Human Factors in Design, Engineering, and Computing, Vol. 159, 2024, 2006–2016
https://doi.org/10.54941/ahfe1005768

Evaluation of New Measures of Spatial Ability and Attention Control for Selection of Naval Flight Students

Joseph T. Coyne[1], Christopher Draheim[1], Ciara Sibley[1], Nick Armendariz[2], Sarah Melick[2], Cyrus Foroughi[1], Alexander P. Burgoyne[3], and Randall W. Engle[4]

[1]Naval Research Laboratory, Washington, DC 23505, USA
[2]Naval Aerospace Medical Institute, Pensacola, FL 32508, USA
[3]Human Resources Research Organization (HumRRO), Alexandria, VA 22314, USA
[4]Georgia Institute of Technology, Atlanta, GA 30332, USA

ABSTRACT

Identifying individuals who have the knowledge, skills, and ability to become pilots has been a persistent challenge for militaries around the world. Unlike airlines who typically select based upon flight experience, militaries need to identify individuals with the capability to learn to fly. The present study evaluated the use of new measures of spatial ability and attention control to determine if they can improve the prediction of early flight training outcomes in U.S. Navy flight students. We administered these measures to 114 students prior to the start of their Naval flight training and then compared their predictive validity to the composite score and spatial ability test from the Navy's current selection battery used to predict ground school performance. The new spatial ability test was significantly correlated with training grades and the number of setbacks whereas the spatial ability test currently in the selection test battery was not. Two of the three new attention control measures were also significantly correlated with training outcomes, and a composite of the attention measures added incremental validity beyond the current selection test's composite score. Overall, the study found positive results for the new tests and thus we argue they should continue to be evaluated for potential use in Navy personnel selection.

Keywords: Spatial ability, Attention control, Individual differences, Pilot selection

INTRODUCTION

Each year, several thousand applicants take the Navy's Aviation Selection Test Battery (ASTB) which is designed to assess whether an applicant has the cognitive capability to become a naval aviator or flight officer. The battery is comprised of measures of crystalized intelligence (subtests for math, verbal, mechanical, and aviation/nautical knowledge) as well as fluid measures (e.g., psychomotor, attention, spatial and multi-tasking abilities). The ASTB is a high stakes test, and motivated test takers can find an abundance of unofficial test preparation material shared online by previous applicants. It is therefore critical that the ASTB's effectiveness is continually evaluated and that items and subtests are periodically replaced to maintain the battery's ability to predict Navy training outcomes. To this end, the present study investigated the viability of three new double conflict tests of attention control as well as a new measure of spatial ability, the terrain orientation task (TOT), for inclusion into the ASTB. This study is part of a research effort focused on identifying new measures for predicting Navy training outcomes.

Attention Control

Attention control is the broad ability to maintain focus on goal-directed information and behaviors. It is especially important in cognitively demanding situations, such as ones involving cognitive interference or distraction from external events and/or internal thoughts. Some researchers argue that attention control is the main driver of cognitive behavior and even the best marker of an individual's cognitive potential (e.g., Draheim et al., 2022). This idea has been supported by many studies demonstrating the importance of attention-related abilities in driving real-world cognitive behavior as well as the close connection between attention and other cognitive constructs, including fluid intelligence and working memory capacity. But individual differences studies of attention control have often failed to find convincing evidence of this, raising questions as to the importance and coherence of attention control as a unitary ability (for a review, see Draheim et al., 2022). Some researchers contend that this is ultimately a measurement issue as the tasks historically used to assess attention control have poor psychometric properties and are not well-suited to assessing individual differences (e.g., Draheim et al., 2019; Hedge et al., 2018). Recent iterative efforts have endeavored to address this by modifying existing attention control tasks and developing new ones, and some of these studies have been quite successful in demonstrating that attention control is a coherent and reliably measurable ability that correlates strongly with other executive functions (see Burgoyne et al., 2023).

The new attention control measures under evaluation in the present study were three "double conflict" tasks, called so because there can be a conflict (i.e., incongruency) in the stimulus and response portions of the tasks. The tests were recently developed by Randall Engle's lab at Georgia Tech as improved and *very* quick (under 3-minutes each) variants of traditional cognitive conflict tasks. A recent validation study from Dr. Engle's lab demonstrated that these double conflict tests are reliable and valid indicators of attention control that also predict individual differences in multi-tasking ability (Burgoyne et al., 2023). We have also found that performance on them correlates with grades in Navy air traffic control students (Coyne et al., 2024). Given these results, we hypothesized these measures will be predictive of training outcomes for Naval flight students.

