# Assessing Intuitive Decision Making with Cognitive Neuroscience-based Methods

**Lisa C. Lucia, Ph.D., Jeffrey M. Beaubien, Ph.D., E. Webb Stacy, Ph.D.**
**Aptima, Inc. Woburn, MA**
**llucia, jbeaubien, wstacy@aptima.com**

**CAPT Ronald Steed (USN, ret.)**
**Mystic, CT**
**ronaldsteed@gmail.com**

## ABSTRACT

Many military jobs pose complex perceptual-cognitive challenges, such as detecting potential collisions among vehicles that are operating in close physical proximity. These situations require rapid and intuitive decision making. Unlike deliberate decision making, intuitive "hunches" are largely automatic and do not require working memory (Evans, 2003; Evans & Stanovich, 2013; Kahneman, 2011). As a result, cognitive neuroscience-based methods such as Diffusion Modeling (DM; Wagenmakers, 2009) and electroencephalography (EEG) are well-suited to the study of intuitive decision making. Previous research by Lucia and colleagues (2017) provided support for a generalized neurological "signature" of intuitive decision making (Luu et al, 2010). This signature held for two different decision tasks, both of which used briefly-presented static imagery as the decision stimulus. In the current study, we examine the generalizability of this neurological signature with a new sample of 22 submariners who performed a similar decision task, but which used brief motion pictures as the decision stimulus. We also analyze the response time (RT) and accuracy data from both studies using DM. Unlike traditional methods for analyzing RT and decision accuracy, which treat them as separate dependent variables (Luce, 1986), DM combines them to model the underlying cognitive processes: processing speed (Delta), response caution (Alpha), stimulus encoding (Tau), and response bias (Beta). While some of the findings generalized across the two studies, others did not. Differences may be due to the type of decision stimuli used: static imagery vs. motion pictures. In some cases, however, the EEG results were able to reliably differentiate experts from novices even when the behaviorally-based measures did not. The paper concludes with practitioner-oriented guidelines for using EEG and DM methods to study intuitive decision making.

## ABOUT THE AUTHORS

**Lisa C. Lucia, Ph.D.** is a Scientist at Aptima, Inc., where she leads projects that involve the development of neuroscience-based training tools, decision support tools, and medical informatics systems. Her research interests span from the brain-bases of human perception and memory to visuospatial abilities and skill learning. She holds a Ph.D. in Cognitive Neuroscience from Tufts University, and a B.S. in Biological Psychology from Bates College.

**Jeffrey M. Beaubien, Ph.D.** is a Distinguished Principal Scientist at Aptima, Inc. For the past 20 years, his work has focused on training and assessing leadership, teamwork, and decision-making skills. His research has been sponsored by the U.S. Navy, the U.S. Army, the U.S. Air Force, and the Telemedicine and Advanced Research Technologies Center, among others. Dr. Beaubien holds a Ph.D. in Industrial and Organizational Psychology from George Mason University, a M.A. in Industrial and Organizational Psychology from the University of New Haven, and a B.A. in Psychology from the University of Rhode Island.

**Webb Stacy, Ph.D.** is a Corporate Fellow at Aptima, Inc., where he is responsible for enhancing Aptima's technology portfolio. Dr. Stacy has an interest in using modern Cognitive Science to improve experiential training. His recent work includes investigating the relationship of simulator fidelity to training effectiveness and developing an approach to optimizing the training value of experiential scenarios. Dr. Stacy holds a Ph.D. in Cognitive Science from SUNY/Buffalo, and a B.A. in Psychology from the University of Michigan.

**CAPT Ronald Steed (USN, ret.)** was previously the Principal at UpScopeConsulting Group, LLC where he served on research and development projects involving command decision making, submarine systems, team dynamics, and human factors. Mr. Steed transitioned from the U.S. Navy's Submarine Force in 2007. His last assignment was as the Commander of Submarine Squadron Two in Groton, CT. Captain Steed holds a M.S. in Engineering Management from Old Dominion University, and a B.S. in Electrical Engineering from The Citadel.

# Assessing Intuitive Decision Making with Cognitive Neuroscience-based Methods

**Lisa C. Lucia, Ph.D., Jeffrey M. Beaubien, Ph.D., E. Webb Stacy, Ph.D.**
**Aptima, Inc. Woburn, MA**
**llucia, jbeaubien, wstacy@aptima.com**

**CAPT Ronald Steed (USN, ret.)**
**Mystic, CT**
**ronaldsteed@gmail.com**

## BACKGROUND

Military missions often involve uncertainty, which poses significant perceptual and cognitive challenges. For example, warfighters must continually update their situation awareness, rapidly detect the onset of critical cues (e.g., the presence of nearby surface vessels when a submarine is surfacing to periscope depth), and respond accurately without hesitation. Such situations do not allow time for extensive deliberation. In these situations, rapid decision making is essential to survival. The need to make accurate intuitive decisions is not unique to the military. It is critical to successful performance in law enforcement, disaster response, medicine, transportation, and aviation.

The dual-process theory of decision making (Evans, 2003; Evans & Stanovich, 2013; Kahneman, 2011) postulates that there are two brain processes. One of these processes, “Type 1,” operates at the unconscious level. This type of decision making is extremely fast, makes minimal demands on working memory, and operates (in part) by pattern matching the current situation to one’s corpus of accumulated prior experiences in long-term memory. All humans engage in a considerable amount of Type 1 processing in their day-to-day lives, for example when driving, reading, and identifying everyday objects. In addition, experts have well-developed domain-specific Type 1 skills. For example, a chess master can quickly look at a populated chessboard, determine each player’s strategy, and project the next two or three best moves – all within a matter of seconds. By comparison, “Type 2” processing operates at the conscious level. It is much slower, requires active attention, places heavy demands on working memory, and involves effortful deliberation or calculation. It is akin to the slow and deliberative decision making approach used by novices (Kahneman & Klein, 2009; Klein, 2008). For example, a novice chess player must carefully study the two players’ positions, reflect on their respective strategies, and consciously recall the relevant decision rules before being able to predict the next move. In practice, Type 1 and Type 2 decision making often work together. For example, the pattern matching component of Recognition Primed Decision making (RPD) model corresponds to intuitive (e.g., Type 1) decision processes, while the mental simulation component corresponds to deliberative (e.g., Type 2) decision processes (Klein, 2008).

Some researchers have suggested that Type 1 processes’ speed advantage is rooted in the ability to subconsciously extract statistical covariation from the environment (Evans, 2010; Kahneman & Klein, 2009; Reber, 2013). Specifically, experts are believed to be adept at leveraging information from long-term memory to selectively focus their attention onto those environmental cues that are most informative or diagnostic, and then use this information during decision making. Support for this idea is provided by electroencephalography (EEG) research. For example, Luu and colleagues (Luu et al., 2010) identified an EEG indicator (an inversion of the P3 component, occurring ~300-400 ms after stimulus onset, e.g., N300) that distinguished performance between object stimuli and non-object stimuli on a partially-obscured everyday object recognition task. These results, conducted with college-aged participants, show the expert object identification skill inherent in healthy, sighted adults.

