Bracketing Human Performance to Support Automation for Workload Reduction: A Case Study

Robert E. Wray(✉), Benjamin Bachelor, Randolph M. Jones, and Charles Newton

Soar Technology, Inc., 3600 Green Court Suite 600, Ann Arbor, MI 48105, USA
{wray, ben.bachelor, rjones, charles.newton}@soartech.com

Abstract. Semi-automated Forces (SAFs) are commonly used in training simulation. SAFs often require human intervention to ensure that appropriate, individual training opportunities are presented to trainees. We cast this situation as a supervisory control challenge and are developing automation designed to support human operators, reduce workload, and improve training outcomes. This paper summarizes a combined analytic and empirical verification study that identified specific situations in the overall space of possible scenarios where automation may be particularly helpful. By bracketing "high performance" and "low performance" conditions, this method illuminates salient points in the space of operational performance for future human-in-the-loop studies.

Keywords: Simulation-based training · Semi-automated forces · Cognitive workload

1 Introduction

Semi-automated Forces (SAFs) are commonly used in training simulation to represent the behavior of enemy, friendly, and neutral entities within the training exercise. Most SAFs in current simulations are adaptive to the doctrinal context of the mission, such as the application of appropriate tactics in the situation. However, they are typically unaware of the learning context, such as the current training objectives, the estimated level of skills of the trainee, and assessment of trainee actions, relative to the training objectives, as the scenario progresses.

When there is a mismatch in the appropriate tactical decision and the learning context, human intervention is required. For example, a pilot trainee might be expected to achieve or demonstrate specific competencies within a defined training event. As a consequence, human intervention during the simulation to enable achievement of training objectives is often a requirement for effective training.

When variation from pre-programmed SAF behavior is appropriate, whether to reflect real-world situations or to reinforce instruction (e.g., consequences of an error), the instructor must recognize this need, choose a course of action, and then issue direction to operators. The instructor desires to control or "steer" the training scenario

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2015, LNAI 9183, pp. 153–163, 2015.
DOI: 10.1007/978-3-319-20816-9_16

toward particular milestones and outcomes. The instructor has, by design, limited control over the trainee, so a dynamic control process is needed to adjust scenario evolution in response to trainee actions.

For tactical aircraft pilot training, we have developed an instructional support tool, the Training Executive Agent (TXA), designed to facilitate and simplify this control problem [1]. This paper describes the results of a software verification experiment to evaluate the functionality of the TXA. The primary goal of the study was to compare the relative quality of training experiences without and with the TXA as scenario complexity increased. In conjunction with this empirical evaluation, analytic methods were used to estimate anticipated operator workload over routine and non-routine control profiles. Together, the results provide high-end and low-end expectations (brackets) for human-in-the-loop experimentation. The importance of this bracketing it helps to identify salient points in a large space of possible training scenarios and differing complexities to sample for actual human-in-the-loop experimentation. We describe goals, methodology and results and consider the potential value of this combination of analytic and empirical methods in the design of supervisory control systems more generally.

2 Context and Motivation

As suggested above, today's training typically requires some direct human control of SAFs to achieve desired training outcomes. SAFs are reactive to the doctrinal context of the mission. However, they unaware of the learning context, such as the current training objectives, the estimated level of skills of the trainee, and an assessment of actions, relative to the training objectives, as the scenario progresses. When there is a mismatch in the tactical context and learning context, human intervention is required. For example, a pilot trainee is expected to achieve or demonstrate specific competencies within a defined training event. As a consequence, human manipulation of the simulation to enable the presentation of situations appropriate for the training objectives is a critical part of the training: a missed bogey (enemy aircraft) may be manually relocated to support a trainee's inefficient or misdirected radar search to ensure that the trainee also has an opportunity to engage ("intercept") the bogey; an extra bogey may be inserted to challenge a trainee that is excelling in demonstrating basic intercept tactics.

