to predict training outcomes.

The study included other measures of attention control, which are also being evaluated for future inclusion in the ASTB. Georgia Technical University (Burgoyne et al., 2023) developed three short measures of attention control which were built on classical attention tasks such as the Stroop task and Flanker task. These new tests were shown to have good

psychometric properties, specifically with respect to reliability and construct validity. Poor psychometric properties from tasks using difference scores had been an issue with a number of attention tasks (Draheim et al., 2019).

From a selection standpoint the tasks appear to measure meaningful individual differences in a short time period. Across multiple studies the tests have been shown to correlate with training outcomes in Naval Flight Students (Coyne et al., 2025, 2024) as well as Naval Air Traffic Control students (Coyne et al., 2025B).

The objective of this study is to assess TOT across both high and low stakes settings to determine whether the test retains its predictive validity. This will be done by comparing TOT scores from the ASTB with TOT scores for the same students both before and after repeated practice on the task. A secondary objective is to continue to assess the effectiveness of the new short attention control measures.

METHOD

A total of 362 US Sailors and Marines completed a lab-based cognitive assessment battery. All participants were training at Naval Air Station Pensacola and came from multiple schoolhouses. The sample included Naval Flight Students (NFS), as well as enlisted students across multiple Navy ratings. Present analyses include both enlisted and officers, however the outcome data and high stakes ASTB TOT data is limited to the 200 NFS who participated in the study. The training outcome data is from the Naval Introductory Flight Evaluation (NIFE), a combination of ground school and preliminary flight training completed before primary flight training. The study was approved by the institutional review board at the Naval Research Laboratory and lab data was collected between May 2025 and November 2025.

Procedure

The study included a demographics survey, eye tracking setup and baseline, an initial assessment of the TOT, a survey on gaming experience and ASTB preparation (for flight students only), a randomized practice of the TOT, a mixed attention control task, followed by a final post training assessment of the TOT.

Terrain Orientation Task (Pre and Post Training)

The Terrain Orientation Task (TOT) is designed to be similar to a terrain association task. The objective is to determine the direction a vehicle is traveling by comparing a reference map with North on the top, with a rotated image of the same area. Participants are told the second image is from a downward facing camera that is attached to an aircraft, and their objective is to determine the direction the vehicle is traveling based

upon the orientation of the camera image. Similar to terrain association, the participants must identify common features in both the reference and camera map in order to identify the aircraft's heading. The images used were taken from two of the ASTB's TOT forms. One form was used for pre training testing and the other for post training testing. Figure 1 shows how a TOT item would appear during the lab testing, but was not an item that was used in data collection.

The pre and post assessments required participants to complete 12 practice items in which feedback is given on every item and they are notified of the correct response option. Following the practice, participants completed 18 scored items. Participants were informed of whether their response was correct or incorrect, but not informed about which option was correct.

Participants had 36 seconds to respond for each item. This insured that participants would all see every item.