A more recent study investigated the generalizability of this neurological signature with a sample of submariners (Lucia et al., 2017). In the first experiment, expert and novice submariners performed the same everyday object recognition task that was used by Luu and colleagues (2010). The data showed between-group differences in performance (experts were more accurate than novices), along with an earlier EEG signature, the P200 (175-275 ms). The P200 is posited to reflect the selective attention processes that are enhanced in experts vis-à-vis novices (Phillips & Takeda, 2009). In the second experiment, expert and novice submariners performed a submarine-specific decision task which required them to make rapid course safety decisions based on briefly-presented periscope images. Again, experts performed more accurately than novices, and a similar P200 group effect was observed. Furthermore, similar to the findings from Luu et al., a second EEG component also revealed significant between-group differences: the N300. This signal occurred between 275-400 ms post stimulus onset, and is thought to reflect knowledge-driven pattern matching (KDPM) at the start of decision processes (Schendan & Ganis, 2015). In sum, the earlier EEG component (P200) may reflect a generalizable skill that facilitated the expert submariners' performance across the two different decision making tasks – one involving everyday object recognition decisions and the other involving domain-specific safety decisions. Additionally, because the later EEG component (N300) was only significant for the submarine-relevant task, it may suggest that cognitive processing during this time may be impacted by the operators' submarine-relevant experience.

A key limitation of these prior studies is their reliance on briefly-presented static images as the decision stimulus. While rapid, intuitive decisions are sometimes made based on static images – such as when a radiologist interprets a patient's x-ray – radiologists have an extended period of time with which to make their decision. As a result, the brief presentation of static images is largely an experimental manipulation. In real life, many intuitive decisions are made after watching a situation unfold over time. For example, submarine sonar operators track a target's location over time in order to predict whether or not it represents a potential collision threat. Similarly, athletes track the position of moving objects, such as balls or pucks, to estimate where they will be at some point in the immediate future. However, motion pictures are not typically used in event-related potential (ERP) research, because video is generally not conducive to time-locking (Maguire et al., 2013). EEG-based measures require millisecond-level accuracy to pinpoint the onset of a critical visual stimulus for inferring perceptual or cognitive processes. Furthermore, in video, most objects maintain their form (i.e., object constancy), but actions which unfold over time have the potential to obscure key features, such as when wave swells partially occlude the periscope's view of a surface vessel.

To our knowledge, only one ERP study with video-based decision stimuli has been conducted to date. In a study of object congruency, Sitnikova and colleagues (Sitnikova, Kuperberg, & Holcomb, 2003) showed participants lead-up videos that expressed the situational context (e.g., a man applying shaving cream to his face), and time-locked the ERP trials to the start of decision videos which showed an action (e.g., shaving) with either a congruent object (e.g., a razor) or an incongruent one (e.g., a rolling pin). Participants were required to rapidly decide whether or each decision video showed a scenario that one would witness in everyday life. Results revealed the expected N400 effect, but it lasted 2-3 times longer than effects typically found in decision studies that use briefly-presented static images (~300-800 ms vs. ~325-600 ms). This prolonged time-course might be explained by the variability in the timing of identification of different objects in the videos, resulting in a smeared N400 effect across trials. Therefore, the first objective of the current study was to examine the extent to which the previous ERP findings (Lucia et al., 2017) generalize to more realistic decision tasks that use video stimuli instead of static images. We hypothesized that this new paradigm would yield similar ERP results (e.g., P200, N300), but that they would have longer durations, similar to those observed by Sitnikova and colleagues (2003).

The second objective of the current study was to use advanced modeling techniques to investigate the cognitive processes involved in intuitive decision making. Specifically, we utilize an approach called Diffusion Modeling (DM) which takes all performance data – including correct and incorrect response times (RTs), along with measures of decision accuracy – into account (Ratcliff, 1985; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002; Wagenmakers, 2009). The basic idea behind the Diffusion Model is that the time required to make a decision in a forced two-choice RT experiment involves quickly gathering information about which response is best until an information threshold is reached, at which point a decision is made (see Figure 1).

There are four parameters of interest in the DM model. Delta measures the speed of information processing, regardless of whether or not the decision was correct. Higher values represent faster decision making. Alpha measures the amount of information needed to make a decision. Higher values mean that more information must be gathered before making a decision. Therefore, higher Alpha values reflect a lower tolerance for uncertainty (i.e., response caution). Tau represents the amount of time spent on non-decision tasks such as stimulus encoding and response execution. Larger values mean that more time is spent in these activities. Unlike Delta and Alpha, whose metrics are arbitrary, Tau is measured in milliseconds. Finally, Beta represents the extent to which the participant exhibits a systematic response bias. For example, in a two-choice decision task, some participants may be more likely to respond "no" vs. "yes." Beta represents this bias. Taken together, these four variables convey the critical cognitive processes that occur during intuitive decision making. DM modeling was applied to the two new video-based experiments presented in this paper. For comparison purposes, it was also applied to the submarine-specific decision task from the prior study (Lucia et al, 2017).

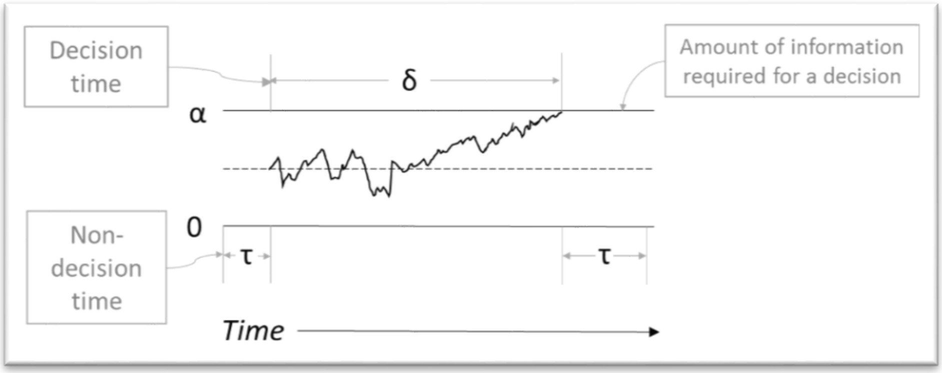

**Figure 1. Selected elements of the Diffusion Model (Wagenmakers, 2009). Delta (δ) represents the participant's processing speed. Alpha (α) represents the amount of information that the participant requires to make a decision. Tau (τ) represents the amount of time that the participant spends in activities other than decision making, such as stimulus encoding or response execution. When response execution is trivial (such as responding via a key press), differences in τ primarily reflect stimulus encoding time. A fourth variable, Beta (β; not shown) represents response bias. Taken together, these four variables convey the critical cognitive processes that occur during intuitive decision making.**

## METHOD

### Participants

Thirty-five active duty and recently-retired submariners from the metropolitan New London, CT area consented to participate in the study. The sample included a mix of enlisted personnel, officers-in-training, and officers, all of whom operated the periscope as a part of their job duties. The active duty participants were students and instructors from the Naval Submarine School (NSS) Submarine Officer Advanced Course (SOAC) who were recruited from the Submarine Learning Center (SLC). The retired participants were recruited from a local consulting firm. All of the participants were adult male English-speakers with normal or corrected-to-normal vision. None had a history of head injuries or were taking medications that would otherwise affect their EEG recordings. Of the 35 participants, three from the Video Preview Block (s4, s6, and s22) and four from the Video Only Block (s4, s19, s22, s27) were excluded from data analyses because they had too few artifact-free EEG trials. This practice is consistent with prior ERP research.