Spatial Ability

Spatial ability measures have been used in military selection since World War I (see Damos, 2011 for a history of aviation selection). Spatial ability continues to be used by the Air Force and Navy to select both manned and unmanned pilots and operators today. The US Navy and Air Force currently use the direction orientation test (DOT) to assess spatial ability in their aviation applicants. The Air Force introduced the DOT as part of the Test of Basic Aviation Skills (TBAS) which became operational in 2006. The first large-scale effort to validate the TBAS and DOT assessed performance of just under 1000 students enrolled in the Air Force's primary flight school (Carretta, 2005). DOT accuracy and response time were both correlated with grades during training as well as attrition. Further, DOT performance provided incremental validity in predicting success (attrition and grades) in training over other TBAS subtests. After the Air Force's validation of the DOT, the Navy added it to the ASTB in late 2013.

Initial Navy data found that DOT was predictive of performance during flight training during the early years of the test's inclusion into the ASTB (e.g., Coyne et al., 2022). However, there were some early data showing limitations of the test, specifically, Momen (2009) found significant improvements in DOT performance when respondents took the test a second time. More recently, Coyne et al. (2022) highlighted several additional problems, including (1) a ceiling effect, (2) significant annual increases in applicant scores since the test was introduced, and (3) a loss of incremental validity in predicting success of student Naval Aviators during primary flight training. Attempts to make the DOT more difficult succeeded, specifically by increasing the number of items and widening the distribution of scores (Coyne et al., 2020; Keiser et al., 2019). However, data on these alternatives suggested those who performed well on the test were using mathematical solutions as opposed to spatial ones. This was deemed problematic given both that the ASTB already has a math subtest and that the DOT was included in the ASTB to assess spatial ability. A further concern is the likelihood that motivated applicants would learn such mathematical strategies and then even then share them with future test takers.

The present study examined the Terrain Orientation Task (TOT), which our lab designed to address the known limitations of the DOT. The TOT is similar to terrain association in that it requires identifying terrain features and landmarks in a reference map in order to determine the direction an aircraft is traveling in a rotated map of the same area. Like the DOT, the TOT has face validity since terrain association and navigation are important skills that pilots need to demonstrate during training. Beyond face validity, the TOT has a number of features which should make it superior to the DOT as a selection test. First, the TOT has the potential for an unlimited number of trials since there is no limit to the number of new maps that can be generated, whereas the DOT only has 48 items and applicants are always tested on all 48 of them. Second, the number, size, and contrast of terrain features as well as angles of rotation were manipulated in the TOT which results in varying trial difficulty. Finally, a more practical benefit of the TOT is the lack of a (simple) mathematical solution to each trial. Since each reference map can be unique, participants must identify new landmarks to use as reference points in both images every time. Alternative forms of TOT can also help provide additional test security.

Given the improvements the TOT has over the DOT, we hypothesized that it will predict success in aviation ground school better than the DOT. Ideally, TOT will provide incremental validity in predicting ground school success beyond the current composite score. However, since the DOT is already part of the composite score and the TOT and DOT were designed to assess the same construct, the TOT may not provide incremental validity.

METHOD

A total of 114 Naval flight students participated in the study. The study was approved by the Naval Research Laboratory institutional review board.