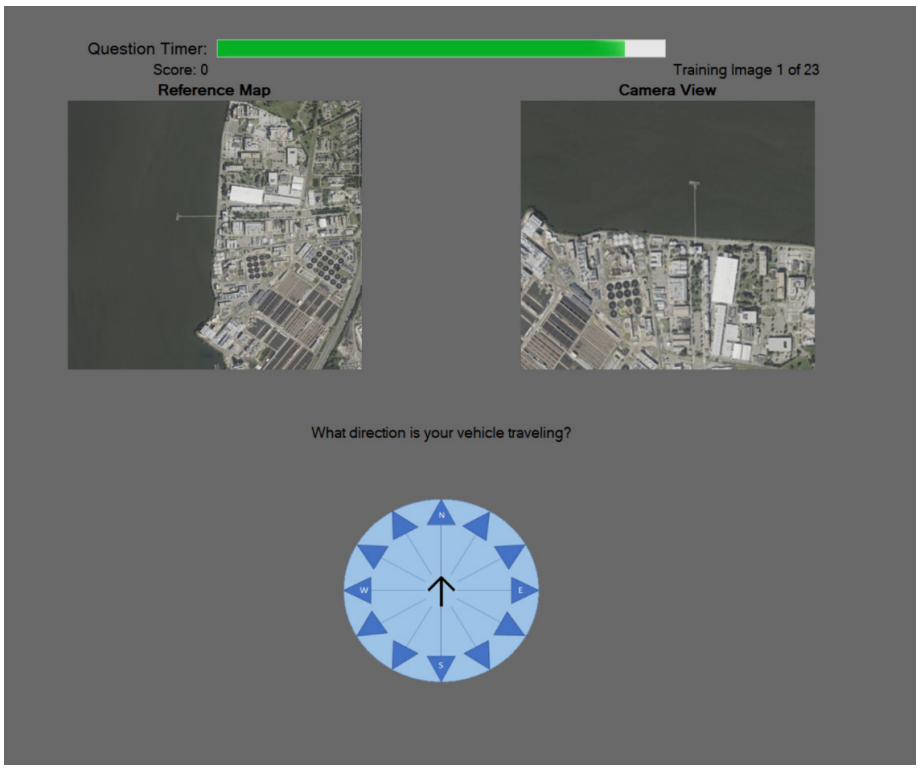


Figure 1: Depiction of a sample item from the TOT. The image on the left is the reference map that is oriented with North at the top. The camera view on the right has rotated 90 degrees to the right. The correct response would be 270 Degrees.

ASTB Preparation Questionnaire

The Naval Flight Students in the sample completed a questionnaire which asked them what resources they used in preparing for the ASTB, and how

useful they found those resources. It also asked whether they had the TOT on any of their official ASTB attempts, and how many hours they spent practicing the TOT, followed by an open-ended question asking how they practiced for the TOT if they previously reported doing so.

TOT Training

The TOT training was almost identical to the TOT. The main difference was that every trial provided feedback indicating what the correct response was by overlaying a green arrow on reference map with the direction of travel and a red arrow with the correct direction on the compass when wrong and a green arrow when they were correct. Additionally, individual's had nine minutes to answer each training item. There were 36 trials in the training. The training used a set of maps that were not from the ASTB. The training was also fully randomized such that trials were randomly generated based upon the map and 12 possible headings. Thus, no two participants received the same set of practice trials.

Attention Control Tasks

The study used an altered version of the Stroop and Flanker tasks described in Burgoyne et al. (2023). The scored trials were the same. That is trials the presence of conflict in both the stimulus and the response were altered creating four stimulus response conflict (congruency) pairs.

For Stroop the task participants were instructed to match the font color (red or blue) of the stimulus with the color's word/name in the response. Conflict was present in either stimulus or response when the font color did not match the word.

For the Flanker task each stimulus and response option contained 5 arrows. The arrows were either all in the same direction (congruent) or the middle arrow was in the opposite direction (incongruent/conflict). The participants were instructed to match the direction of the outside arrows of the stimulus with the middle arrow of the response options.

Participants practiced the Stroop task for 30 seconds, followed by practicing Flanker for 30 seconds, followed by practicing a version in which the trials were intermixed in a block. Unlike the version used in Burgoyne et al., anytime participants made an incorrect response during practice the timer paused, and they were reminded of the task rules. Following the three practices participants completed two 90 second scored testing phases in which Sroop and Flanker trials were mixed together. The original conflict trials in Burgoyne et al. did not mix trial types. Participants were given 1 point for a correct response and lost a point for an incorrect response. Thus, they were rewarded for both speed and accuracy. The objective was to get the most points in each 90 second window. Figure 2 depicts a sample trial from the two task types.