### Procedure

In many respects, the current study's design and method mirrored the previous study (Lucia et al., 2017), except for the use of video-based decision stimuli rather than briefly-presented static images. As in the previous study, upon arrival at the study location, the investigator explained the experimental tasks to the participants and answered all of their questions. Written informed consent was then obtained using a protocol that was approved by the Office of Naval Research (ONR). The data collection session began by having the participants complete a brief demographic questionnaire. Next, the participants completed the Periscope Operator Adaptive Training tool (POAT; Landsberg et al., 2012), which was used as a pre-test. Participants were then fitted with a Quick-20 Dry EEG Headset (Cognionics, Inc.; Figure 2C) and were seated 65 cm in front of a 22" LCD computer monitor. After a set of practice trials, the participants then completed two experiment blocks (Video Preview, Video Only) in order, with a brief rest in between. None of the decision stimuli used during the practice trials were repeated during the experiment blocks. Finally, the EEG headset was removed, and the participants completed a post-experiment survey. The entire procedure lasted approximately 90 minutes.

Expertise is defined as the ability to demonstrate consistent levels of superior performance on a set of domain-specific tasks (Ericsson, Krampe, & Tesch-Romer, 1993). However, certain tasks such as periscope operations decay quickly without practice. Therefore, we could not use the participants' prior Navy qualifications or service ratings to differentiate "novices" from "experts." Similarly, we could not rely on proxy variables such as the participants' rank or total years of Navy service. Therefore, consistent with prior work (Lucia et al., 2017), POAT test scores were used to assign the participants into expertise conditions ("novice" vs. "expert") for subsequent statistical analysis. POAT test scores were calculated as an average of per trial performance scores; per trial, scores were computed by multiplying points earned for decision accuracy by a weighting factor based on response speed. Higher scores represent better performance (e.g., more accurate angle on the bow decisions, speedier responses). The range of possible POAT test scores spans from 0 points (worst performance) to 100 points (best performance). In the current study, the participants' POAT scores ranged from 0.59 to 0.85 (*M* = 0.69, *SD* = 0.07), which is also consistent with prior research (Lucia et al., 2017). However, when attempting to use a median split to divide the participants into expertise groups for statistical comparison (cf. Lucia et al., 2017), there were no significant between-group differences. Therefore, we created an "extreme groups" comparison design by removing the ten middle-level performers – those with POAT scores ranging from 0.67 to 0.71 – thereby allowing us to compare the highest-performing (e.g., "experts") with the poorest-performing (e.g., "novices") participants (see Table 1).

**Table 1. Group Demographics per Block**

| **Block** | **Group** | **POAT Test Score**<br>M (SD) | **Age**<br>(years)<br>M (SD) | **Navy experience**<br>(years)<br>M (SD) | **Retirees per group**<br>(#) | **Retirees' years since active duty**<br>M (SD) |
|---|---|---|---|---|---|---|
| **Video Preview** | Novice (n=10) | 62.3 (1.9) | 39.6 (9.7) | 13.8 (7.5) | 4 | 17.3 (7.2) |
| | Expert (n=12) | 75.7 (4.5) | 35.6 (7.9) | 13.4 (6.8) | 5 | 8.2 (3.1) |
| **Video Only** | Novice (n=9) | 62.1 (1.9) | 40.8 (9.5) | 14.5 (7.6) | 4 | 17.3 (7.2) |
| | Expert (n=12) | 75.9 (4.4) | 35.6 (7.9) | 13.1 (6.9) | 6 | 8.3 (2.8) |

Not surprisingly, this approach resulted in the expert group having significantly higher POAT scores than the novice group (*p*s < .001). There were no significant differences between the two groups as a function of age (*p*s > .20), or years of Navy service (*p*s > .67). For the non-active duty participants, novices had been retired longer than the experts (*p*s < .09).

### Decision Tasks

EEG studies generally do not utilize video clips because it can be difficult to time-synchronize the point during which the critical visual input is present. Because EEG technology allows the observation of millisecond-by-millisecond brain activity, time-synchronization is critical. For this reason, we used two slightly different decision tasks, the second of which was "riskier" than the first from the perspective of obtaining accurate EEG measurements. For both tasks, participants were informed about the submarine's current course, speed, and depth; they then answered via button press whether or not it was safe to continue. The researcher emphasized that participants should rely on their impressions, and that they did not to indicate the specific action they would need to take if it was unsafe to continue. Submarine-specific images and videos were taken from the POAT training tool (Landsberg et al., 2012). Each stimulus depicted a ship in the open water as seen through a submarine's periscope, with views that varied as a function of vessel type, distance from the submarine, angle on the bow, sea state, and time of day (see Figure 2).

Video Preview Task. In the Video Preview (VP) task block, participants watched a brief video clip, which was then followed by a static decision image (see Figure 2A). In this sense, it was similar to the approach used by Sitnikova and colleagues (2003), however the video preview showed the exact environmental conditions for the upcoming decision trial. It was included because a foggy nighttime scene might prompt the search for different visual cues than a clear daytime scene. Each trial began with a cue (a fixation cross lasting 500 ms) alerting participants to get ready for the video clip. Following the fixation, the preview clip was shown for 2 seconds. Participants were instructed to use the preview as a guide for what to expect in terms of the environmental conditions. After a variable-length blank screen (lasting from 1.5 and 2.5 seconds, generated randomly), a static image of a vessel in the open water was presented for 400 ms and then immediately followed by a blank screen. This blank screen remained until the participant responded. As soon as possible after the onset of the picture, participants were encouraged to indicate their response by pressing a button for Yes (i.e., safe to proceed on current course, speed, and depth) or No. The response interval was from the onset of static decision image until the button press response. A total of 69 trials assessed decision making performance in this block (safe: 48; not safe: 21). On average, the VP task block took about 14 minutes to complete.

Video Only Task. In the Video Only (VO) task block, participants observed a video clip lasting approximately 2 seconds, and were asked to respond as quickly as possible (see Figure 2B). The video stimuli depicted a single vessel in a realistic and active view through the periscope. For example, the vessel appeared to move, and atmospheric effects such as fog sometimes partially obscured it. Each trial began with a variable-length blank screen (lasting from 1.5 and 2.5 seconds, generated randomly). Next, a cue (a fixation cross lasting 500 ms) alerted participants to the upcoming visual stimulus. Following the fixation, a periscope video clip was shown for a maximum of 2 seconds, followed by a blank screen. Participants were instructed to indicate their decision by pressing a button for Yes (i.e., safe to proceed on current course, depth, and speed) or No as soon as possible after the video started to play. The response interval was from the onset of each video clip until the button press response. A total of 69 trials assessed decision making performance in this block (safe: 35; not safe: 34). On average, the VO task block took about 10 minutes to complete.