When deviation from the pre-programmed SAF behavior is appropriate, the instructor must recognize this need, choose a course of action, and then issue tell operators to modify the simulation. From a control perspective, the instructor desires to drive the training scenario toward particular milestones and outcomes. The instructor has, by design, limited control over the trainee, so a dynamic control process is needed to adjust scenario evolution in response to trainee actions. Further, the simulation presents only a tactical summary of the situation to the instructor, which requires the instructor to maintain a separate mental representation of the learning context and possible implications of the current situation on the learning context.

From this assessment, the instructor determines which adjustments are necessary and acts to execute them. In the operational training context, the recommended adjustments are verbally communicated to a human operator who then issues command to individual

SAFs, monitors their execution, and makes low-level adjustments to behavior. A consequence of this distribution of control, however, is that the operators may not necessarily understand the intent of the instructor's directives because the learning context that motivates the adjustments may remain implicit (e.g., not verbalized).

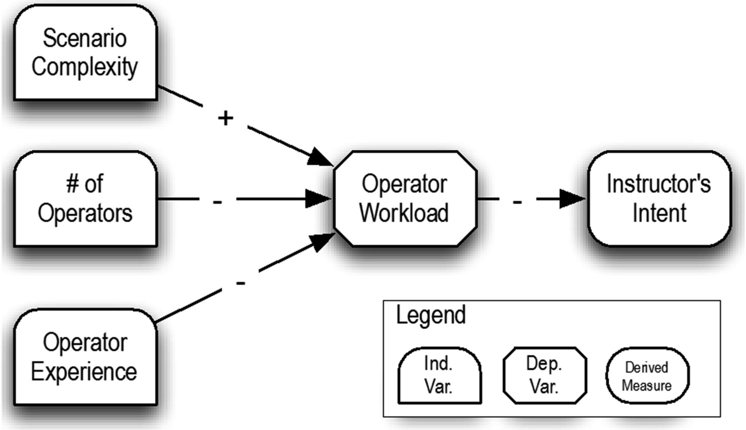


Fig. 1. Summary of relationships between independent and dependent variables for constructive simulation.

From this perspective, the targeted functions of the TXA are two-fold. First, it provides a more explicit representation of the scenario and learning context to the instructor and operator. Second, it shifts the manual control process for individual SAFs to a supervisory control process [2] in which an operator (or instructor/operator) uses the summary information from the TXA to make comparatively infrequent or periodic adjustments to the system via a higher-level action representation. In this role, the TXA acts as a mediator or controller of individual SAFs, similar to the role of a real-time strategy game player [3]. However, unlike previous SAF control approaches, the TXA makes decisions and recommendations based on both the tactical situation (as do most controllers do) and the learning context (uncommon).

The hypothesized relationships between human-control variables in the context of constructive simulation are illustrated abstractly in Fig. 1. We assume that quality of training is consonant with meeting instructor intent; that is, if the training scenario delivers the experiences that the instructor intends to deliver, the training experience is high quality. Meeting instructor intent is negatively correlated with operator workload: as workload increases, the ability of the overall system to match instructor intent decreases. Operator workload is itself a dependent variable. It positively correlates (goes up) with scenario complexity, and it negatively correlates with number of operators and operator experience.

3 Experimental Goals and Methodology

The primary use case for TXA testing and evaluation is a current U.S. Navy program of instruction. The training focuses on teaching new pilots (i.e., pilots recently graduated flight school) how to fly a specific platform. The initial stages of training focus on cockpit familiarization; later stages focus progressively on making good decisions relative to the platform capabilities; that is, how to use the platform in relatively simple tactical contexts. The TXA evaluation focuses on the later stages of training where a mission context is important to the training experience.

Currently, 2–4 people support individual pilot training in a simulation. An instructor guides the process and directs operators to adapt a training scenario. An expert pilot may fly as a lead or wing (depending on the training goals) with the trainee pilot when the trainee has advanced to section-level tactics. In some cases, the expert pilot is also the instructor. Simulation operators control the pilot simulation and other technology coordination/interoperation.