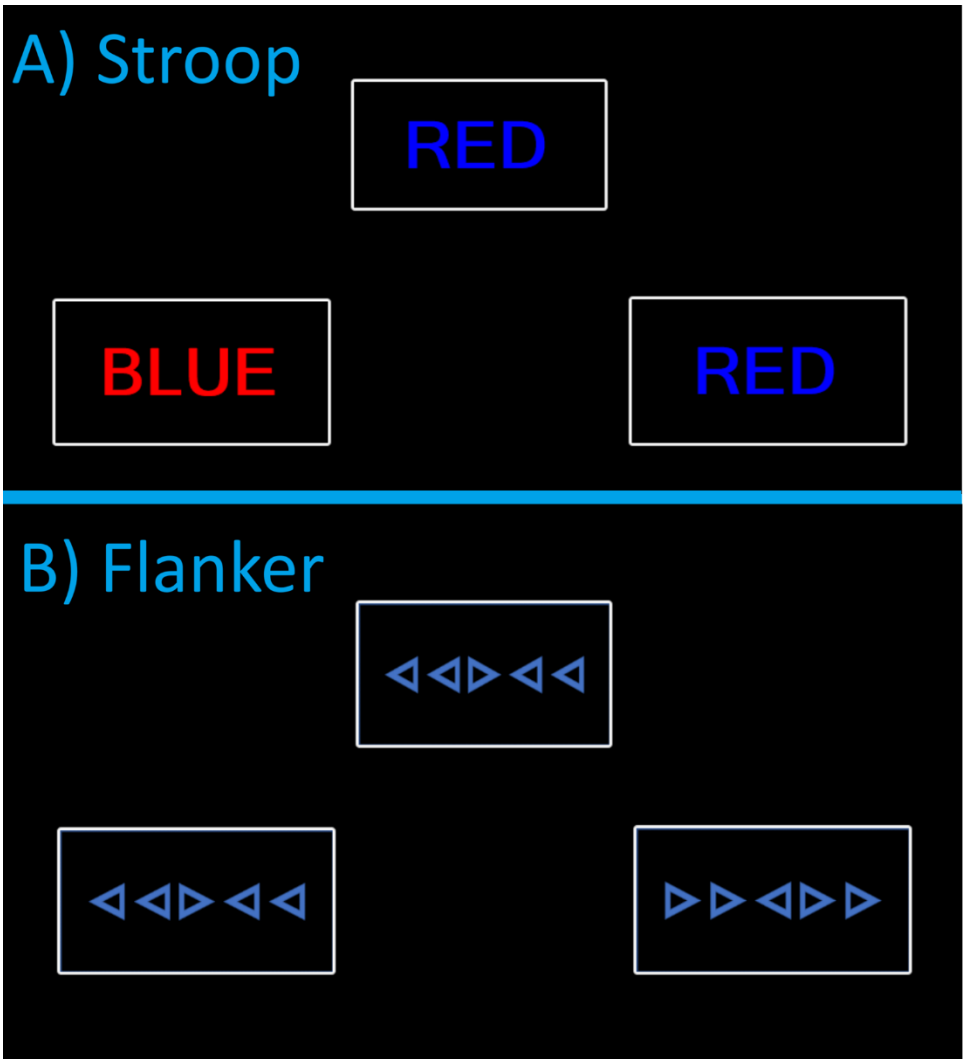


Figure 2: Depiction of the two attention control trial types. Both sample trials are fully incongruent.

RESULTS

Lab Performance Dependent Variables

The dependent variable for the TOT tasks was total number of correct items. The dependent variable from the two attention control tasks was the score (+1 point for a correct response; –1 point for an incorrect response) at the end of each 90 second testing block.

NIFE and ASTB Performance

For the 200 flight students, training outcome data as well as data from each student's final ASTB score were linked with the data from the lab. All flight students had a NIFE outcome; however, we were unable to obtain 1 student's

ASTB scores. NIFE training outcome variables include attrition, academic grade, and number of failed exams. Attrition was a binary completion variable, academic grade was the average grade across the academic components of NIFE, number of fails (0-2) were the number of academic components which had to be repeated. The ASTB data used in the analysis was the number of correct TOT items.

Of the 200 SNA/SNFOs who completed the study nine attrited from NIFE for reasons unrelated to performance (e.g., medical, legal); these students were not included in the attrition analysis. Thirteen attrited for academic reasons (i.e., failures or motivation/drop on request). Students who attrite do not have a final NIFE grade for analysis. The overall attrition rate was 11.0% and the academic attrition rate was 6.5%. Table 1 Presents means and standard deviations for all of the experiment and ASTB data, broken out by student type and ASTB TOT exposure.

Table 1: Means and standard deviations for the TOT and attention control tasks by student type and TOT ASTB exposure.

	Enlisted			No Prior TOT			ASTB TOT			Everyone		
	Mean	SD	N	Mean	SD	N	Mean	SD	N	Mean	SD	N
ASTB TOT							11.29	3.73	145			
TOT Pre Test	6.64	3.43	161	10.63	3.43	56	10.07	3.67	143	8.62	3.95	360
TOT Post Test	8.52	3.89	152	12.30	2.90	56	11.50	3.59	142	10.33	3.96	350
Mixed AC 1	17.49	11.50	159	31.38	11.95	56	27.97	11.46	144	23.86	12.90	359
Mixed AC 2	21.24	12.27	158	33.84	12.00	55	32.29	10.74	143	27.63	12.94	356

Correlations between the tests from the experiment, NIFE training outcomes and the ASTB are presented in Table 2. The correlations presented are not corrected for range restriction.