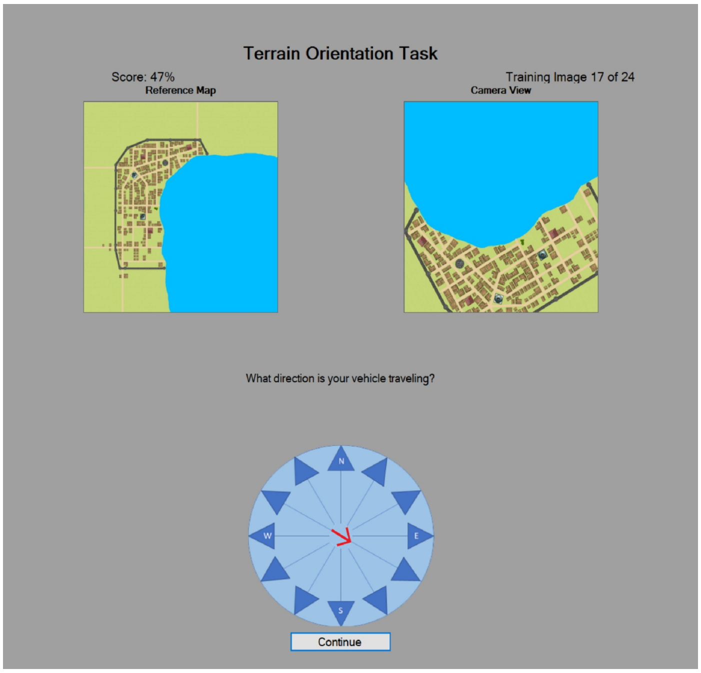
Procedure

Data from this study were part of a larger effort that included a number of cognitive and psychophysiological tests outside the scope of this paper. The study at large also included Sailors and Marines training for other Navy jobs (e.g., air traffic controller). Data from this study were included in Exp. 3 of Robison et al. (2022); however, that effort was not limited by ground school outcome data but rather excluded 20% of the participants for issues related to eye tracking data quality. Since eye tracking data is not of interest in this paper, flight students with poor eye tracking data could be included in the present analyses.

Terrain Orientation Task

The objective of the Terrain Orientation Task (TOT; Figure 1) is to determine the direction an unmanned aerial vehicle (UAV) is traveling by comparing a reference map, which is oriented with North at 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. All maps used in the present study were randomly generated computer game maps of low resolution.

Participants were given 24 practice trials that increased in difficulty and decreased in feedback across each trial. Specifically, trials 1–8 only had four potential response options (i.e., the four cardinal directions), trials 9–17 had eight potential response options (i.e., the cardinal and the intercardinal directions), and trials 18–24 had 12 potential response options (i.e., the four cardinal directions and 30-degree offsets of each). During the first 12 practice trials, participants were given animated feedback when they responded incorrectly. This was done by animating how a UAV would rotate in the map image in order to achieve the camera image. The UAV icon would rotate in 15-degree increments until it reached the correct orientation, while simultaneously the camera image changed to reflect the image of the downward facing camera as the UAV rotated. The rotation was either clockwise or counter clockwise depending on the shortest direction to the correct response. The animation ended when the UAV reached the correct response. The response box also depicted a red arrow indicating the appropriate response. For the final 12 practice trials, the only feedback was the correct response via a red directional arrow. After completing the practice trials, participants then completed the same 24 experimental trials in a fixed order with no performance feedback. All 24 trials had 12 response options. Participants were told to respond as quickly and as accurately as possible, even though there was no time limit per trial or overall time limit for the task.



What direction is your vehicle traveling?

Figure 1: Depiction of feedback during the final 12 practice trials of TOT. The red arrow indicates an incorrect response was originally given and the arrow's direction points to the correct response option.

Attention Control Tasks

In each attention control task, a single stimulus and two response options were presented and both the stimulus and response could be either congruent or incongruent. Trials were randomly generated such that the stimulus was congruent on 50% of trials and the response congruence was independent of the stimulus and was also congruent on 50% of the trials. All tests were speeded response tests with new items appearing immediately after each response. Participants had 30 seconds of scored practice, followed by a chance to reread the instructions before beginning the 90 seconds of scored trials. The tasks were run in E-Prime software and identical to the tasks used in Burgoyne et al. (2023). Participants received feedback on all trials, and both their score and remaining time were continuously displayed on the screen. The dependent variable was the number of correct trials minus the number of incorrect trials. The difference in the three tasks was the source of the incongruence, as explained in detail below. Figure 2 depicts a sample trial from each of the three tasks.

Double Stroop. This task is a double conflict version of the color Stroop. Participants had to identify the font color of the stimulus word (either red or blue). The task was to select the response option with the semantic meaning that matched the font color of the stimulus word. The stimulus was either congruent (with the font color matching the word) or incongruent (with the font color different from the word). Participants then chose between two response options: the word red or blue. The responses were also either congruent or incongruent.