Simulation operator(s) also control and guide SAFs and set simulation parameters and variables to support the training (at direction of the instructor). The TXA is targeted primarily at reducing the workload of this simulation operator, by improving misionization and behavior fidelity and reducing the number of interventions needed to achieve behavioral changes made at the request of the instructor. In other words, the TXA should enable a high span of control while maintaining high scenario quality/satisfy instructor's intent. The TXA is not designed to improve the fidelity that can be provided by human operators but rather provide force multiplication for the operator, making it possible for a human operator to maintain high levels of scenario quality in more complex scenarios than is currently feasible with manual control.

We conducted a software test evaluation to begin to assess the specific relationships illustrated in Fig. 1. We assumed in this initial test design that the number of available

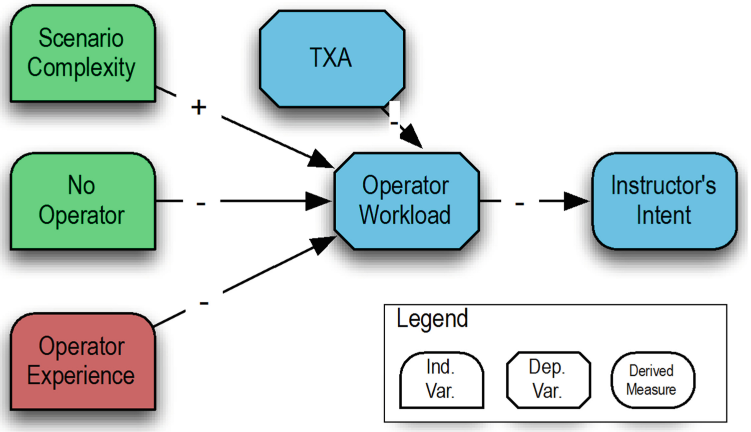


Fig. 2. The TXA is hypothesized to reduce workload and improve the ability of the training system to meet instructor's intent.

operators and the level of operator experience/capability are scarce resources that cannot be significantly improved without increasing training costs. The implication then is that increased scenario complexity will increase operator workload, which in turn will decrease presentation quality (match with instructor intent), thus decreasing overall training quality.

The further hypothesis is that the TXA's functions will positively influence the quality of presentation or match to instructor intent (Fig. 2). By providing operators and instructors with high-level control actions, operator workload will decrease which will improve (or maintain) presentation quality.

We manipulated scenario complexity as the primary independent variable. To conduct actual tests, we developed 20 aviation-training scenarios, reflecting differences in subjective complexity as determined by subject matter experts. Each test case scenario was composed from individual setups, representing tactical situations one might encounter in a training program and the training goals associated with each scenario. For example, in the example in Fig. 3, the training goal is to give the trainee (blue pilot) the experience of combating two distinct groups of aggressor (red) aircraft. The instructor's goal is to present these groups as two successive but independent engagements for the trainee.

For each setup, we had subject matter experts define criteria for assessing the overall quality of the training experience as presented to a trainee. For example, in the situation illustrated in Fig. 3, maintaining some distance between the groups is important, but the trailing group needs to remain close enough to the blue aircraft that when the first engagement is completed the trainee has the experience of a "successive" engagement. Subject matter experts provided specific criteria and parameters. We used a scenario monitoring tool, the Goal Constraint System [4] to encode these criteria and automatically "score" individual setups.