Table 2: Correlation table of TOT, attention control, NIFE training outcomes and ASTB.

	Grade	Setbacks	Attrite	ASTB TOT	TOT Pre	TOT Post	AC 1	AC 2
NIFE Grade								
NIFE Setbacks	–0.13							
NIFE Attrite	NA	0.57***						
ASTB TOT	0.21*	0.03	0.01					

(Continued)

Table 2: Continued.

	Grade	Setbacks	Attrite	ASTB TOT	TOT Pre	TOT Post	AC 1	AC 2
TOT Pre Training	0.25***	-0.07	-0.1	0.61***				
TOT Post Training	0.29***	0.02	-0.03	0.56***	0.68***			
Mixed AC 1	0.37***	-0.13	-0.03	0.16	0.40***	0.44***		
Mixed AC 2	0.40***	-0.12	-0.11	0.18*	0.41***	0.51***	0.80***	

Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05.

The study was also interested in the extent to which training and practice could improve TOT scores. Rather than separate the score by type of Sailor (e.g., enlisted vs officer) we partitioned the data into quartiles based upon their initial scores on the TOT. The pre and post TOT data were then analyzed using a Linear Mixed-Effects Regression to determine if there were any differences in score changes based upon the initial score quartile. There was a significant main effect of training ($F = 113.41$, $p < .001$) with scores pre training being significantly lower than post training. There was also a significant interaction of starting quartile and training ($F = 22.29$, $p < .001$). Scores significantly improved for the lower three quartiles, but not for the top quartile. The interaction of starting quartile and training is shown in Figure 3.

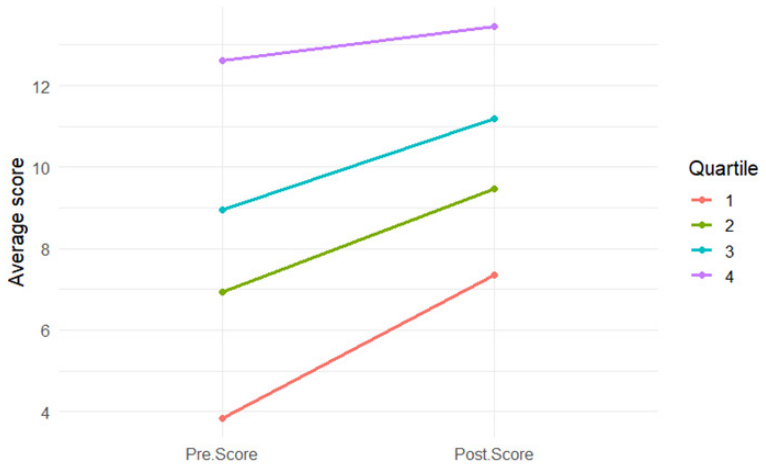


Figure 3: TOT Scores before and after training broken out by starting quartile scores before training.

ASTB Preparation Results

Overall, there was ASTB preparation survey data for 197 of the flight students. Seven percent of the data did not match, that is their reporting of taking the TOT and whether they did take the TOT was inaccurate. Some discrepancy is to be expected given that the gap between the start of

flight training and when they took the ASTB was on average 1 year and 9 months. Of the students who reported taking the TOT, 64% (90 students) said they practiced the task prior to the exam, and over 53% (48 students) reported using the unofficial practice app.

DISCUSSION

The objective of this paper was to evaluate the effectiveness of the TOT in both a high and low stakes setting for predicting Navy flight training outcomes. The study also looked at how repeated practice on the TOT impacts scores on the task as well as its predictive validity. Several studies have shown that TOT scores in a low stakes lab setting are predictive of flight training outcomes (Coyne, 2024, 2025). However, this is the first study which tested a group of flight students who had already taken the TOT within the Navy's ASTB as part of the selection process. This allowed for an examination of TOT in both high and low stakes setting for predicting training outcomes.