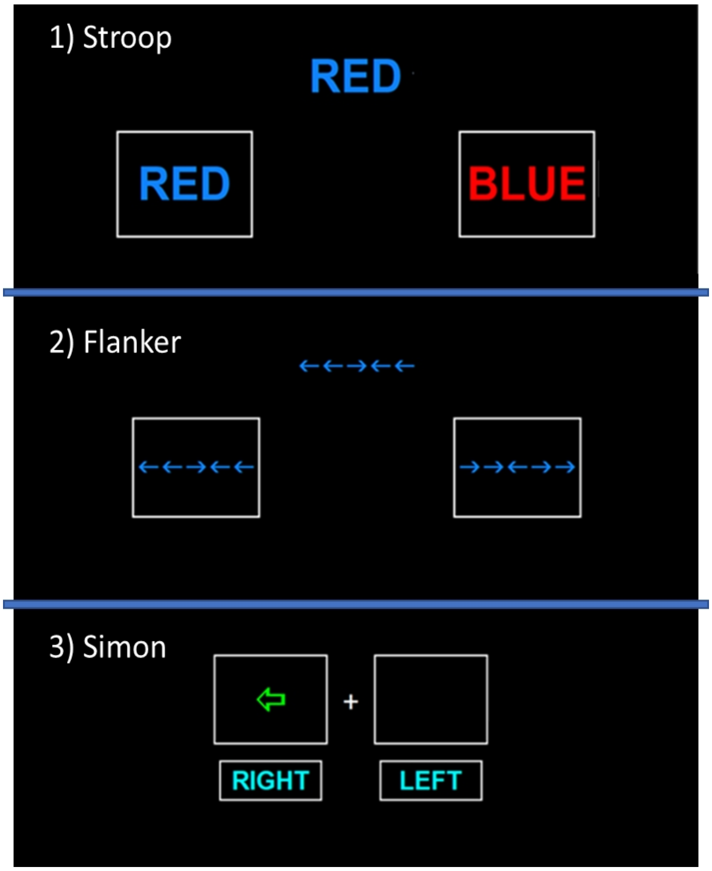


Figure 2: Depiction of the three double conflict attention control tasks.

Double Flanker. As with the double Stroop task, participants were shown a single stimulus and two response options below it. The stimulus and response options each consisted of five arrows which were either congruent (e.g., <<<<) or incongruent (e.g., <<><<). The task was to match the direction of the outside arrows of the top arrow set with the direction of the central arrow on the bottom.

Double Simon. A left or right facing arrow was either presented on either the left or right side of the screen. The stimulus was congruent when the arrow appeared on the side of the screen it pointed towards. The two response options were the words “LEFT” or “RIGHT”. The response was congruent when the word matched the side of the screen of which it was presented. The task was simply to select the response that indicated which direction the arrow was pointing. Practice was 30 seconds with 90 seconds of scored trials, with feedback displayed throughout both.

RESULTS

Lab Performance Dependent Variables

Three dependent variables were computed for the TOT: total correct, median response time (RT), and rate of correct score (RCS). Total score was the total number of correct trials (with a max score of 24), median RT was the median response time across the 24 trials, and RCS was computed by dividing the participant’s total score by their total RT. One participant’s TOT data was removed as their median RT was over 4 SDs above the mean. The dependent variable for all attention control tasks was how many more correct responses they made than incorrect responses. Due to software issues, all three attention control tasks did not always launch. Two participants’ flanker scores were removed for being below 0, but otherwise no outliers were identified in the attention control tasks. We computed a composite score (composite AC) by averaging each participant’s available z scores for the attention control tasks.

Naval Introductory Flight Evaluation (NIFE) and ASTB Performance

Training outcome data as well as data from each student’s final ASTB score were linked with the data from the lab. Training outcomes were attrition, academic grade, and number of failures, in NIFE. NIFE is an 8.4 week course which includes both classroom instruction on basic fundamentals of aviation, and an initial exposure to flight training, NIFE is meant to prepare students for Primary Flight Training (Chief of Naval Air Training [CNATRA], 2020). Attrition was a binary completion variable, academic grade was the average grade across the academic components of NIFE, and number of failures (0-2) was the number of academic components which had to be repeated. The ASTB data used in the analysis was the academic qualifying rating (AQR), a proprietary composite score used to predict NIFE training outcomes, and DOT factor, a proprietary weighting of speed and accuracy from the DOT subtest of the ASTB. A total of 113 participants completed TOT, but outcome data for four students were excluded from analysis as they dropped out from NIFE for reasons unrelated to performance (e.g., medical). Of the remaining 109 students, 12 (11%) attrited for performance reasons. Table 1 shows means, standard deviations, and number of valid data points for all of the experiment data. Uncorrected zero-order Correlations between the ASTB, evaluation tests, NIFE training outcomes are presented in Table 2. Information regarding how DOT is weighted within the AQR composite is excluded for test security purposes.