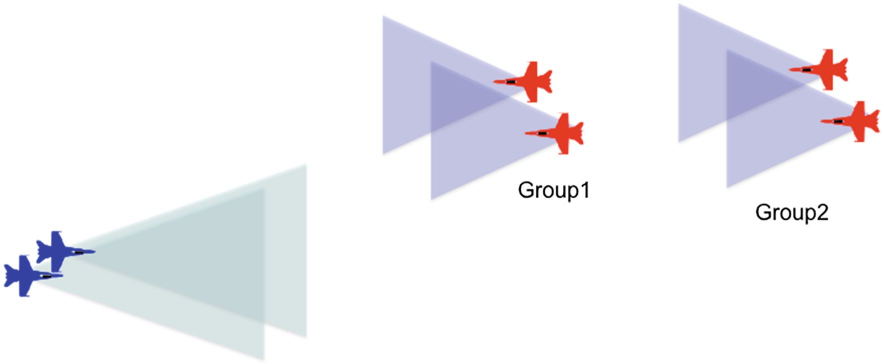


Fig. 3. Experimental testcases are composed from relatively simple tactical situations

We then developed (simple) computational models of trainees, representing "normal" or expected trainee actions and "novice" trainee actions that could diverge from the pre-programmed scenario assumptions. For example, in the situation illustrated in

Fig. 3, a “normal” trainee would directly engage the first group. A “novice” trainee might take doctrinally unnecessary evasive action to attempt to separate the two groups. The role of these models is to provide some sampling of the variation one might expect to encounter in the training program.

Finally, based on the setups and testcases, we developed analytic models of the workload demands for each setup. We used the Goals, Operators, Methods, and Selection Rules (GOMS) modeling paradigm [5, 6]. We developed a methodology and estimates of workload for each setup and testcase [7]. For the purposes of this paper, the results of this analysis provided estimates of the time it would take an operator to monitor a setup and to intervene under certain anomalous conditions. Table 1 summarizes the prediction for the setup illustrated in Fig. 3. It says the minimum amount of time the operator can attend to the setup and it remain close to 100 % of its performance quality is about 22 s over the course of 1 min of execution time. Based on this lower bound, the analysis tells us that the operator could successfully manage up to two of these setups simultaneously (i.e., managing three setups perfectly would require 66 s of operator monitoring and action/minute).

Table 1. Example operator performance bounds predicted by GOMS for Fig. 3 setup

	Lower bound	Max. number of manageable setups
Successive intercepts	22.2 s	2

4 Empirical Test Results

The overall results of the experiment are depicted in Fig. 4. Numbers on the X-axis identify each test case scenario, in order of estimated scenario complexity. The Y-axis value represents the presentation quality from observed behavior for that test case, for the control (No TXA) and test (TXA) conditions. This graph indicates confirmation of both experimental hypotheses. First, with some minor variation, the observed presentation quality for each scenario in the “No TXA” condition correlates negatively with the estimated complexity for the scenario. Second, the observed presentation quality for each scenario in the TXA condition is largely independent of scenario complexity and maintains an overall high value. This result indicates the contribution of the TXA in dynamically managing and tailoring entity behaviors to maintain an instructor’s goal for the exercise in response to variations (e.g., mistakes) in trainee actions.

Figure 5 presents the experimental results organized by estimated complexity. These figures also contain the quantitative presentation quality scores omitted in Fig. 4. For these summaries, we divided the test-case results into three categories. Low complexity test cases are those test cases that we predict a human operator would find manageable to perform the same types of tailoring as the TXA performed to maintain high presentation qualities. That is, in a future experiment with humans in the loop, we predict human performance to be comparable to automated TXA performance.

We predict medium complexity scenarios to be near the edge of manageability for human operators. We predict that human operators would, for the most part, not be able