Understanding how TOT performs across high and low stakes setting is of critical importance to the Navy. There is mounting evidence to no longer include the Direction Orientation Task (DOT), the ASTB's older spatial ability subtest, in the selection battery. Previous analysis of DOT scores over time found an increase in scores over time and a loss of incremental validity for the exam. Related to this study previous work found that there was no correlation between high stakes DOT, and any low stakes spatial tasks presented in a lab setting including the DOT. These findings have called the DOT subtest's construct validity into question (Coyne et al., 2025). The low stakes spatial tasks including DOT were predictive of flight training outcomes whereas the high stakes DOT was not. Previous results for the DOT are in contrast with those found for the TOT in this study. The high and low stakes versions of the TOT performed similarly in predicting training outcomes. That is, they were all significantly correlated with academic grades, but not correlated with attrition or test failures within this sample. The lack of a correlation with academic attrition is not surprising given the relatively low rate of attrition in the sample. Additionally, all instances of the TOT were significantly correlated with each other. The Mixed attention control tasks were significantly correlated with NIFE Grades, but only one of the two tasks were correlated with the high stakes ASTB TOT. Overall, the data suggests that the TOT is maintaining both predictive validity and construct validity across the different test environments.

While practice on the TOT significantly increased test scores, the correlation between pre and post training remained high ($r = .68$). Further the tests' ability to predict flight training outcomes was nearly identical. A potential limitation of the items used in this analysis is that they may not have been sensitive enough for individuals with higher levels of ability. The top students in the task did not have a significant increase in test performance. The addition of some more difficult items may increase the ability to differentiate individual performance at the upper levels.

The introduction of an unofficial practice TOT so quickly after its introduction to the ASTB means that many of the participants already had substantial practice on the test prior to recording any of the lab TOT scores

used in this analysis. The multiple test forms and large bank of test items for the TOT appear to have allowed the test to maintain its initial utility.

The study also continued to add to the growing body of evidence that show measures of attention control are correlated with flight training outcomes. Further the study's results provides continued support for the use of the TOT as a predictor of training outcomes for U.S. Naval Flight Students. While previous research only had low stakes lab-based TOT data available this is the first study to use data from both the low and high stakes environment. The TOT was able to maintain predictive validity despite repeated practice in the lab as well as any preparation they may have done prior to taking the TOT within the ASTB. These results are important as they suggest TOT can retain its predictive validity in a high stakes testing environment. Overall, the results are promising for the continued use of the TOT within the ASTB, particularly as new images are introduced to maintain the test's item novelty.

ACKNOWLEDGMENT

The research was funded by Natalie Steinhauser's Prepared Warfighter Program at the Office of Naval Research.

REFERENCES

Ackerman, P. L. (2022). Intelligence process vs. content and academic performance: A trip through a house of mirrors. *Journal of Intelligence*, 10(4), 128–147.

Barron, L. G., & Rose, M. R. (2013). Relative validity of distinct spatial abilities: An example with implications for diversity. *International Journal of Selection and Assessment*, 21(4), 400–406.

Burgoyne, A. P., Tsukahara, J. S., Mashburn, C. A., Pak, R., & Engle, R. W. (2023). Nature and Measurement of Attention Control. *Journal of Experimental Psychology: General*. Advance online publication. https://dx.doi.org/10.1037/xge0001408

Coyne, J. T., Brown, N. L., Foroughi, C. K., Sexauer, E., & Rovira, E. (2020) The use of non-spatial strategies in the Direction Orientation Task. *Proceedings of the Human Factors and Ergonomics Society*, 64, 802–806.

Coyne, J. T., Drollinger, S., Brown, N., Foroughi, C., Sibley, C. & Phillips, H. (2022) Limitations of current spatial ability testing for military aviators. *Military Psychology 34*(1), 33–46.

Coyne, J. T., Draheim, C., Sibley, C., Foroughi, C., Strong, K., NeSmith, R., Burgoyne, A. P., & Engle, R. W. (2024, April). *Evaluation of short attention tests for selecting Navy air traffic controllers*. Poster presented at the 39th Society for Industrial and Organizational Psychology annual conference, Chicago, IL, United States.

Damos, D. L. (2011) A Summary of the Technical Pilot Selection Literature. Air Force Personnel Center (AFCAPS-FR-2011-0009). Retrieved November 11, 2022 from https://apps.dtic.mil/sti/pdfs/ADA553707.pdf

Draheim, C., Mashburn, C. A., Martin, J. D., & Engle, R. W. (2019). Reaction time in differential and developmental research: A review and commentary on the problems and alternatives. *Psychological Bulletin, 145*(5), 508–535.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186.

Lohman, D.F. (1993, July). Spatial Ability and G. Paper presented at the first Spearman Seminar, University of Plymouth, UK.