Table 1. Means and standard deviations for TOT and three attention control tasks.

	Mean	SD	N
TOT N Correct	15.34	4.80	113
TOT Mdn RT	12.19	4.98	113
TOT RCS	0.05	0.02	113
Simon	69.04	6.65	112
Flanker	41.88	14.07	105
Stroop	36.43	13.69	110
Composite AC	−0.03	0.78	114

Table 2. Correlation table of TOT, attention control, NIFE training outcomes and ASTB. Correlations with an asterisk (*) are significant at *p* <.05.

	Grade	Attrite	Setback	TOTN Ct	TOT RT	TOT RCS	Simon	Flanker	Stroop	Composite
Grade										
Attrite	−0.45*									
Setbacks	−0.66*	0.58*								
TOT N Correct	0.25*	−0.04	−0.13							
TOT RT	−0.17	0.13	0.16	0.20 *						
TOT RCS	0.30*	−0.11	−0.19*	0.59*	−0.54*					
Simon	0.17	−0.07	−0.11	0.23∗	−0.07	0.23∗				
Flanker	0.38*	−0.07	−0.16	0.17	−0.05	0.23*	0.21			
Stroop	0.28*	−0.04	−0.17	0.23*	0.02	0.23*	0.34*	0.28*		
AC Composite	0.39*	−0.07	−0.20*	0.30*	−0.03	0.31*	0.70*	0.74*	0.78*	
AQR	0.50*	−0.04	−0.32*	0.34*	−0.08	0.36*	0.16	0.32*	0.30*	0.38*
DOT Factor	−0.06	−0.01	0.03	−0.09	−0.05	0.05	0.08	0.04	0.05	0.05

We conducted follow-up hierarchical regression analyses to determine if the TOT RCS and composite AC scores added incremental validity to the prediction of NIFE performance beyond the ASTB AQR composite score. NIFE attrition rate was excluded from analysis as no variables in this sample predicted attrition. In the base model, AQR scores were significantly correlated with NIFE grades and accounted for 24% adjusted variance ($\beta = 0.50, p < .01$). Adding TOT RCS and composite AC added an additional 4% adjusted variance to the model. However, only AQR ($\beta = 0.39, p < .01$) and composite AC ($\beta = 0.25$, $p < .05$) predicted significant variance in the model. For NIFE setbacks, the base model with AQR was significant (*adj.* $R^2 = .09$, $p < .01$) and neither TOT ($\beta = -0.08$, $p = 0.42$) nor composite AC ($\beta = -0.08, p = 0.56$) significantly increased the variance accounted for above AQR.

DISCUSSION

The goal of this study was to evaluate the TOT and three new attention control tasks for predicting Navy flight training outcomes. Our evaluation provides preliminary evidence that TOT predicts important Navy training outcomes, specifically grades and number of failures in the academic component of NIFE. Additionally, the double Stroop and flanker tasks were also predictive of NIFE grades. Neither the ASTB nor any of the lab tests evaluated in this study were significantly correlated with attrition. This is a common finding given that attrition rate is low and a binary variable.

Regression analyses clarified that although the TOT, and a composite attention control measure all correlated with NIFE grades, only the composite attention control measure added incremental prediction to grades above and beyond AQR. Despite this, TOT had significant correlations with NIFE outcomes and should continue to be assessed for potential use.

The success of the attention control tasks in this study is important. The tests not only added incremental validity, but all three take under nine minutes to complete. The flanker, and other double conflict attention tests, have also been shown to predict training outcomes in the Navy's Air Traffic Control school, another cognitively demanding military training program and with a high rate of attrition (Coyne, et al., 2024). As such, the flanker and other attention tasks show promise and should continue to be evaluated for their potential in military selection.

