INSTITUTE FOR DEFENSE ANALYSES

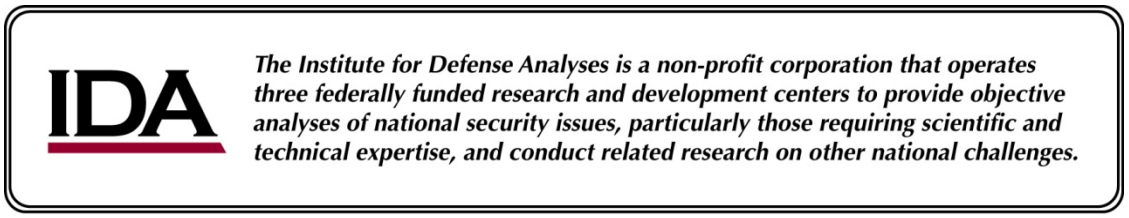
A Multi-Method Approach to Evaluating Human-System Interactions during Operational Testing

Dean Thomas, *Project Leader*
Heather Wojton
Chad Bieber
Daniel Porter

November 2017
Approved for public release.
IDA NS D-8857
Log: H 2017-000658

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.



About This Publication

The quality of human-system interactions is a key determinant of mission success for military systems. However, operational testers rarely approach the evaluation of human-system interactions with the same rigor that they approach the evaluation of physical system requirements, such as miss distance or interoperability. Often, testers evaluate human-system interactions solely using survey instruments (e.g., NASA-Task Load Index (NASA-TLX)), excluding other methods entirely. In this paper, we argue that a multi-method approach that leverages methodological triangulation provides greater insights into human-system interactions observed during operational testing. Specifically, we present data from an operational test in which a multi-method approach was used. Ten attack helicopter pilots identified and responded to threats under four conditions: high vs. low threat density and presence vs. absence of a threat detection technology. Testers recorded two primary measures of pilot workload: time to detect first threat and the NASA-TLX. Pilots took significantly longer to detect threats under low threat density than high threat density when the threat detection technology was absent. However, there was no difference in time to detect threats when the threat detection technology was present. The NASA-TLX data showed a similar pattern of results, suggesting that the observed effect is a result of pilot workload rather than the method used to measure workload – i.e., survey instrument vs. behavioral metric. Triangulating methods in this way provides a more rigorous and defensible test of the research question, and when combined with qualitative methods, provides useful information for identifying whether degradations in performance should be addressed through additional training or interface redesign.

For more information:
Dean Thomas, Project Leader
dthomas@ida.org • (703) 845-6986

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Approved for public release; distribution is unlimited.

INSTITUTE FOR DEFENSE ANALYSES

IDA NS D-8857

A Multi-Method Approach to Evaluating Human-System Interactions during Operational Testing

Dean Thomas, *Project Leader*
Heather Wojton
Chad Bieber
Daniel Porter

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

A multi-method approach to evaluating human-system interactions during operational testing

Heather Wojton, Chad Bieber, Daniel Porter
Institute For Defense Analyses
Alexandria, VA

ABSTRACT

The quality of human-system interactions is a key determinant of mission success for military systems. However, operational testers rarely approach the evaluation of human-system interactions with the same rigor that they approach the evaluation of physical system requirements, such as miss distance or interoperability. Often, testers evaluate human-system interactions solely using survey instruments (e.g., NASA-Task Load Index (NASA-TLX)), excluding other methods entirely. In this paper, we argue that a multi-method approach that leverages methodological triangulation provides greater insights into human-system interactions observed during operational testing. Specifically, we present data from an operational test in which a multi-method approach was used. Ten attack helicopter pilots identified and responded to threats under four conditions: high vs. low threat density and presence vs. absence of a threat detection technology. Testers recorded two primary measures of pilot workload: time to detect first threat and the NASA-TLX. Pilots took significantly longer to detect threats under low threat density than high threat density when the threat detection technology was absent. However, there was no difference in time to detect threats when the threat detection technology was present. The NASA-TLX data showed a similar pattern of results, suggesting that the observed effect is a result of pilot workload rather than the method used to measure workload – i.e., survey instrument vs. behavioral metric. Triangulating methods in this way provides a more rigorous and defensible test of the research question, and when combined with qualitative methods, provides useful information for identifying whether degradations in performance should be addressed through additional training or interface redesign.

ABOUT THE AUTHORS

Dr. Heather Wojton is a Research Staff Member at the Institute for Defense Analyses (IDA) supporting the Director, Operational Test and Evaluation in the Office of the Secretary of Defense (OSD/DOT&E). She provides expertise in the evaluation of human-system interactions, survey methods, and research design. She currently aids in the test and evaluation of a broad range of major defense systems, including both training and operational aircraft, and information systems. Prior to taking a position at IDA, she obtained her PhD in Experimental Psychology from the University of Toledo where her research focused on how social and contextual information shapes human cognition and behavior. Dr. Wojton is currently interested in measuring how trust affects people's behavior toward complex technologies and improving methods for measuring common human-system constructs, including workload and usability.

Dr. Chad Bieber is a Research Staff Member at the Institute for Defense Analyses supporting the Director, Operational Test and Evaluation in the Office of the Secretary of Defense (OSD/DOT&E). He provides expertise in operational test and evaluation of large aircraft as well as human-systems interaction and autonomous systems. He was an instructor pilot in the US Air Force with 2500 hours in the C-5 and T-1 aircraft and created a system of multiple autonomous UAVs while completing his PhD in Aerospace Engineering from North Carolina State University. Dr. Bieber is currently interested in measuring how people interact with complex and autonomous systems.

Approved for public release; distribution is unlimited.

A multi-method approach to evaluating human-system interactions during operational testing

Heather Wojton, Chad Bieber, Jonathan Snavely
Institute For Defense Analyses
Alexandria, VA
hwojton@ida.org, cbieber@ida.org, jsnavely@ida.org

There are three broad approaches to research practiced within the scientific community: quantitative, qualitative, and multi-method approaches. Multi-method approaches are most widely practiced within the social, behavioral, and human sciences where the combination of both quantitative and qualitative techniques is recognized as valuable in providing comprehensive, defensible answers to research questions (Campbell & Fiske, 1959; Johnson, Onwuegbuzie, & Turner, 2007). Despite its popularity in academia, however, multi-method approaches are absent in many areas of applied research where drawing incorrect conclusions from data is costly both in terms of monetary resources and human lives. During the operational testing of weapon systems, for instance, the military evaluates the quality of human-system interactions almost exclusively using survey instruments. Consideration and inclusion of methods with differing sources of measurement error, such as behavioral measures or observational techniques, is quite rare. The purpose of this paper is to identify the shortcomings of a single-method approach to evaluating human-system interactions and offer an alternative, multi-method approach that is more defensible, yields richer insights into how operators interact with weapon systems, and provides a practical method for identifying when the quality of human-system interactions warrants correction through either operator training or redesign.

HUMAN-SYSTEM INTERACTIONS IN OPERATIONAL TESTING

The quality of human-system interactions is a key determinant of mission success for weapon systems (Tillman, Fitts, Woodson, Rose-Sundholm, & Tillman, 2016). Though the future promises autonomy, most weapon systems employed by the U.S. military currently require inputs from operators to conduct missions (Defense Science Board, 2012, 2016). Even "unmanned" vehicles, such as the Predator and Reaper, are controlled by operators from remote ground stations. Poor human-system interactions – particularly, those that are ineffective or inefficient – are more likely to produce errors and lead to operator fatigue, which can degrade mission accomplishment and place operators at greater risk of injury or death. The Costa Concordia cruise ship sunk in 2012, for example, because the captain manually diverged from the route selected by the ship's automated navigation system, crashing into a shallow reef and killing 32 passengers (Levs, 2012). In 2009, Turkish Airlines flight 1951 crashed because the pilots continued to rely on the plane's automatic pilot after an altitude measuring instrument failed, killing 9 people, including all 3 pilots (Hoff & Bashir, 2015). Similarly, in 2009, Air France flight 447 crashed when the automatic pilot disconnected following pitot system icing and the crew was unable to correctly analyze the malfunction. (BEA, 2012). It is critical, therefore, that the Department of Defense and the Services prioritize the need for positive human-system interactions in the acquisition and modernization of new and existing weapon systems.