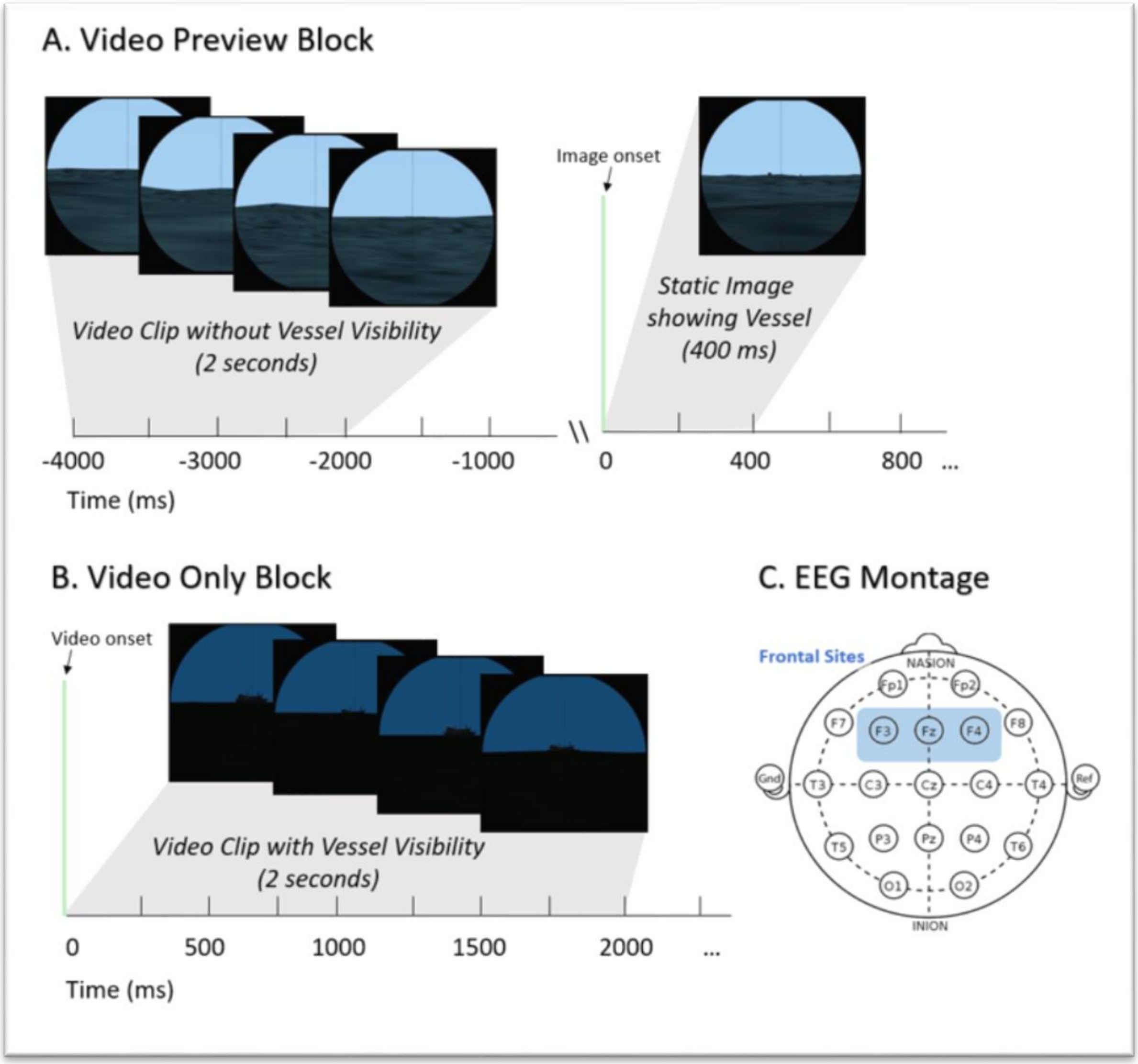

**Figure 2. Decision stimuli & EEG montage. (A-B) Experiment timing and sample decision stimuli are presented for the Video Preview (VP) and Video Only (VO) decision tasks, respectively. (C) Also depicted is the layout of electrode sites across the scalp (top-down view, nose on top) for the Quick-20 Dry EEG Headset. ERP analyses focused on frontal sites labeled as F3, Fz, and F4 (as in Lucia et al., 2017).**

### ERP Analyses

We investigated brain activity patterns that have been associated with selective attention and knowledge-based pattern matching. Consistent with previous work (Luu et al., 2010; Lucia et al., 2017), ERP mean amplitudes and peak positive and negative latencies were calculated at frontal sites F3, Fz, and F4 (see Figure 2C) to assess anterior effects for the P200 and the N300 components. Visual inspection of the ERPs across the two tasks showed that P200 and N300 peaks were delayed in the VO task by ~300 ms. For the VP task, P200 was assessed between 200-290 ms and N300 between 250-350 ms. In the VO task, however, 'P200' was assessed between 500-575 ms and 'N300' was assessed between 550-650 ms.

### Diffusion Modeling Analysis

To investigate the cognitive processes involved in decision making, performance data (RTs and accuracy rates) were subjected to Diffusion Modeling (DM; Wagenmakers, 2009) using the RWeiner package (Wabersich & Vandekerckhove, 2014). Diffusion analysis parameters (Delta, Alpha, Tau, and Beta) were computed separately for the VO and VP tasks. For comparison purposes, data from the submarine-specific task in the previous study (Lucia et al., 2017) were also analyzed using DM.

### Statistical Analyses

The EEG and RT analyses focused on correct trials only, as per standard practice. The DM parameters (Delta, Alpha, Tau, and Beta) were generated using both accuracy and RT data for all responses. Statistical tests of these calculated variables included the between-subjects factor of expertise group. Additionally, the EEG analyses also included the within-subject factor of channel location (F3, Fz, and F4). Before presenting the EEG and DM analyses, however, we present the raw performance data (RT and accuracy) separately, as per Lucia et al., 2017. The purpose behind these visualizations is to demonstrate that the RT and accuracy data from the current study are consistent with the submarine specific-task from the previous study.

## RESULTS

### Accuracy and Response Time

In the VP task, participants tended to make their responses within approximately 1.3 seconds (novices: $M$ = 1363.67 ms, $SD$ = 226.65; experts: $M$ = 1297.34 ms, $SD$ = 190.75). In the VO task, participants tended to make their responses within approximately 1.2 seconds (novices: $M$ = 1270.54 ms, $SD$ = 227.99; experts: $M$ = 1141.36 ms, $SD$ = 222.19). There were no statistically significant differences in mean RTs between the two expertise groups in either the VP task, $t$ (20) = .73, $p$ = .47, or in the VO task, $t$ (19) = 1.30, $p$ = .21 (see Figure 3). These RT results are comparable to the Periscope task (from Lucia et al., 2017) which used only static images as the decision stimuli. The results suggest that the participants were following the instructions to respond with a quick, intuition-based response rather than with an extended, deliberation-based response.