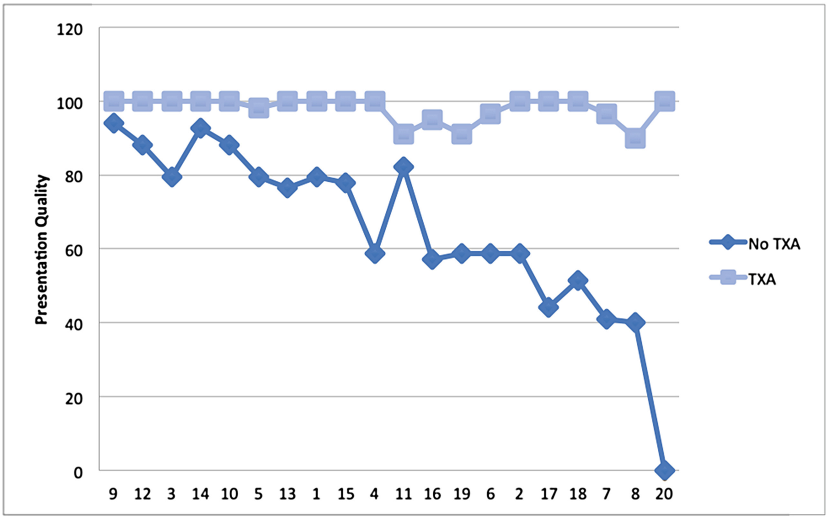


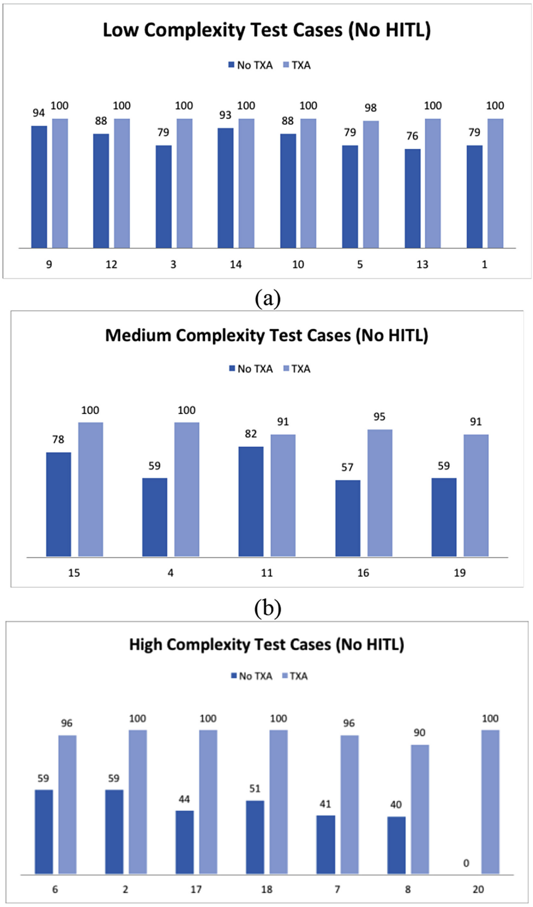
Fig. 4. Overall summary of the verification results with and without TXA automation

to produce the same quality of presentation as TXA is able to produce, for these test cases, although human performance results in this band may differ widely based on expertise and innate capacities. However, high complexity test cases are predicted to be of such significant complexity that they are completely beyond the abilities of human operators to manage them. We predict low presentation quality scores by human operators for these scenarios.

We see that relatively high quality scores are observed even without the assistance of the TXA for the low-complexity scenarios (Fig. 5a). For the medium-complexity scenarios, we see larger differences in presentation quality for most of the scenarios (Fig. 5b). For the high-complexity scenarios, there is a significantly large difference in quality between the No TXA and TXA conditions (Fig. 5c). The most complex scenario (Scenario 20) produces the worst possible quality score in the No TXA condition, and the best possible quality score in the TXA condition.

5 Bracketing for Human-in-the-Loop Experimentation

The results of these experiments suggest some potential value of automation. However, comparing a system that requires intervention to achieve its goals without the means to deploy interventions, as in the “No TXA” case, offers little insight for estimating the operational impact of the automation. Further, replicating this experiment with human operators, across all the scenarios, would be prohibitively costly. Instead, we can combine the empirical results and some additional analysis using the GOMS approach to attempt to “bracket” expected human performance. This bracketing heuristic is adopted from Kieras and Meyer [8]. The intention is that the GOMS analysis provides a priori predictions of expected performance across the range of scenarios. We can then



(a)

(b)

(c)

Fig. 5. Quantitative experimental results for low-complexity test cases (a), medium-complexity (b) and high complexity (c).

choose specific scenarios to test the predictions. But we choose scenarios that would also help most fully contrast the impact of automation without exploring the full set of scenarios.