In the 1980s, Congress formally recognized the important role that human-system interactions play in mission success by establishing the Director, Operational Test and Evaluation (DOT&E) whose primary goal is to evaluate how weapon systems perform under realistic combat conditions when employed by trained operators (DoD Authorization Act of Fiscal Year 1984; Defense Directive 5141.2). DOT&E sets policy regarding the design, conduct, and analysis of operational tests, and works closely with each Services' operational test agency to implement these policies and ensure that operational tests are both rigorous and defensible. In total, DOT&E has issued 4 policy memos since 2014 (http://www.dote.osd.mil/guidance.html) directed at improving the validity of survey instruments designed to evaluate the quality of human-system interactions during operational tests. Specifically, these policies direct operational test agencies to:

1. Construct survey instruments according to best practices identified within the academic literature on survey methodology (e.g., Dillman, Smyth, & Christian, 2009)
2. Leverage survey instruments from industry and academia that are recognized as valid measures of key constructs, such as system usability (Bangor, Kortum, & Miller, 2009; Borci, Federici, Bacci, Gnaldi, & Bartolucci, 2015), task workload and operator fatigue (Charlton, 1991; Gawron, Schifflet, & Miller, 1989; Hart, 2006)

Approved for public release; distribution is unlimited.

3. Administer survey instruments systematically across test conditions according to established social science research methods and statistical design of experiments techniques (Montgomery, 2012; Somekh & Lewin, 2011).

These policies are slowly being adopted by each Services' operational test agency. To our knowledge, however, no guidance currently exists regarding alternative methods for evaluating human-system interactions.

Such heavy reliance on survey instruments at the expense of other methods is partly due to a lack of social science and human factors expertise within the operational testing community. A recent study funded by DOT&E found that in 2015, 63 percent of operational test agencies' professional and technical staff members held degrees in the biological and physical sciences, engineering, computer science, mathematics and statistics (Snavely, Wojton, Bieber, & Freeman, 2017). By contrast, less than 4 percent of professional and technical staff members held degrees in the social sciences, with only 1 percent holding a degree in psychology – the social science whose subject matter is most directly relevant to evaluating how operators interact with weapon systems. The fact that social scientists make up such a small proportion of the operational test community means that the community has limited knowledge of the kinds of methods that currently exist to evaluate human thought and behavior and how best to implement these methods.

Survey instruments that are well-constructed and administered appropriately can yield important insights into operators' experiences while operating weapon systems. In fact, survey instruments are commonly used in the social sciences, particularly psychology, to measure how people think and feel about objects or events (Visser, Krosnick, & Lavrakas, 2000). Typically, participants respond to a series of questions or statements using a rating scale (for example, 1 = *Strongly Disagree*, 5 = *Strongly Agree*) or some other close-ended response option (No/Yes, True/False); though, open-ended, essay-type questions may also be included. Researchers then compute a composite score (e.g., sum, average) from the individual ratings to create a quantitative measure of some underlying construct, such as system trust or usability (Furr & Bacharach, 2014). Survey instruments, like all methods of measurement, however, are subject to measurement error, which can lead researchers to draw erroneous conclusions.

Specifically, survey instruments are susceptible to five primary sources of measurement error: coverage error, sampling error, nonresponse error, instrument-induced error, and participant-induced error (Hansen, Hurwitz, & Madow, 1953; Visser et al., 2000). *Coverage error* is bias that occurs because the pool of potential participants from which the sample is drawn fails to include some portion(s) of the population of interest. *Sampling error* arises due to random differences that exist between the characteristics of the sample and the population from which it was drawn. *Nonresponse error* is bias that results when data are not collected from all participants in the sample. *Instrument-induced error* is bias that is due to some aspect of the survey instrument itself, such as ambiguous or confusing question wording, biased questions, or poorly constructed response options. *Participant-induced error* arises as a result of the participants own behavior, such as misreporting their true feelings or failing to pay close attention to how questions are worded. Given that the population of operators for a specific weapon system is likely to be relatively homogenous (e.g., similar training, embedded in a similar organizational structure and culture), the risk of drawing erroneous conclusions due to coverage or sampling error is greatly reduced. However, bias that arises due to nonresponse, instrument- and participant-induced errors have the potential to affect the conclusions that testers draw about the quality of human-system interactions during operational testing. Implementing a multi-method approach would greatly reduce this risk.

KEY CHARACTERISTICS OF MULTI-METHOD APPROACHES

Multi-method approaches reduce the risk that researchers will draw erroneous conclusions from data by assuring that the observed effect is the result of the construct of interest rather than a bi-product of a particular measurement method (Campbell & Fiske, 1959). This assurance is born out of a process called *triangulation* in which two or more methods are used to measure a construct, each with unique sources of measurement error. For example, a tester might administer a survey instrument to measure system usability during an operational test and later use that data to demonstrate that poor system usability reduces the likelihood of mission success. It could be argued, however, that the observed relationship between system usability and mission success is a bi-product of the method used to measure system usability (i.e., a survey instrument) rather than the construct itself. Demonstrating that the relationship between system usability and mission success holds when using a second method to measure system usability, such as task completion time, would greatly reduce if not eliminate this possibility. Arguably, the convergence of findings from two or more measurement methods, with their respective sources of error, provides a more rigorous and defensible test and should, therefore, increase our confidence in the findings (Bouchard, 1976; Webb, Campbell, Schwartz, & Sechrest, 1966). Additionally, triangulation tends to produce richer data, is more comprehensive in scope making it possible to uncover any contradictions in the data, and forces researchers to think

critically and creatively about the data collection methods that are most useful for addressing their research questions (Jick, 1979; Rossman & Wilson, 1985).

Researchers can triangulate methods in four ways (Denzin, 1978). The first, referred to as *data triangulation*, requires that researchers use multiple sources of data in a study. Often, the only source of data on new or upgraded weapon systems with representative operators comes from a few operational test events. Some of these test events occur early in system development and are limited by the lack of representative operators and the immaturity of the system. This form of triangulation may, therefore, be difficult to achieve in operational testing. The second, *investigator triangulation*, requires that multiple researchers collect, analyze, and interpret data. This form of triangulation already occurs to a limited extent during operational testing. Although the Services' operational test agencies are the primary data collectors during test events, both the operational test agencies and DOT&E independently evaluate the adequacy of the test plan before the test begins, and independently analyze and interpret test data. However, there are few instances during test events in which testers examine the degree to which ratings from multiple, independent observers agree on what happened under a particular set of test conditions – for instance, the degree to which a human-system interaction was successful. Methods such as this are helpful in quantifying the amount of measurement error associated with observational techniques and survey instruments, and produce more robust findings. The third, *theory triangulation*, requires that researchers present multiple interpretations of the data stemming from different theoretical perspectives. In general, it is good practice for operational testers to present multiple possible explanations for findings when data on the mechanisms underlying an observed effect are not clear or were not collected. The fourth and arguably, most rigorous form of triangulation is *methodological triangulation*. Methodological triangulation occurs when researchers use multiple methods to study a research problem. In operational testing, this could mean that testers evaluate the quality of human-system interactions using a combination of techniques including: survey instruments, independent observers, behavioral and physiological metrics, and structured- or semi-structured interviews (among others). As mentioned above, this form of triangulation assures that findings are a product of the construct being measured rather than an artifact of the method chosen to measure the construct.