There were also no statistically significant between group differences in mean accuracy scores for the VP task, $t$ (20) = .94, $p$ = .36, or for the VO task, $t$ (19) = -.43, $p$ = .67 (see Figure 4). However, the accuracy results did differ from those of the 2017 submarine-specific task, which showed that experts were significantly more accurate than novices. In all three cases, the accuracy rate across tasks had a very constrained range (58%-61% mean accuracy), which suggests that the experimental tasks were of similar difficulty. Because random guessing on a two-choice decision task would result in 50% accuracy, the difficulty level across all three tasks could be considered "medium" to "high."

### Event-Related Potentials (ERPs)

Video Preview Task. In the VP task, scalp ERPs at focused frontal sites (see Figure 5 and Table 2) during the P200 appeared to be more positive and later peaking in experts than in novices. Mixed factor ANOVAs examining P200 mean amplitude and peak latency (between 200-290 ms) showed no main effects of expertise group with regard to amplitude, $F$ (1,60) = .09, $p$ = .77, but trended toward a significant difference in terms of latency, $F$ (1,60) = 3.43, $p$ = .07, although it unexpectedly peaked later in experts than novices. This pattern of P200 effects differs prior work with static images (from Lucia et al., 2017). On that task, P200 amplitude was greater in experts compared to novices, but its latency did not differ between the expertise groups. The N300 (between 250-350 ms) appeared to be more negative and peak earlier in novices than in experts. For this component, mixed factor ANOVAs showed no main effects of expertise group for amplitude or latency, *F* (1, 60) = 2.12, *p* = .15 and *F* (1, 60) = 1.01, *p* = .32, respectively. Again, this pattern of effects differs from the prior work with static images (from Lucia et al., 2017). On that task, the N300 peaked earlier in experts but the amplitude did not differ between groups.

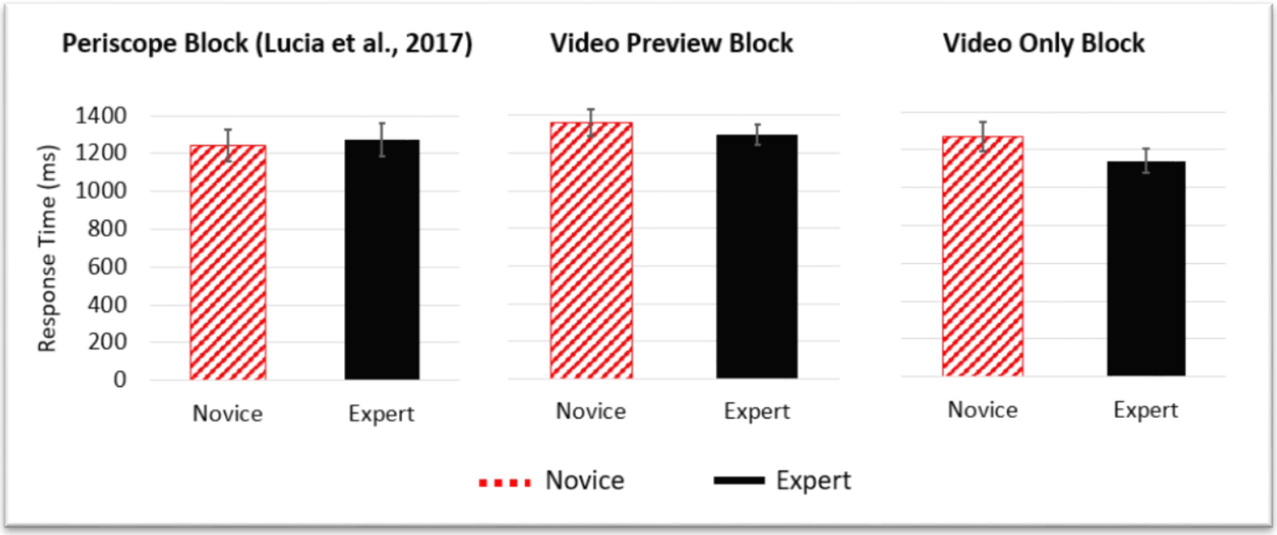

**Figure 3. Mean Response Times (RTs) per Decision Task. This figure shows mean RT for the static periscope image task (left; see Lucia et al., 2017), the Video Preview task (middle), and the Video Only task (right). There were no statistically significant differences between expertise groups when using brief motion pictures as the decision stimuli. This finding is similar to the previous study, which used static images as the decision stimuli.**

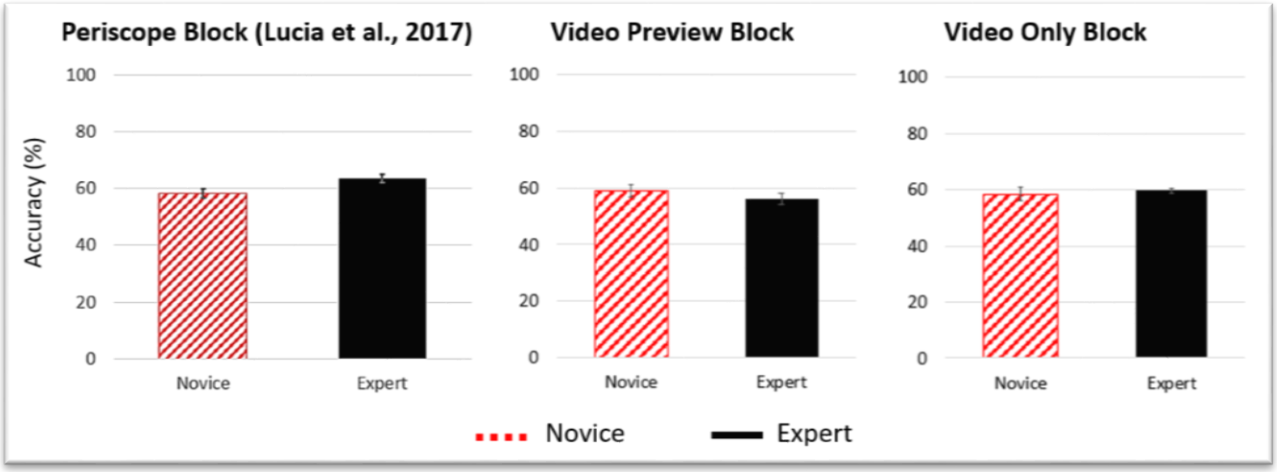

**Figure 4. Mean Accuracy Scores per Decision Task. This figure shows mean accuracy scores for the static periscope image task (left; see Lucia et al., 2017), the Video Preview task (middle), and the Video Only task (right). There were no statistically significant differences between expertise groups when using brief motion pictures as the decision stimuli. However, in the previous study that involved static images only, experts were statistically more accurate than novices. The lack of differences and moderate accuracy performance across expertise groups and stimulus conditions suggest that the three experimental tasks were all "medium" to "high" in difficulty.**