Our results provide even more evidence that the ASTB's current spatial ability test, the DOT, has lost its effectiveness (i.e., predictive validity). The DOT is part of the ASTB composite score (AQR) used to predict ground school performance and yet it was not significantly correlated with any NIFE outcomes in our sample. Further, DOT performance did not significantly correlate with any of the lab tests, including the TOT which is a similar task designed to measure the same ability. This lack of a relationship may be because students taking the ASTB can find resources on the DOT and prepare for it by practicing non-spatial strategies, and thus for motivated examinees it no longer assesses spatial ability. While the DOT factor alone is not used to directly predict NIFE, it is included in the AQR composite and our results suggest this should be reassessed. Despite the limitations of the DOT, AQR continues to significantly predict NIFE outcomes.

One significant limitation of the present work is that TOT performance was not compared with either other measures of spatial ability or DOT performance from an experimental setting (i.e., with naive participants). While the TOT has face validity, as it is similar to terrain association which is important for flying/navigating an aircraft, future work should compare TOT with other spatial ability measures to assess its construct validity. Another limitation of the current work is the TOT used here did not have a time limit. As such, it is likely that participants differentially emphasized the importance of speed vs. accuracy. Such differences in speed vs. accuracy emphasis would impact all performance variables of the task, especially total score and median RT because these only index speed or accuracy in isolation (see Goldhammer, 2015). While the DOT factor score is a composite of both speed and accuracy, the DOT within the ASTB also has no time limits. For future versions of TOT, it might be worthwhile to impose either an overall or per-item time limit to better control for individual differences in speed-accuracy trade-offs.

Any new test added to the ASTB will face the same challenges as the DOT. The DOT was initially predictive of aviation success because examinees were naive to the test. However, as more applicants took the ASTB and TBAS, some shared strategies, developed practice cue cards, and even created unofficial practice versions. This may explain why data suggest that the approach and strategies used to complete the DOT have shifted over time. Unequivocally, there has been a significant increase in the DOT factor scores and a loss of incremental validity. Thus, any new cognitive ability test (e.g., spatial ability and attention) being administered in a lab setting has the advantage of naive participants, and it will be more likely to capture the construct it is intended to measure. The TOT has the potential to be more robust than the DOT given its ability to constantly add new items. Obviously, its ability to be more resilient to practice effects, and to be able to withstand time in general, has yet to be empirically validated.

In conclusion, this study provides preliminary evidence that the TOT and double flanker tasks can predict training outcomes for US Naval flight students. However, the ability of these tests to reliably measure individual differences in cognitive ability and to withstand repeated practice still needs to be established.

ACKNOWLEDGMENT

This research was funded by Dr. Jeffery Morrison's Command Decision Making Program and LCDR Michael Natali's Manpower, Personnel, Training, and Education Program at the Office of Naval Research.

REFERENCES

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

Carretta, T. R. (2005). Development and Validation of the Test of Basic Aviation Skills.

(TBAS). Air Force Research Laboratory (AFRL-HE-WP-TR-2005-0172) retrieved
December 1, 2022 from https://apps.dtic.mil/sti/pdfs/ADA442563.pdf

Chief of Naval Air Training (2020). Naval Introductory Flight Evaluation. CNATRAINST 1542.178A. Corpus Christi, TX: Department of the Navy. retrieved December 23, 2022 from https://www.cnatra.navy.mil/local/docs/mcg/1542.178.pdf

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.

Draheim, C., Pak, R., Draheim, A. A., & Engle, R. W. (2022). The role of attention control in complex real-world tasks. *Psychonomic Bulletin & Review*, *29*(4), 1143–1197.

Goldhammer, F. (2015) Measuring ability, speed, or both? Challenges, psychometric solutions, and what can be gained from experimental control. *Measurement 13*(3-4) 133–164.

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.

Keiser, H. N., Moclaire, C. M., King, K. M., Brown, N. L., Foroughi, C. K., Sibley, C., & Coyne, J. T. (2019). Updating the direction orientation task: an aviation selection tool. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 63, No. 1, pp. 1414–1418). Sage CA: Los Angeles, CA: SAGE Publications.

Momen, N. (2009). The effects of alternative input devices and repeated exposures on the Test of Basic Aviation Skills (TBAS) performance. *Military Medicine*, 174, 1282–1286.

Robison, M. K., Coyne, J. T., Sibley, C., Brown, N. L., Neilson, B., & Foroughi, C. (2022). An examination of relations between baseline pupil measures and cognitive abilities. Psychophysiology, *59*(12), e14124.