Additionally, methodological triangulation in which researchers combine both quantitative (e.g., survey instruments, behavioral metrics) and qualitative (e.g., interviews, observational techniques) methods to address a research question are better able to explain the relationships that researchers observe in their data (Collins, Onwuegbuzie, & Sutton, 2006; Cook, 1985; Reichardt & Cook, 1979). This capability is due to the complimentary natures of quantitative and qualitative methods. In particular, quantitative methods tell us that an effect exists and how large that effect is whereas qualitative methods explain why that effect exists. For example, testers might use quantitative methods to determine whether a relationship exists between operators' trust in a threat detection system and the probability that operators will turn off the system. By contrast, testers might use qualitative methods to understand what makes operators distrust the system and how those factors contribute to their decision to turn the system off. Such insights are difficult to glean from quantitative (qualitative) methods alone. Thus, integrating both quantitative and qualitative methods into operational test events will increase the likelihood that testers will be able to explain any relationships that they observe between the quality of human-system interactions and mission-level outcomes, and if structured properly, may be able to identify whether problems would be best addressed through additional operator training, modification of existing tactics, techniques, and procedures, or interface redesign.

APPLICATION OF THE MULTI-METHOD APPROACH IN OPERATIONAL TESTING

Currently, the operational test community relies almost exclusively on survey instruments, a single-method approach, to evaluate the quality of human-system interactions during operational testing. Although some operational test agencies claim to conduct interviews and focus groups, typically they are not conducted according to best practices, are implemented inconsistently during test events, and are not evaluated in a systematic or comprehensive manner. While it is possible for testers to learn a great deal about how operators interact with weapon systems using survey instruments, we argue that a multi-method approach is preferable to a single-method approach for several reasons. First, multi-method approaches assure that any observed effects are a result of specific human-system interaction constructs rather than the method used to measure them, reducing or eliminating the risk that testers will report erroneous effects. Second, multi-method approaches serve as a more rigorous and defensible test of human-system interactions than single-method approaches and should, therefore, give testers greater confidence in the effects they observe. Third, multi-method approaches are more comprehensive and yield richer datasets. Finally, multi-method approaches, particularly those that combine both quantitative and qualitative methods, increase the likelihood that testers will be able to explain any observed effects and identify whether poor human-system interactions can be addressed through training or if the interface needs to be changed. Given these

advantages, the operational test community would benefit from adopting a multi-method approach whenever possible during operational testing.

To illustrate the advantages more clearly, we present a case study in which operational testers implemented a multi-method approach. During the test, operational testers examined how helicopter pilots' workload changed with the aid of a threat detection technology under conditions of high and low battlefield density. Workload was measured using both a validated survey instrument (the NASA-Task Load Index) and a behavioral metric (how quickly operators detected a threat), two quantitative methods with independent sources of measurement error. Unfortunately, qualitative methods were not systematically collected during this test event. However, inclusion of multiple quantitative methods to evaluate workload still provides many advantages above a single-method approach, including the ability to triangulate findings and a dataset that yields greater insights into how operators' experiences impact performance. A more detailed description of the test event and associated findings is provided below.

ATTACK HELICOPTER CASE STUDY

Overview of Test Design and Procedure

The test consisted of 22 operationally realistic attack helicopter missions. The goal of these missions was to detect and destroy threats in the environment with or without the aid of a new threat detection technology. Each mission included groups of 2 helicopters with 2 pilots in each helicopter. The test was a 2(Technology: absent, present) X 2(Threat Density: low, high) D-optimal design, controlling for the time of day that the mission took place (day or night) as prior testing has demonstrated that pilots may find some piloting tasks more difficult at night. The number of missions under each set of test conditions is provided in Table 1.

Table 1. Number of missions conducted under each set of test conditions

	Technology Absent		Technology Present	
	Low Threat Density	High Threat Density	Low Threat Density	High Threat Density
Day	3	2	6	3
Night	1	2	2	3

During each mission, testers captured data on how quickly the team of pilots detected the first threat in the environment by electronically recording how long it took the pilots to detect a potential target and determine that the target was a threat using the helicopter's targeting software. This resulted in a total of 22 observations, one for each mission. Teams of pilots were changed randomly to ensure that observed effects were not attributable to differences in team dynamics. Directly following each mission, the pilots completed a short survey designed to assess their workload during the mission, resulting in a total of 74 observations from 10 pilots. Testers were unable to collect survey data on 14 occasions as a result of participant choice or confusion, and operational factors. Each pilot completed between 3 and 10 missions during the test, with the majority completing between 7 and 8 missions. The number of surveys completed in each condition is provided in Table 2.

Table 2. Number of surveys collected under each set of test conditions

	Technology Absent		Technology Present	
	Low Threat Density	High Threat Density	Low Threat Density	High Threat Density
Day	12	8	20	12
Night	4	4	4	10

Measures

NASA Task Load Index (NASA-TLX; Hart & Staveland, 1988). The NASA-TLX is a 6-item scale that measures workload across 6 dimensions, including: mental demand, physical demand, temporal demand, perceived performance, frustration, and effort. One item is designed to address each of these dimensions. Scale scores range from 0 to 100 and reflect the level of workload that users experienced using a specific system to accomplish a specific task or set of tasks. The NASA-TLX has been used to measure workload in more than 1600 studies over the

past 20 years (Hart, 2006) and is commonly used to measure task workload in operational testing. The NASA-TLX demonstrates high test-retest reliability (r = .83; Hart & Staveland, 1988) and internal consistency (α > .80; Xaio, Wang, Wang, & Lan, 2005).

Threat Detection Task. Task completion time or the efficiency with which operators complete a task is a widely recognized behavioral measure of workload (for reviews see Gopher & Donchin, 1986; Hancock & Meshkati, 1988). Theoretically, operators that are under higher levels of workload (mental or physical) will take longer to complete tasks than those who are under lower levels of workload. In aviation, for example, pilots might take longer to complete a checklist or recognize a change on the instruments during an emergency than during normal operations, suggesting that the pilot is experiencing higher workload during an emergency. Following this logic, operational testers measured how quickly pilots detected a threat in the environment under different operational conditions. Time to detect a threat was captured electronically by the helicopter's targeting software. The time began when the helicopters reached the combat area and ended when the team of pilots "locked on" to the threat.

Results

The two measures of workload, the NASA-TLX and time to detect first threat, were evaluated separately and compared qualitatively because they were collected at the individual and group levels, respectively. Results are presented and discussed for each measure below in turn.

NASA-TLX Results: The 6-items from the NASA-TLX demonstrated high levels of internal consistency (α >.70) and consequently, were averaged to compute a single measure of workload for each pilot. Pilots reported a relatively low level of workload (M = 22.43, SD = 9.05) across test conditions. However, workload scores were higher under some conditions than others. Descriptive statistics for each condition are provided in Table 3. As you can see, pilots reported the lowest levels of workload when the threat detection technology was absent and the threat density was high and the highest levels of workload when the threat detection technology was absent and the threat density was low. Workload scores when the threat detection technology was present fell between these extremes under conditions of both high and low threat density.

Table 3. Descriptive statistics for the NASA-TLX

	Technology Absent		Technology Present	
	Mean	SD	Mean	SD
Low Threat Density	24.88	11.88	23.25	7.92
High Threat Density	17.67	8.00	22.36	7.97

A mixed effects model was used to determine whether pilots' workload scores differed statistically after controlling for the time of day that the mission took place. A mixed effect model was chosen to account for dependency in the data that occurred because the same pilots provided multiple ratings of their workload throughout the test. Although a repeated measures ANOVA is also designed to deal such dependencies, it cannot deal with differences in the number of ratings provided by each pilot.