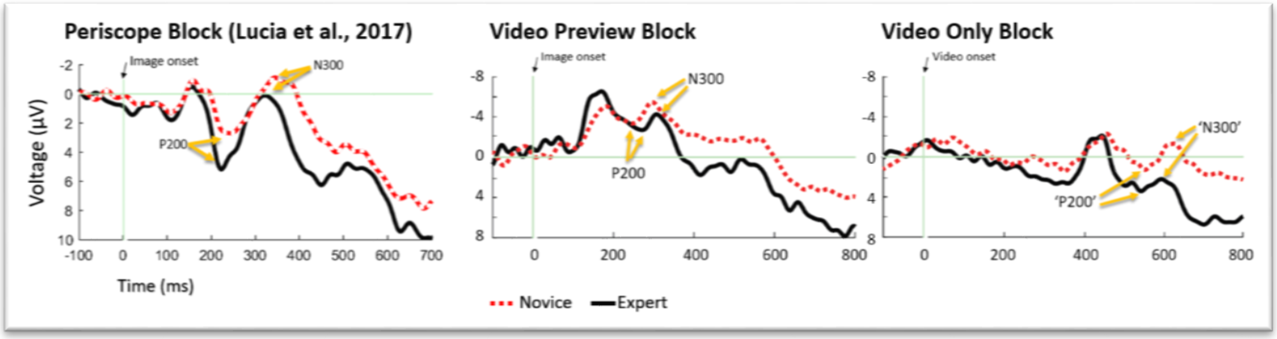

**Figure 5. Electrophysiological Results per Decision Task. This figure shows average neural activation at frontal scalp sites (F3, Fz, and F4 averaged together), with negative values plotted up (as per convention), from 100 ms prior to the stimulus onset through 700 or 800 ms. Neural activation patterns are displayed for the static periscope image task (left; see Lucia et al., 2017), the Video Preview task (middle), and the Video Only task (right).**

Video Only Task. In the VO task, scalp ERPs at focused frontal sites (see Figure 5 and Table 2) began to differ between groups at the 'P200,' which seemed to be more positive in experts than in novices. Mixed factor ANOVAs comparing 'P200' mean amplitude and peak latency (between 500-575 ms) across channel locations (F3, Fz, F4) and group (novice, expert) showed a main effect of expertise for amplitude, *F* (1,57) = 5.02, *p* = .03, but not for latency, *F* (1,57) = .57, *p* = .45. This pattern of 'P200' effects matches the results from prior work with static images (from Lucia et al., 2017); on that task, P200 amplitude was greater in experts compared to novices, but its latency did not differ between groups. The 'N300' effects were similar to those from the periscope static images task from prior work (from Lucia et al., 2017). Specifically, the 'N300' appeared to peak earlier in experts compared to novices. Mixed factor ANOVAs revealed significant main effects of expertise for both 'N300' amplitude and latency (between 550-650 ms), *F* (1, 57) = 9.85, *p* = .003 and *F* (1, 57) = 18.80, *p* < .001, respectively.

**Table 2. Event Related Potential (ERP) Analyses per Decision Task**

| | | | **P200 amplitude** | | **P200 latency** | | **N300 amplitude** | | **N300 latency** | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Block** | **N** | **df** | ***F*** | ***p*** | ***F*** | ***p*** | ***F*** | ***p*** | ***F*** | ***p*** |
| **Periscope Image (Lucia et al., 2017)** | 22 | 1,60 | 3.93 | 0.05* | 2.41 | 0.13 | 1.95 | 0.17 | 3.95 | 0.05* |
| **Video Preview** | 22 | 1,60 | 0.09 | 0.77 | 3.43 | 0.07 | 2.12 | 0.15 | 1.01 | 0.32 |
| **Video Only** | 21 | 1,57 | 5.02 | 0.03* | 0.57 | 0.45 | 9.85 | <0.01* | 18.8 | < 0.001* |

* *p* ≤.05

**Diffusion Modeling (DM)**

Performance data (RTs and accuracy rates) were subjected to DM for each task separately. DM analysis was also performed on the submarine-relevant decision task from the previous study for comparison purposes (see Table 3). The first parameter, Delta, represents the participants' average speed of processing. Higher scores suggest faster processing. For the Periscope Image task in prior work, experts revealed a higher Delta than novices, *t* (20) = -2.76, *p* = .02. By comparison, there were no between group differences in Delta for either the VP or VO tasks (*p*s > .86).

The second parameter, Alpha, represents the amount of information required to make a decision. Higher Alpha values suggest that more information is required to make a decision. This can be interpreted as response caution. There were no between group differences in Alpha (*p*s > .40) for any of the three decision tasks. When considering Alpha and Delta in combination for the Periscope Image decision task (from Lucia et al., 2017) – experts were more accurate than novices, but did not take longer to respond – it appears that experts used the *same amount of information* as the novices, but they simply *processed it faster and more thoroughly*.

The third parameter, Tau, represents non-decision time. This parameter is a measure of how long it takes participants (in milliseconds) to encode stimuli and plan and execute a response. Given that these decision tasks all used button presses, we can safely assume that the response planning and execution component is negligible. Therefore, Tau should be interpreted here as a measure of stimulus encoding. Across all three decision tasks, there were no significant between group differences (ps > .20).

The fourth and final parameter, Beta, represents response bias, such as when a participant naturally tends to select one of the two response options – such as "Yes, it is safe to continue on the current course and heading" or "No, it is not safe to continue on the current course and heading" – significantly more than the other. Across all three decision tasks, there were no significant between group differences (ps > .20).

**Table 3. Diffusion Model (DM) Analysis per Decision Task**

| **Block** | **Group** | **Delta**<br>M (SD) | | **Alpha**<br>M (SD) | | **Tau**<br>M (SD) | | **Beta**<br>M (SD) | |
|---|---|---|---|---|---|---|---|---|---|
| **Periscope (Lucia et al., 2017)** | *Novice (n=12)* | 0.20 (0.12) | * | 1.59 (0.31) | ns | 0.66 (0.19) | ns | 0.50 (0.04) | ns |
| | *Expert (n=10)* | 0.42 (0.23) | | 1.68 (0.28) | | 0.62 (0.13) | | 0.48 (0.05) | |
| **Video Preview** | *Novice (n=10)* | 0.18 (0.30) | ns | 1.57 (0.19) | ns | 0.79 (0.22) | ns | 0.53 (0.07) | ns |
| | *Expert (n=12)* | 0.16 (0.29) | | 1.66 (0.31) | | 0.67 (0.21) | | 0.50 (0.06) | |
| **Video Only** | *Novice (n=9)* | 0.25 (0.33) | ns | 1.61 (0.16) | ns | 0.68 (0.16) | ns | 0.49 (0.07) | ns |
| | *Expert (n=12)* | 0.23 (0.13) | | 1.53 (0.30) | | 0.60 (0.18) | | 0.51 (0.05) | |

* $p < .05$; ns = not significant at the $p < .05$ level

## DISCUSSION

A primary objective of the current study was to investigate whether the neural signatures of intuitive decision making, which were observed in prior research using briefly-presented static images, would be also observed when using brief motion pictures as the decision stimuli. To this end, we developed two new decision tasks – the Video Preview (VP) and Video Only (VO) decision tasks – which used brief (~ 2 second) video clips as the decision stimuli. We assumed that because video clips contain more and more useful information that do static images – for example, video clips can present trajectories over time, which static images cannot – they should elicit similar, although not necessarily identical measures of performance and cognitive functioning.