Figure 1 illustrated the prediction that a training scenario's presentation quality will decrease as the estimate of the operator's workload increases for a particular scenario. However, there is a "workload threshold" below which the operator will be able to maintain a close to perfect presentation quality. We would like to determine where this threshold occurs because testing scenarios "below the threshold" is unlikely to show much value for automation. According to the GOMS analysis, this threshold can be determined primarily by "the number of times an operator switches attention between setups." This specific value for any individual is dependent on the switching strategy they choose for that individual testcase as well as operator expertise and tendencies. For

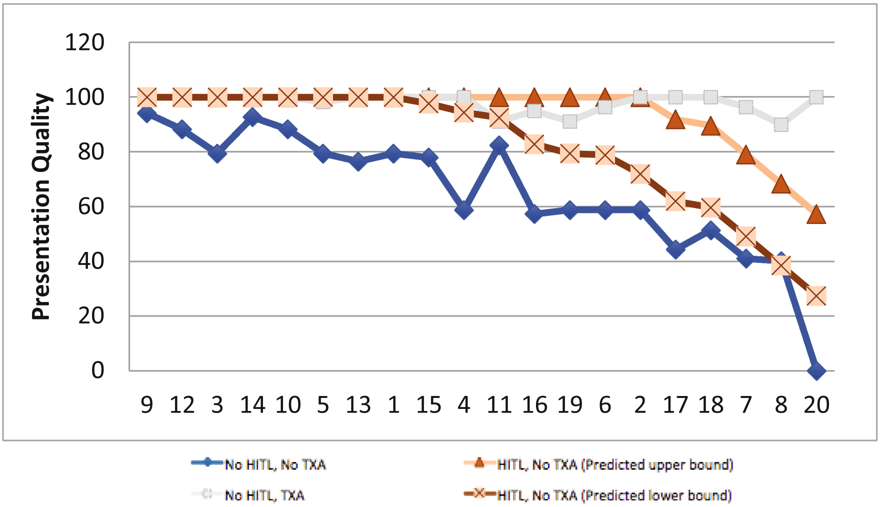


Fig. 6. Bracketing predictions of expected human performance

example, some operators might cycle deliberately thru each testcase while others might attempt to anticipate time-critical events in individual setups. These factors are targets for a future more refined GOMS analysis.

Rather than attempt to explicitly model the impact of these different strategies (as is done by Kieras and Meyer), we instead generated thresholds based on pilot-study observations of operator performance (both on the test cases and more generally). Figure 6 illustrates the predicted operator performance bracket from the GOMS analysis for the "No TXA" condition. The brackets were generated from the workload analysis illustrated in Table 1. The upper bound assumes that an operator can perfectly use all available time (60 s). Thus, if a testcase with three setups had limits of 20 s, 30 s, and 10 s, the upper bound assumes that the operator would manage all of these cases without error. We expect that most operators will use some time within each minute to perform switching. Thus, this upper bound is likely to be liberal in its estimate of operator performance. The lower bound assumes that the operator is half as efficient as the high bound (30 s of work per 60 s of simulation time).

The brackets suggest that even for the "medium" complexity cases, some operators may be able to maintain very high presentation quality. Pilot studies focused on test cases between 11 and 6 (11, 16, 19, 6) should help to determine how optimistic the upper performance bound is and localize the actual "bend in the curve" in human performance data. Our expectation is that the operator's sense of subjective workload would be quite high under these conditions in comparison to the low workload conditions, even if presentation quality differences were not evident. As they stand prior to any pilot testing, the brackets suggest that human in the loop studies should focus primarily on the high complexity cases. Further, we should work with subject matter experts to evaluate points in the curriculum that require or would benefit from presentation of scenarios with comparable levels of complexity.