The fixed effects – presence of the threat detection technology, threat density, and time of day – were regressed on NASA-TLX scores simultaneously, with pilot entered as a random effect. Restricted maximum likelihood estimation ("REML") was used to estimate the model. Together, the fixed and random effects accounted for 47.74 percent of the variance in pilots' workload scores, with the fixed effects accounting for nearly half of that value (marginal $R^2 = 0.21$). The regression coefficients for the fixed effects are presented in Table 4.

Table 4. NASA-TLX model results

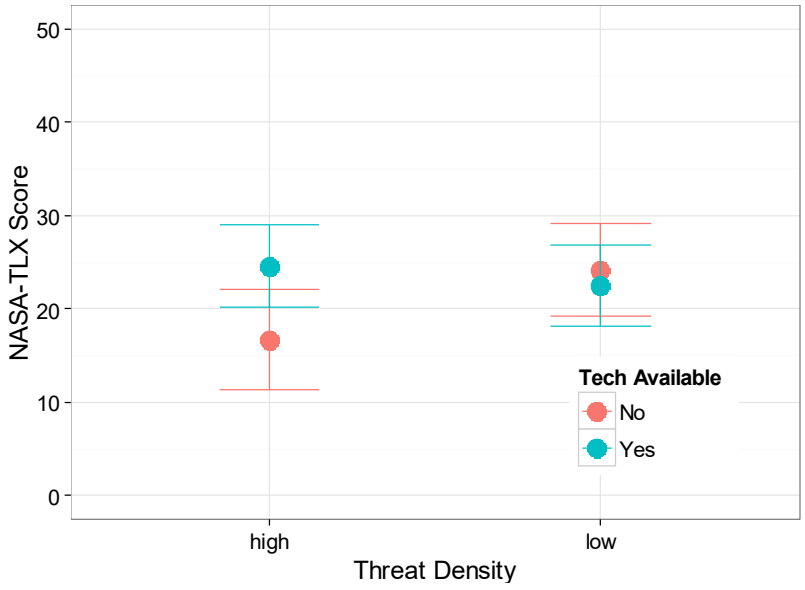
	Coefficient	SE	t-value
Time of Day	-7.56***	1.84	-4.10
Threat Density	7.50**	2.69	2.79
Technology Presence	7.89**	2.74	2.88
Threat Density X Technology Presence	-9.56**	3.43	-2.79

***p<.001, **p<.01

Pilots rated their workload significantly higher during the day than during the night and under conditions of low threat density than high threat density. In general, it is harder to detect threats when there are fewer threats in the environment. Unexpectedly, however, pilots also rated their workload higher when the new threat technology was available in the cockpit. The reason for this finding becomes clearer when we consider the nature of the interaction between threat density and technology presence. This interaction is presented in Figure 1.

In particular, pilots reported similar levels of workload under conditions of high and low threat density when the threat detection technology was present. When the threat detection technology was absent, however, pilots reported higher levels of workload when threat density was low than when it was high. These findings suggest that the threat detection technology is helping pilots manage their workload when threat density is low, but is actually contributing to the difficulty of detecting threats when threat density is high. The reason for this is unclear given that qualitative data was not systematically collected during the test event. However, the qualitative data that is available suggests that elements of the interface may be driving this effect. In particular, when the system detects a potential threat, an icon pops up that has to be manually investigated by the pilot. To do so, the pilot must hover over the icon with the cursor and select it to read information about the threat. When threat density was high, icons cluttered the screen, making it more difficult for pilots to perform the detection task using the technology than simply looking out the window.

Figure 1. Threat Density X Technology Presence Interaction



Note: error bars represent 95 percent confidence intervals

One way to mitigate this unintended effect of the technology would be to alter tactics, techniques, and procedures so that pilots turn off or ignore the technology when the threat density is high. Before making this recommendation, however, it is important to verify that this effect actually alters their behavior, hindering performance. To examine this possibility, we must look at whether pilots' ability to detect threats demonstrated a similar pattern of results.

Threat Detection Task Results: Data collected on the amount of time it took pilots to detect a threat was normalized (converted to z-scores) to protect sensitive information. This process simply places the data in standard deviation units and does not affect the magnitude of any observed effects. To normalize the data, the mean detection time was subtracted from each individual observation and divided by the standard deviation of all observations. This places the data on a scale where the mean detection time is 0 and the standard deviation of the distribution is 1. Negative values represent detection times that were quicker than the mean whereas positive values represent detection times that were slower than the mean. Descriptive statistics are provided in Table 5. Consistent with the NASA-TLX data presented above, pilots were slowest a detecting a threat under low threat density when the threat detection technology was absent and were quickest at detecting a threat under high threat density when the threat detection technology was absent. Detection times when the technology was present fell between these two extremes.

Table 5. Descriptive statistics for the threat detection task

	Technology Absent		Technology Present	
	Mean	SD	Mean	SD
Low Threat Density	1.51	1.28	-0.35	0.70
High Threat Density	-0.64	0.38	-0.11	0.31

A linear regression model was used to determine whether threat detection time differed statistically by condition after controlling for the time of day that the mission took place. A mixed effects model was originally considered to account for the fact that the same pilots completed missions throughout the test; however, the random effect of pilot did not significantly improve model fit ($p > .90$) and was therefore, discarded in favor of a simpler, fixed effects only model.

The three predictors – presence of the threat detection technology, threat density, and time of day – were regressed on threat detection times simultaneously. Together, these predictors accounted for 47.13 percent of the variance in threat detection time, a similar number to that reported above for NASA-TLX scores. That said, the pattern of results for each predictor differed somewhat from those reported above. The threat density by technology presence interaction was the only significant predictor of time to detect a threat. None of the main effects remained significant predictors after accounting for this interaction. The regression coefficients for the threat detection model are presented in Table 6.

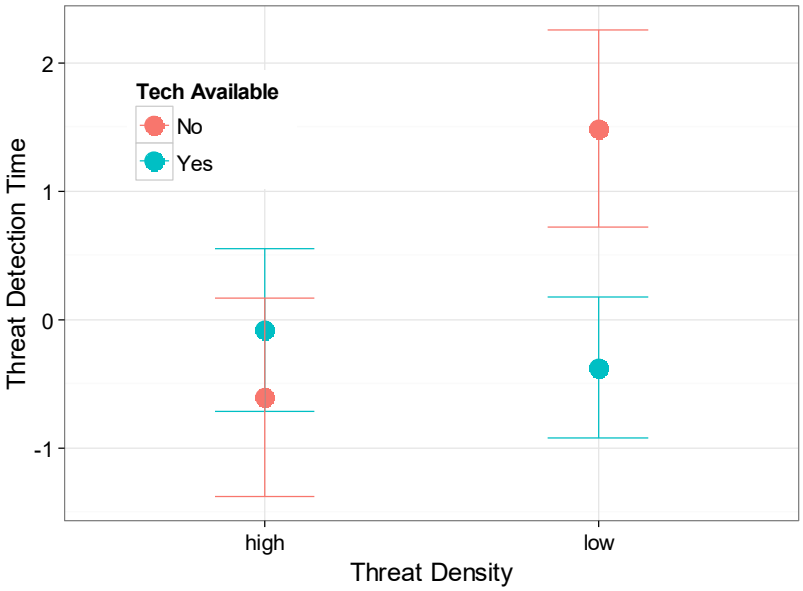
Table 6. Threat Detection Task model results

	Coefficient	SE	t-value
Time of Day	-0.20	0.33	-0.61
Threat Density	-0.30	0.40	-0.74
Technology Presence	-0.53	0.47	-1.13
Threat Density X Technology Presence	2.39**	0.65	3.70

***$p<.001$, **$p<.01$

Mirroring the NASA-TLX findings, time to detect a threat was similar under conditions of high and low threat density when the threat detection technology was present. When the threat detection technology was absent, however, pilots took longer to detect threats when threat density was low than when it was high. These findings are presented in Figure 2. Again, these findings suggest that the threat detection technology is helping pilots manage their workload when threat density is low, but is actually contributing to the difficulty of detecting threats when threat density is high. The fact that we were able to replicate this same pattern of results using both a survey instrument and a behavioral metric gives us confidence that these results reflect reality rather than chance or measurement error – the threat detection technology improves workload under some conditions, but not others. Furthermore, it serves as a more rigorous test of these effects and is, therefore, more defensible than reporting results from either of these measures of workload alone and provides multiple pieces of evidence to support the idea that pilots may benefit from altering their tactics, techniques, and procedures when using the threat detection technology under conditions of high threat density.