A secondary objective of the current study was to use advanced modeling techniques to investigate the processes underlying intuitive decision making. Specifically, we utilized DM which takes all performance data – including correct and incorrect RTs, along with percent accuracy – into account (Ratcliff, 1985; Ratcliff & Rouder, 1998; Ratcliff & Tuerlinckx, 2002; Wagenmakers, 2009) and generates parameter estimates of the underlying decision making components: information processing speed (Delta), amount of information required to make a decision (Alpha), non-decision time such as stimulus encoding (Tau), and response bias (Beta). We analyzed the performance data from the current study's two experimental tasks along with performance data from the submarine-relevant task in our prior study (Lucia et al., 2017).

### Generalizability of Research Findings

In terms of the demographic variables – such as POAT pre-test scores, age, and years of Navy service – there were no significant differences between the participants from the current and prior study. Similarly, the raw RT and accuracy measures seemed entirely consistent with one another. The participants all took approximately 1.2 seconds to respond, which suggests that they were following the instructions to respond with quick, intuitive-type decisions. Moreover, they responded with roughly 60% accuracy, regardless of the specific decision task (Static Periscope Image, Video Preview, and Video Only), which suggests that the three tasks were of roughly equal (i.e., medium to high) difficulty levels. Even the effective sample size was consistent across the two studies, with each expertise group averaging around 10-12 participants.

The expertise main effect from the previous study – in which experts were significantly more accurate than novices – was not replicated with either of the two video-based decision tasks. The results were surprising given that the pre-test was working as intended. Therefore, we expected the experts to outperform the novices and the inability to detect an effect was surprising. In retrospect, given the constrained range of performance (8-11% greater accuracy than chance levels of performance), across the three decision tasks and expertise groups, we believe that the three tasks were all moderately to highly difficult, and that this difficulty may have resulted in a substantial amount of "guessing" (rather than intuitive "hunches"). Moreover, what appears to be driving the statistically significant accuracy results for the 2017 study was the fact that the standard errors were smaller than in the 2018 study (i.e., there was less variability within each expertise group, respectively), which suggests that there was potentially more guessing involved on the video-based decision tasks. Given the way that the DM analyses are calculated – they leverage the raw RT and accuracy data – and given that the RT and accuracy scores were so similar across groups and tasks, it is not surprising therefore that the DM results did not reveal different findings than the standard analyses of RT and decision accuracy which are treated as separate dependent variables.

In terms of ERP analyses, the results were a bit more nuanced. In the VP task, which was meant to represent the least level of "risk" from an ERP assessment perspective, we did not detect any experience-dependent differences in neural signals of intuition. However, in the VO task, which was meant to be the most realistic because it used no static images whatsoever, the ERP results appeared more promising. Specifically, the cascade of ERPs was unusual in that previously-studied components did not appear within expected time frames. Supposing that this was an artifact of the starting characteristics of the video stimuli (cf. Sitnikova, Kuperberg, & Holcomb, 2003), we identified late versions of previously-studied components and called them 'P200' and 'N300.' Examination of these delayed ERPs revealed very similar patterns compared to the Static Periscope Image results from past work; the 'P200' was larger, and the 'N300' was earlier, in experts than in novices.

We believe that there is something about the use of video stimuli that may be causing the difference in the results. Although additional data will be required to verify this hypothesis, we believe that the use of video-based decision stimuli may have "leveled the playing field" between the experts and novices in terms of performance. For the VO task, we suspect that the continuous stream of decision making information (viewable for up to 2 seconds or until they made their response) leveled the performance playing field for the two expertise groups. By comparison, in the Static Periscope Image task (Lucia et al., 2017), decision making information was only present for 400 ms. The VO task serves as a good example of the generalizability of the neural signal of intuition. This is especially interesting as an indicator of expertise, since this ERP signal did not coincide with behavioral performance differences between groups – in essence, the ERP signal was sensitive in ways that standard measures of RT and accuracy (and by extension, DM) were not. If that is in fact what happened, then ERP-based measures could potentially be used in training related applications – such as Intelligent Tutoring Systems (ITS) – to dynamically modify the presentation of instruction as learners become more expert-like, at least in terms of their brain activity.

By comparison, the VP task – which provided a visual context for the upcoming image-based decision task – revealed no expert-novice differences for any of the dependent variables: accuracy, RT, the DM parameters, or the ERP signals. The lack of differences is particularly intriguing, because the VP was less "risky" from an EEG data collection perspective, because the actual decision was linked to a still image (rather than a video), thereby making the time synchronization between the EEG and the decision stimulus even more precise. Therefore, we expected to find between-group performance differences and ERP effects. Clearly, additional research is needed to better understand what is occurring here. That being said, we believe that it is important to present non-significant results, lest we contribute to psychology's ongoing "replication crisis" (Ferguson & Heene, 2012). By only presenting statistically significant findings, researchers ultimately do the field a disservice by removing a key tool – falsification – that helps to support a scientific field's evolving body of knowledge.

### Recommendations for Future Work and Lessons Learned

In an effort to keep the entire administration time roughly equivalent to that of the previous study (Lucia et al., 2017), we purposely limited the number of decision trials per task block. The 2017 study presented each participant with approximately 140 decision trials. By comparison, the current study averaged about half the number of decision trials per task block. This was due to the logistical constraint that each decision trial took substantially longer due to the use of video stimuli. Had we included the same number of decision trials, the time spent performing each task would likely have doubled. As a result, at this time, it is impossible to determine whether or not the effects were due to the decision stimuli (i.e., static image vs. video) or the difference in trial numbers (i.e., lower trial numbers reduce our power to detect a difference). In retrospect, if our goal was not to make such a close comparison with the previous study, we would have significantly increased the number of decision trials per participant. We believe that doing so would provide more precise mean estimates and smaller standard errors for the accuracy and RT measures, which in turn influence the accuracy of the DM parameters. While there is no single rule of thumb for determining the precise number of decision trials – some simulations recommend including 500 or more decision trials per participant (Ratcliff and Childers, 2015) – we therefore recommend that researchers attempt to collect as many trials per participant as possible. The number of decision trials in the current study, around 65 per task, was at the low end of the spectrum. Fortunately, each task block only required about 15 minutes of the participants' time. Therefore, tripling the number of decision trials would not be an onerous burden on each participant, especially if only one decision task is included (not two separate ones, as we used in the current study).