6 Discussion and Conclusions

The primary conclusions of this experimental analysis are that the two experimental hypotheses are tentatively confirmed: presentation quality in the No TXA case is generally negatively correlated with scenario complexity, and presentation quality in the TXA case is consistently high. This analysis also provides corroborative evidence that the GOMS-based complexity estimates are reasonable. They allow us to predict results for future human-in-the-loop experiments using the same set of test-case scenarios. As a consequence of these verification and bracketing steps, subsequent human-in-the-loop studies can be targeted to the most important and salient scenarios and experimentation conditions, which will save time and reduce experimentation cost, especially in comparison to a human-in-the-loop study over all the usage conditions. Thus, in addition to the specific experimental results, the experimental methodology we document may also be a contribution for future evaluations of automation over a large space of potential usage conditions.

Although the experimental results suggest the validity of the primary hypotheses, they also provide useful information for improving the technology, experiments, and analyses in future work. For example, the quality measures in the experimental results do not completely correlate with the test-case complexity scores. It is not necessarily the case that these two measures must correlate, but the fact that there are inconsistencies between the mappings and the subsequent results point to potential opportunities to examine assumptions. These assumptions include the criteria for scoring presentation quality, as well as assumptions about estimating scenario complexity. Although we worked with SMEs to generate the presentation quality criteria, the rationale for some of the scoring formulas was not always straightforward. For example, we conceived of presentation quality as a continuous function, but some types of setups are more “binary” in nature (“perfect” or “unacceptable” was the way the SME framed the scoring in this case). Because there are both “binary” and “continuous” forms of presentation quality in the setups, some additional effort to analyze the qualitative differences between these two classes of setups may be justified, which would potentially result in refinements to the scoring criteria. Additionally, we should further analyze the question of whether presentation quality ought to correlate negatively with scenario complexity in all cases. It may be that complexity and cognitive workload manifest themselves in different ways for different types of setups.

Acknowledgments. This work is supported by the Office of Naval Research project N00014-1-C-0170 Tactical Semi-Automated Forces for Live, Virtual, and Constructive Training (TACSAF). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Department of Defense or Office of Naval Research. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon. We would like to thank collaborators and sponsors at NAWC TSD and ONR who have provided insights and operational perspectives in the development of TXA: LCDR Brent Olde, Ami Bolton, Melissa Walwanis, and Heather Priest.

References

1. Wray, R.E., Woods, A.: A cognitive systems approach to tailoring learner practice. In: Laird, J., Klenk, M. (eds.) Proceedings of the Second Advances in Cognitive Systems Conference, Baltimore, MD (2013)
2. Sheridan, T.B.: Humans and Automation: System Design and Research Issues. Wiley, New York (2002)
3. Wray, R.E., van Lent, M., Beard, J.T., Brobst, P.: The design space of control options for AIs in computer games. In: IJCAI 2005 Workshop on Reasoning, Representation, and Learning in Computer Games, pp. 113–118. US Naval Research Laboratory (2005)
4. Jones, R.M., Bachelor, B., Stacy, W., Colonna-Romano, J.: Automated monitoring and evaluation of expected behavior. In: International Conference on Artificial Intelligence, Las Vegas (in preparation)
5. John, B.E.: Why GOMS? Interactions **2**, 80–89 (1995)
6. John, B.E., Kieras, D.E.: The GOMS family of user interface analysis techniques: comparison and contrast. ACM Trans. Comput. Hum. Interact. **3**, 320–351 (1996)
7. Jones, R.M., Wray, R.E., Bachelor, B., Zaientz, J.: Using cognitive workload analysis to predict and mitigate workload for training simulation. In: Proceedings of the Applied Human Factors and Ergonomics Conference 2015, Las Vegas (2015)
8. Kieras, D.E., Meyer, D.E.: The role of cognitive task analysis in the application of predictive models of human performance. In: Schraagen, J.M.C., Chipman, S.E., Shalin, V.L. (eds.) Cognitive Task Analysis. Lawrence Erlbaum, Mahwah, NJ (2000)