Figure 2. Threat Density X Technology Presence Interaction



Note: error bars represent 95 percent confidence intervals

DISCUSSION & RECOMMENDATIONS

The purpose of this paper was to identify the shortcomings of a single-method approach to evaluating human-system interactions during operational testing and offer an alternative, multi-method approach that is more defensible, yields richer insights into how operators interact with weapon systems, and provides a practical implications for identifying when the quality of human-system interactions warrants correction through either operator training or redesign. Single-method approaches place testers at risk of drawing erroneous conclusions from their data, particularly when using more liberal standards for determining when factors significantly predict an outcome of interest as is commonly the case in operational testing. For example, testers commonly use a standard of $\alpha = .20$ when deciding whether a factor, such as threat density, significantly predicts some mission-level outcome. This means that they are at risk of drawing erroneous conclusions 20 percent of the time due to random error alone. A multi-method approach would markedly reduce this risk.

In fact, multi-method approaches would benefit the operational test community in several ways. As mentioned above, multi-method approaches assure that any observed effects are a result of specific human-system interaction constructs rather than the method used to measure them, reducing or eliminating the risk that testers will report erroneous effects. They also produce datasets that are more comprehensive and richer than those obtained when implementing single-method approaches and serve as a more rigorous and defensible test of human-system interactions, giving testers greater confidence in the effects they observe. These benefits were illustrated in the attack helicopter case study when we were able to replicate the same pattern of results using both a survey instrument and a behavioral metric to measure pilot workload. Systematically integrating qualitative methods into the test design would have made it easier to explain why the threat detection technology reduced pilot workload under some conditions, but not others and identify whether modifying tactics, techniques, and procedures might further improve performance or if adjustments to the interface are needed. Given these benefits, we recommend that the operational test community adopt a multi-method approach whenever feasible during test events, and make an effort to systematically integrate both quantitative and qualitative methods into their evaluation of human-system interactions.

REFERENCES

Bangor, A., Kortum, P., & Miller, J. (2009). Determining what individual SUS scores mean: Adding an adjective rating scale. *Journal of Usability Studies*, 114-123.

BEA (2012). Final report on the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro–Paris. Paris: BEA

Borsci, S., Federici, S., Bacci, S., Gnaldi, M., & Bartolucci, F. (2015). Assessing User Satisfaction in the Era of User Experience: Comparison of the SUS, UMUX, and UMUX-LITE as a Function of Product Experience. *International Journal of Human-Computer Interaction*, 484-495.

Bouchard, T. (1976). Unobtrusive measures: An inventory of uses. *Sociological Methods and Research*, 267-300.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 81-105.

Charlton, S. (1991). *Aeromedical human factors OT&E handbook.* Kirkland, AFB NM: Air Force Operational Test and Evaluation Center.

Collins, K., Onwuegbuzie, A., & Sutton, I. (2006). A model incorporating the rationale and purpose for conducting mixed methods research in special education and beyond. *Learning Disabilities: A Contemporary Journal*, 67-100.

Cook, T. (1985). Postpositivist critical multiplism. In Shotland, & Mark, *Social science and social policy* (pp. 21-62). Beverly Hills, CA: Sage.

Defense Science Board. (2012). *The role of autonomy in DoD systems.* Department of Defense.

Defense Science Board. (2016). *Summer study on autonomy.* Department of Defense.

Denzin, N. (1978). *The research act: A theoretical introduction to sociological methods.* New York, NY: Praeger .

Dillman, D., Smyth, J., & Christian, L. (2009). *Internet, mail, and mixed-mode surveys.* Hoboken, New Jersey: John Wiley & Sons, Inc.

Donchin, E., & Gopher, D. (1986). Workload - An examination of the concept. In Boff, Kaufman, & Thomas, *Handbook of Perception and Human Performance* (pp. 1-49). New York: Wiley.

Furr, M., & Bacharach, V. (2014). *Psychometrics: An Introduction.* Sage Publications.

Gawron, V., Schiflett, S., & Miller, J. (1989). Cognitive demands of automation in aviation. In Jensen, *Aviation Psychology* (pp. 240-287). Brookfield, VT: Gower.

Hancock, P., & Meshkati, N. (1988). *Human Mental Workload.* Amsterdam: North Holland Press.

Hansen, M., Hurwitz, W., & Madow, W. (1953). *Sample Survey Methods and Theory.* John Wiley & Sons, Inc.

Hart, S. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting* (pp. 904-908). Santa Monica: HFES.

Hart, S., & Staveland, L. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In Hancock, & Meshkati, *Human Mental Workload* (pp. 239-250). Amsterdam: North Holland Press.

Hoff, K., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, 407-434.

Jick, T. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly*, 602-611.

Johnson, R., Onwuegbuzie, A., & Turner, L. (2007). Toward a definition of mixed methods research. *Journal of mixed methods research*, 112-133.

Levs, J. (2012, January 15). *What caused the cruise ship disaster?* Retrieved from CNN: http://www.cnn.com/2012/01/15/world/europe/italy-cruise-questions/index.html

Montgomery, D. (2012). *Design and analysis of experiments.* John Wiley & Sons, Inc.

Reichardt, D., & Cook, T. (1979). Beyond qualitative versus quantitative methods. In Cook, & Reichardt, *Qualitative and quantitative methods in evaluation research* (pp. 7-32). Beverly Hills, CA: Sage.

Rossman, G., & Wilson, B. (1985). Numbers and words: Combining qualitative and quantitative methods in a single large scale evaluation study. *Evaluation Review*, 627-643.

Snavely, J., Wojton, H., Bieber, C., & Freeman, L. (2017). *Status of the Operational Test Agency Workforce FY06-FY16.* Alexandria, VA: Institute for Defense Analyses.

Somekh, B., & Lewin, C. (2011). *Theory and methods in social research.* London & Thousand Oaks, CA: Sage Publications.

Tillman, B., Fitts, D., Woodson, W., Rose-Sundholm, R., & Tillman, P. (2016). *Human factors and ergonomics design handbook.* McGraw Hill Education.

Visser, P., Krosnick, J., & Lavrakas, P. (2000). Survey Research. In Reis, & Judd, *Handbook of research methods in social and personality psychology* (pp. 223-252). New York, NY: Cambridge University Press.

Webb, E., Campbell, D., Schwartz, R., & Sechrest, L. (1966). *Unobtrusive measures.* Chicago, IL: Rand McNally.

IDA

The Wisdom of Crowds: Improving Operational Testing Through Multiple Methodologies

The Institute for Defense Analyses

Approved for public release; distribution is unlimited.