ERP assessment systems, such as the Cognionics Quick-20 Dry EEG Headset that we used in the current study, are extremely sensitive to factors such as physical movement, and which can affect the quality of the resulting neurological signal. In the current study, we were forced to exclude 3-4 participants' data per block (9-11% of the participants) because there were too few artifact-free trials. This is consistent with previous research, especially when ERP components are prominent at frontal sites where muscle activity from eye blinks can mask the signal coming from the brain (Luck, 2014; Lopez-Calderon & Luck, 2014). A pillar in the EEG/ERP community, Steve Luck of UC Irvine often quotes Hansen's axiom that "there is no substitute for clean data" (Luck, 2014), and this is indeed true. When environment and movement artifacts contaminate the signal coming from the brain, sometimes there is no amount of filtering or artifact-rejection or -correction that is capable of reviving the brain's EEG signal from the background noise. Key lessons learned here are that researchers should take feasible steps to collect clean data in the first place, and recruit ~25% more participants than needed in anticipation of losing some due to data artifacts. To minimize the collection of noisy data, researchers should: 1) eliminate ambient electrical noise (e.g., turn off or remove unnecessary electronics, run sessions in an electrically shielded room); 2) start up equipment ~30 minutes prior to data collection to allow it to warm up and stabilize; 3) ensure that the ambient temperature and humidity level are at a comfortable range (to prevent sweating); 4) select and apply an EEG cap/headset that appropriately fits the participant (to ensure proper electrode placement and reduce participant desire to adjust/stretch it); 5) have a conversation with participants explaining how movements such as blinking, sneezing, or talking will affect collected data, and; 6) ensure participants that there will be timed rest periods to recover from sitting so still and performing the task. As a note to the reader, we followed all of these best practices, and yet we still had to exclude 9-11% of our participants from the statistical analyses. This attests to the inherent difficulty of conducting ERP-related research.

In both the current and previous studies, we relied on a single assessment of periscope operator performance – the Periscope Operator Adaptive Training tool (POAT; Landsberg et al., 2012) – to generate pre-test scores for classifying participants into novice and expert categories, respectively. While the POAT has amassed some evidence supporting its use as a training tool, and we have no doubt that the POAT scores do help to differentiate novices from experts, this approach was chosen because of expediency. Expertise is generally defined as the ability to consistently demonstrate superior levels of task-related performance (Ericsson et al., 1993). It is not determined based on the results of a single test administration. In an ideal world, we would have administered the POAT on multiple separate occasions, and then used the aggregate data to help classify submariners into expertise categories. However, access to military samples is extremely time-limited, especially when the researchers have no control over the participant's schedule availability. Therefore, we strongly recommend other researchers to include multiple measures of proficiency when classifying participants into expertise groups (whenever possible), although we fully recognize that this is difficult to implement in practice.

## Summary and Conclusion

This study is one of the first systematic attempts to use video-based stimuli for assessing intuitive decision making skills using ERP-based methods. We firmly believe that video provides more, and potentially more useful, information for decision making than that which is conveyed using briefly-presented static imagery. However, it is also more challenging from a neurocognitive assessment perspective because of the millisecond-level precision that is required for assessment. The initial findings described here are promising. For example, we were able to identify theoretically meaningful amplitude and latency effects that were associated with selective attention and knowledge-driven pattern matching, respectively. Curiously, the ERP effects were observed even when more standard metrics of RT – including accuracy, response time, and diffusion modeling – failed to differentiate novices from experts. Additional research is needed to determine if these ERP effects generalize to other samples and task types before they could be used for other purposes, such as for tailoring the content displayed in Intelligent Tutoring Systems (ITS).

## ACKNOWLEDGEMENTS

Funding for this study was provided by the Office of Naval Research to Aptima, Inc. The views expressed in this paper are solely those of the authors, and do not necessarily reflect the opinions of their sponsors at the Office of Naval Research, the Naval Air Warfare Center Training Systems Division, or any other Department of Defense agency, unless stated in official directives. The authors wish to thank the reviewers of this paper for their insightful feedback.

## REFERENCES

Bolte, A. & Goschke, T. (2005). On the speed of intuition: Intuitive judgments of semantic coherences under different response deadlines. *Memory & Cognition*, 33 (7): 1248-1255.

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*(3), 363-406.

Evans, J. S. B. T. (2003). In two minds: Dual process accounts of reasoning. *Trends in Cognitive Sciences, 7*(10), 454-459.

Evans, J. S. B. T., & Stanovich, K. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives in Psychological Science, 8*(3), 223-241.

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science’s aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Kahneman, D. (2011). *Thinking: Fast and slow*. New York: Farrar, Strauss, & Giroux.

Kahneman, D. & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist*, 64 (6), 515-526.

Klein, G. (2008). Naturalistic decision making. *Human Factors, 50*(3), 456-460.

Landsberg, C. R., Mercado, A. D., Van Buskirk, W. L., Lineberry, M., & Steinhauser, N. (2012). Evaluation of an adaptive training system for submarine periscope operations. In *Proceedings of the 56th Annual Meeting of the Human Factors and Ergonomics Society*, 56, 2422.

Lopez-Calderon, J., & Luck, S. (2014). ERPLAB: an open-source toolbox for the analysis of event-related potentials. *Frontiers in Human Neuroscience*, 8, 213

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. New York: Oxford University Press.

Luck, S. J. (2014). *An Introduction to the Event-Related Potential Technique, 2nd Edition*. Cambridge, MA: MIT Press.

Lucia, L. C., Beaubien, J. M., Steinhauser, N., & Steed, R. (2017, December). Assessing Submariners’ Intuitive Decision making Skills Using Neurocognitive Methods. In *Proceedings of the 2017 Interservice/Industry Training, Simulation, and Education Conference*. Arlington, VA: National Training and Simulation Association.

Luu, P., Geyer, A., Fidopiastis, C., Campbell, G., Wheeler, T., Cohn, J., & Tucker, D. M. (2010). Reentrant processing in intuitive perception. *PloS One*, 5, 9523.

Maguire, M. J., Magnon, G., Ogiela, D. A., Egbert, R., & Sides, L. (2013). The N300 ERP component reveals developmental changes in object and action identification. *Developmental Cognitive Neuroscience*, 5: 1-9.

Phillips, S. & Takeda, Y. (2009). An EEG/ERP study of efficient versus inefficient visual search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 31, 383-388.

Picton, T. W., Bentin, S., Berg, P., Donchin, E., Hillyard, S. A., Johnson, R. Jr., Miller, G. A., Ritter, W., Ruchkin, D. S., Rugg, M. D., & Taylor, M. J. (2000). Guidelines for using human event-related potentials to study cognition: recording standards and publication criteria. *Psychophysiology*, 37, 127–152.

Ratcliff, R., & Childers, R. (2015). Individual differences and fitting methods for the two-choice diffusion model of decision making. *Decision, 2*(4), 237-279.

Ratcliff, R. & Rouder, J.N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, 9, 347-356.

Ratcliff, R. & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reactions times and parameter variability. *Psychonomic Bulletin & Review*, 9, 438-481.

Ratcliff, R. (1985). Theoretical interpretations of the speed and accuracy of positive and negative responses.

*Psychological Review*, 93, 212-225.

Reber, P. J. (2013). The neural basis of implicit learning and memory: a review of neuropsychological and neuroimaging research. *Neuropsychologia*, 51 (10), 2026-42.

Schendan, H. E. & Ganis, G. (2015). Top-down modulation of visual processing and knowledge after 250 ms supports object constancy of category decisions. *Frontiers in Psychology*, 6 (1), 1289.

Snodgrass, J. G. & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, 6 (2), 174–215.

Wagenmakers, E-J. 2009. Methodological and empirical developments for the Ratcliff diffusion model of response times and accuracy. *European Journal of Cognitive Psychology*, 21 (5): 641-671.

Wagenmakers, E.-J., van der Maas, H. L. J., & Grasman, R. P. P. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, 14, 3-22.