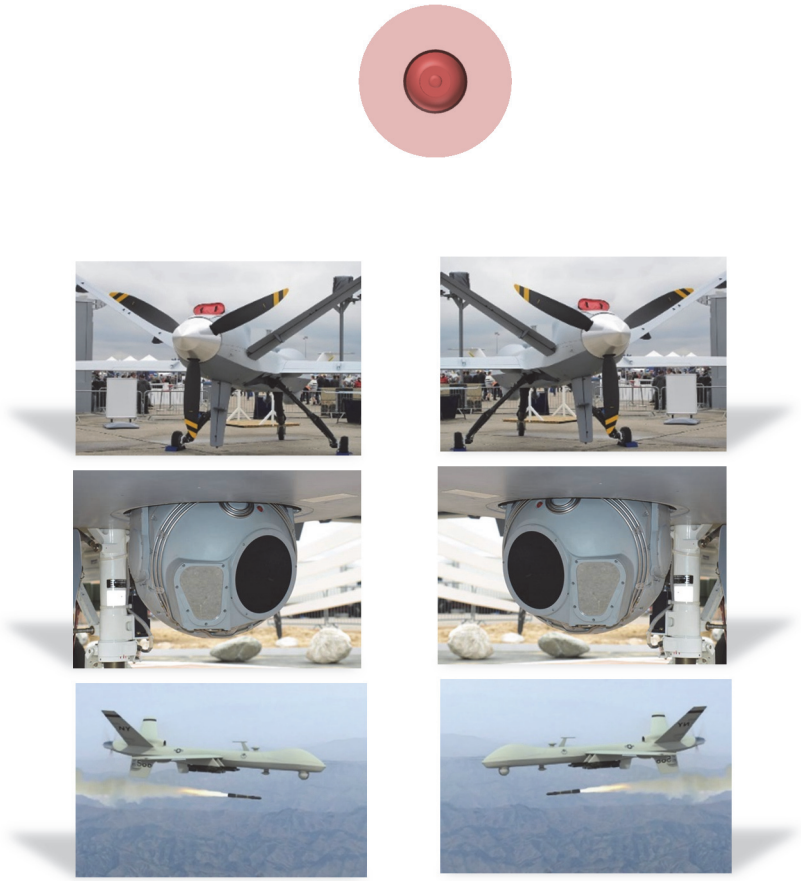
IDA | Operator-in-the-Loop Testing

- How well will a system perform?

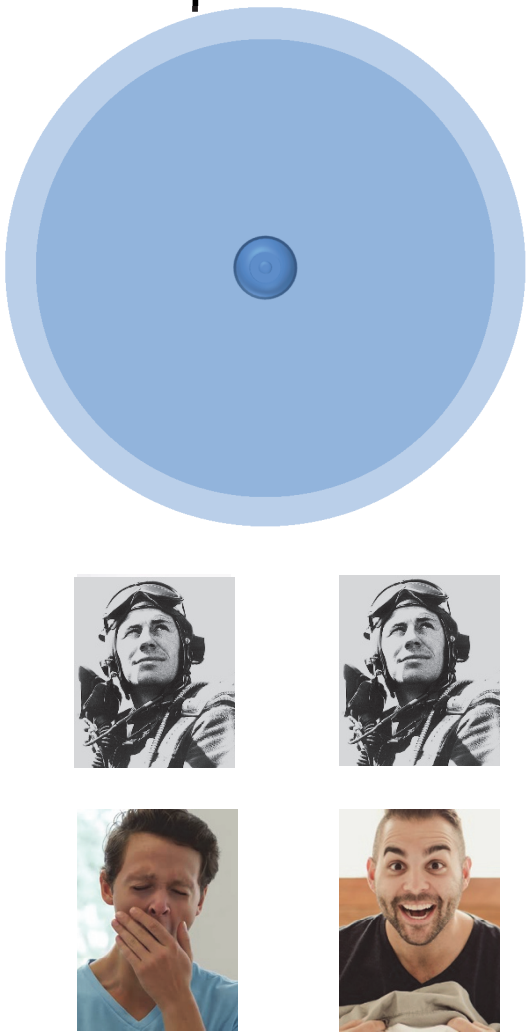
Approved for public release; distribution is unlimited.

IDA | Confidence in Operational Capabilities

Mechanical



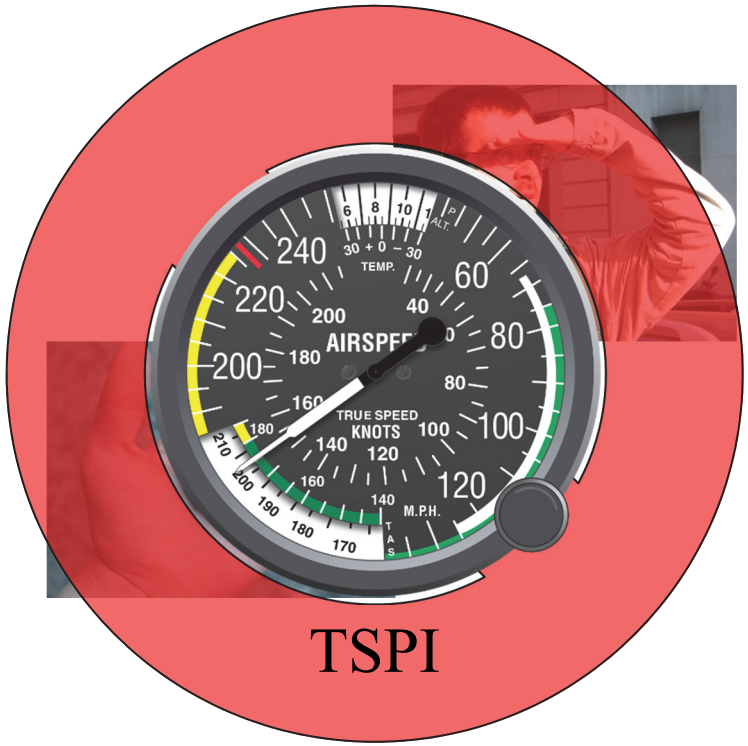
Operator



IDA | Error in Measurement

Mechanical

Operator

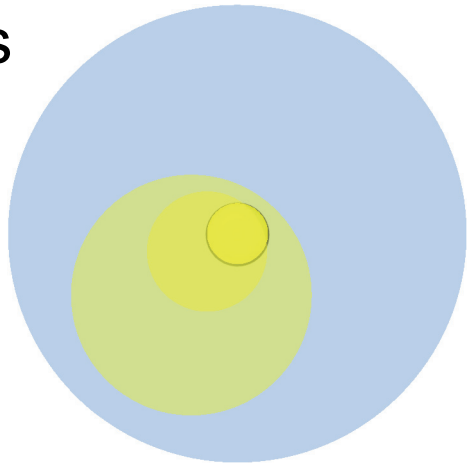


IDA | Reducing Operator Measurement Error

Operator

Problem: Trusting Automatic Warnings

11. I am rarely surprised by how the system responds. 1 2 3 4 5 6 7



stem?"

Not at all — Extremely

1 2 3 4 5 6 7

IDA | Reducing Operator Measurement Error

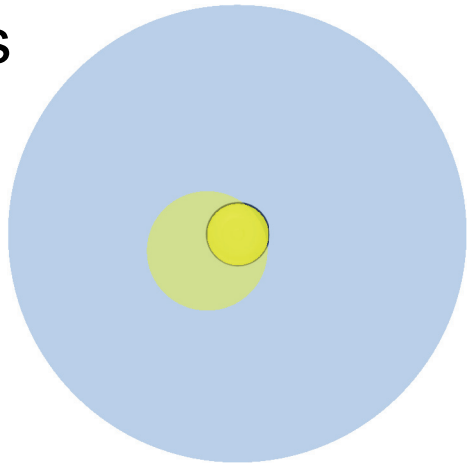
Problem: Trusting Automatic Warnings

SP3 Scale

Directions: Read each statement carefully and indicate the extent to which you agree or disagree using the scale provided.

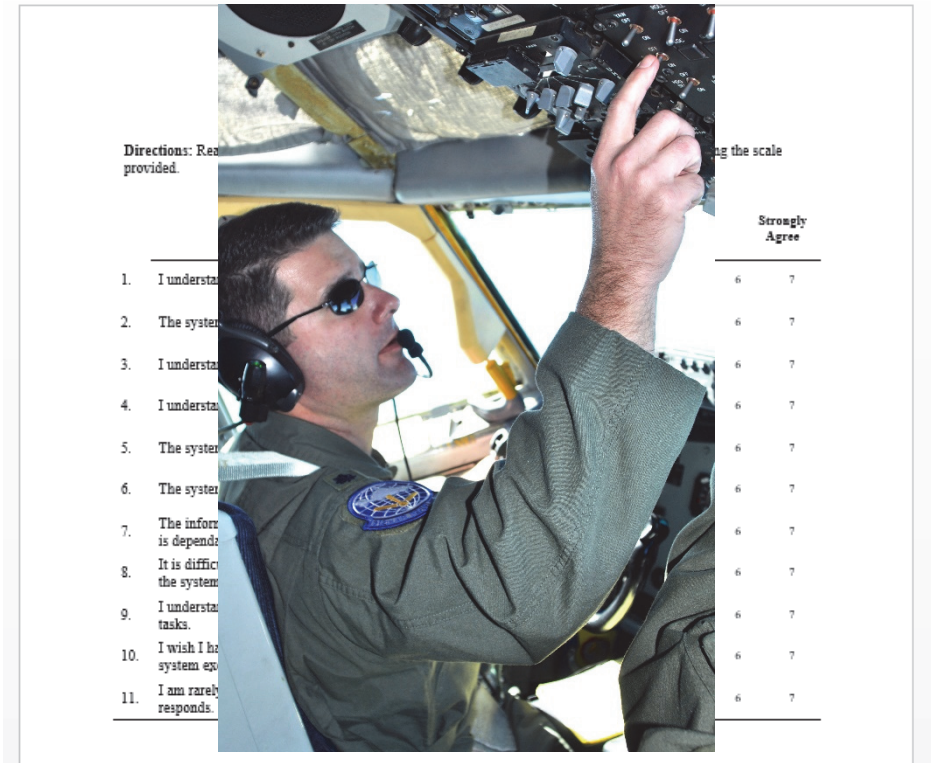
		Strongly Dis						Strongly
	I u tand the system should d	1	2	3	4	5	6	
	Th tem h eve my goal	1	2	3	4	5	6	
	tand nitations of the sys	1	2	3	4	5	6	
	tand of the sy	1	2	3	4	5	6	
5.	The system performs consistently.	1	2	3	4	5	6	7
6.	The system performs as expected.	1	2	3	4	5	6	7
7.	The information I receive from the system is dependable.	1	2	3	4	5	6	7
8.	It is difficult to know when I should trust the system.	1	2	3	4	5	6	7
9.	I understand how the system executes tasks.	1	2	3	4	5	6	7
10.	I wish I had more control over how the system executes tasks.	1	2	3	4	5	6	7
11.	I am rarely surprised by how the system responds.	1	2	3	4	5	6	7

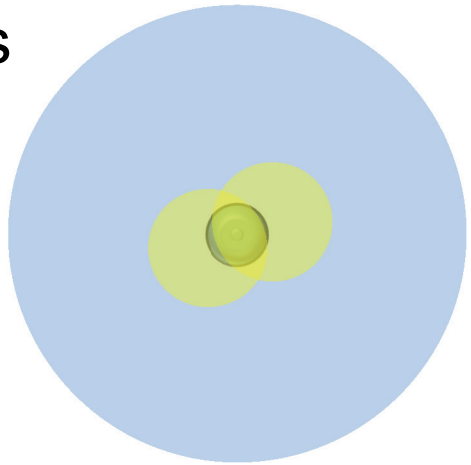
NETFLIX



IDA | Reducing Operator Measurement Error

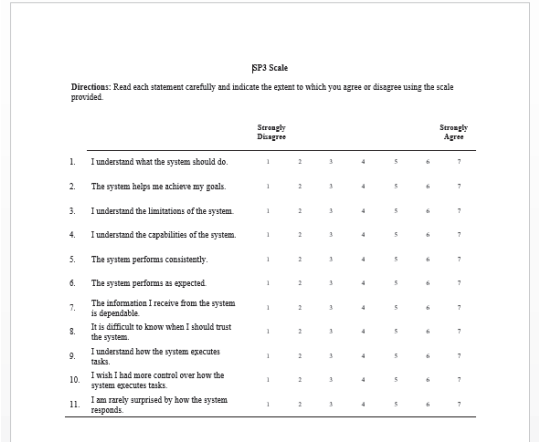
Problem: Trusting Automatic Warnings





IDA | Reducing Operator Measurement Error

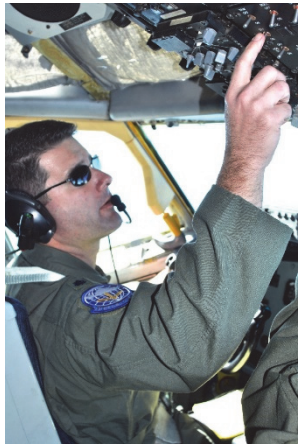
Problem: Trusting Automatic Warnings



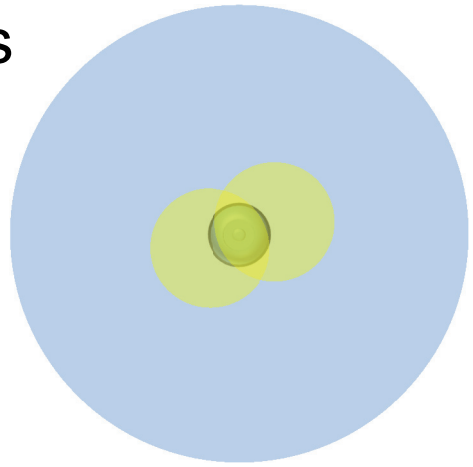
JSP3 Scale

Directions: Read each statement carefully and indicate the extent to which you agree or disagree using the scale provided.

		Strongly Disagree						Strongly Agree
1.	I understand what the system should do.	1	2	3	4	5	6	7
2.	The system helps me achieve my goals.	1	2	3	4	5	6	7
3.	I understand the limitations of the system.	1	2	3	4	5	6	7
4.	I understand the capabilities of the system.	1	2	3	4	5	6	7
5.	The system performs consistently.	1	2	3	4	5	6	7
6.	The system performs as expected.	1	2	3	4	5	6	7
7.	The information I receive from the system is dependable.	1	2	3	4	5	6	7
8.	It is difficult to know when I should trust the system.	1	2	3	4	5	6	7
9.	I understand how the system executes tasks.	1	2	3	4	5	6	7
10.	I wish I had more control over how the system executes tasks.	1	2	3	4	5	6	7
11.	I am rarely surprised by how the system responds.	1	2	3	4	5	6	7



Operator



Triangulation

IDA | Triangulation Case Study

- New Data Link with Target Data for an attack helicopter
- Helicopter Mission: Detect and Destroy
- Does new tech help the mission?
  - High vs. Low Threat Density

Approved for public release; distribution is unlimited.

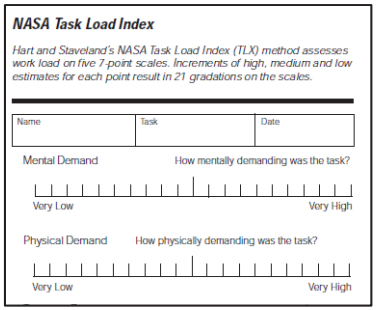
IDA | Workload Reduction?

- Tech should make work easier
  - Externalized cognition
- Lighter workload should help mission
- How can we test this?

Approved for public release; distribution is unlimited.

IDA | Measuring Workload

- NASA Task Load Index (TLX)
  - Self-report
  - Robust, commonly used metric



- Reaction Time
  - Behavioral Triangulation
  - Objective, operationally relevant

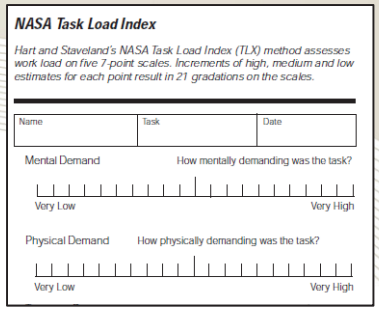
Approved for public release; distribution is unlimited.

IDA | Test Design

	No Tech	New Tech
Low Threat Density		
High Threat Density		

Approved for public release; distribution is unlimited.

IDA | NASA-TLX Results



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name	Task	Date

Mental Demand — How mentally demanding was the task?

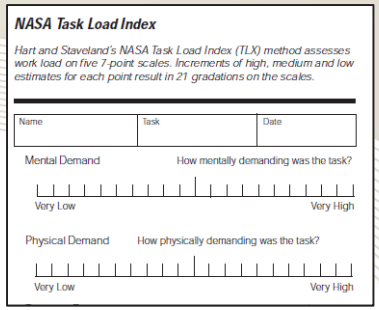
Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

	No Tech	New Tech
Low Threat Density		
High Threat Density		

IDA | NASA-TLX Results


NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name | Task | Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

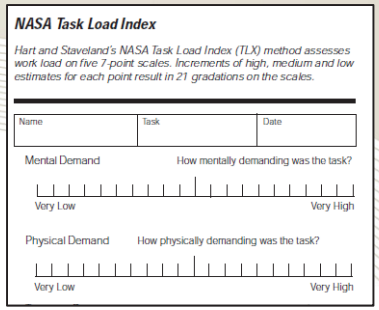
Physical Demand — How physically demanding was the task?

Very Low — Very High

	No Tech	New Tech
Low Threat Density	24.88 (11.88)	
High Threat Density	17.67 (8.00)	

Standard deviations in parentheses

IDA | NASA-TLX Results


NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name | Task | Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

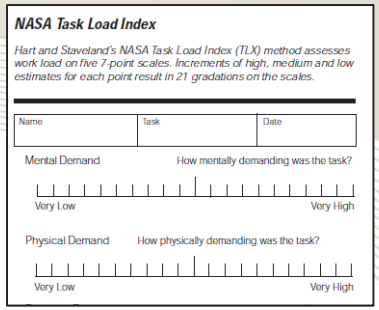
Physical Demand — How physically demanding was the task?

Very Low — Very High

	No Tech	New Tech
Low Threat Density		23.25 (7.92)
High Threat Density		22.36 (7.97)

Standard deviations in parentheses

IDA | NASA-TLX Results


NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name | Task | Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

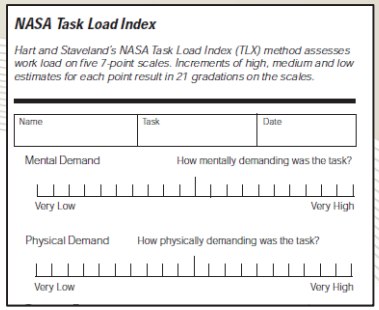
Very Low — Very High

	No Tech	New Tech
Low Threat Density	24.88 (11.88)	23.25 (7.92)
High Threat Density	17.67 (8.00)	22.36 (7.97)

What conclusions would you draw from these results?

Standard deviations in parentheses

IDA | NASA-TLX


NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name | Task | Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

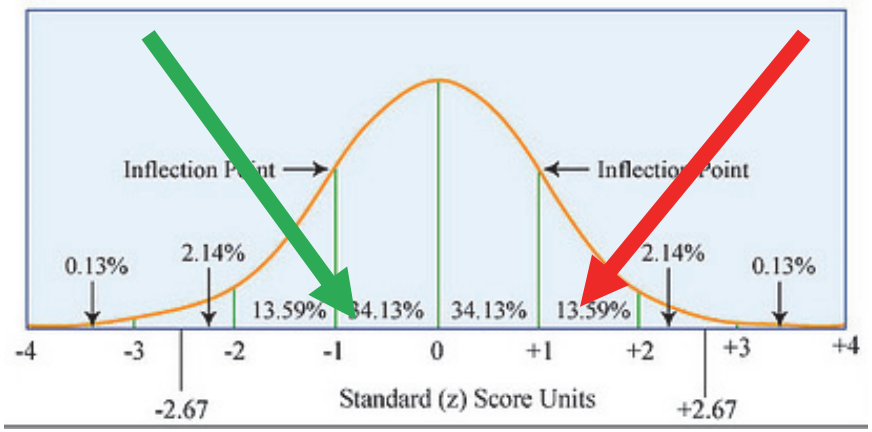
Very Low — Very High

- TLX alone implies that new tech:
  - Hurts in high threat density
  - No benefit in low threat density

Approved for public release; distribution is unlimited.

IDA | Reaction Time

	No Tech	New Tech
Low Threat Density		
High Threat Density		



IDA | Reaction Time

	No Tech	New Tech
Low Threat Density	1.51 (1.38)	
High Threat Density	-0.64 (0.38)	

Note: Normalized values, SD in parentheses

Approved for public release; distribution is unlimited.

IDA | Reaction Time

	No Tech	New Tech
Low Threat Density		-0.35 (0.70)
High Threat Density		-0.11 (0.31)

Note: Normalized values

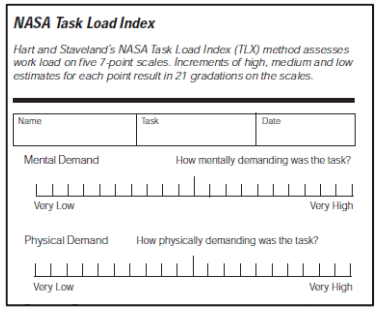
IDA | Reaction Time

	No Tech	New Tech
Low Threat Density	1.51 (1.38)	-0.35 (0.70)
High Threat Density	-0.64 (0.38)	-0.11 (0.31)

Note: Normalized values

IDA | Triangulation

	No Tech	New Tech
Low Threat Density	1.51 (1.38)	-0.35 (0.70)
High Threat Density	-0.64 (0.38)	-0.11 (0.31)



NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Name | Task | Date

Mental Demand — How mentally demanding was the task?

Very Low — Very High

Physical Demand — How physically demanding was the task?

Very Low — Very High

	No Tech	New Tech
Low Threat Density	24.88 (11.88)	23.25 (7.92)
High Threat Density	17.67 (8.00)	22.36 (7.97)

Approved for public release; distribution is unlimited.

IDA | Triangulated Results

- New technology aids mission success
- BUT under certain conditions
  - Creates higher workload and slower reaction in target rich environment
  - But ***more*** effective with fewer targets
    - Not due to workload
- Triangulation added interpretation

Approved for public release; distribution is unlimited.

IDA | Further Triangulation

Approved for public release; distribution is unlimited.

IDA | Conclusions

- More effort to reduce measurement error with operators
- Multiple methodologies reduce error
  - Noise cancels, signal reinforces
- Qualitative + Quantitative is powerful
  - Can move towards causality

Approved for public release; distribution is unlimited.

IDA | Thank you

Approved for public release; distribution is unlimited.