# A New Approach to Scoring Dynamic Decision Making Performance on High Fidelity Simulators: Reliability and Validity Issues

Gunnar E. Schrah, Oleksandr S. Chernyshenko, Michael R. Baumann, & Janet A. Sniezek
Department of Psychology, University of Illinois at Urbana-Champaign

Vadim Bulitko, Scott Borton, & David C. Wilkins
Beckman Institute, University of Illinois at Urbana-Champaign

Poster Presented at the 15th Annual Conference for the Society for Industrial and Organizational Psychology, New Orleans, Louisiana, April, 2000

Please direct any questions or comments to Gunnar Schrah at the University of Illinois, Psychology Department, 603 East Daniel Street, Champaign, IL, 61820, or at gschrah@s.psych.uiuc.edu

Abstract

Advances in technology allow us to simulate tasks with increasing fidelity, providing opportunities for sophisticated training and performance measurement systems. The realism incorporated in such systems also increases the complexity of measuring individual performance. This research describes the development of an artificial intelligence based performance measure utilized in the DC-TRAIN Damage Control Simulator, and some preliminary validation evidence.

.

## Introduction

Presently high-fidelity simulators are primarily utilized for training and assessment of individuals and teams for occupations in which real-world performance is too costly and/or too dangerous (e.g., naval air defense: Driskell & Johnston, 1998; space flight: Flexman & Stark, 1987; tactical aviation: Salas, Bowers & Rhodenizer, 1998). With continued advances in electronic technologies we will certainly see an increase in the use of this approach applied to training, performance measurement and prediction. Additionally the use of "low fidelity" simulators has increased demonstrating that a less realistic simulation can also prove useful (Hanson, Borman, Mogilka, Manning, & Hedge, 1999). The degree of fidelity is a certainly continuum and not a dichotomy. The inclusion of a simulator into the high-fidelity category is usually restricted to systems that simulate complex tasks dynamically, i.e. the simulation changes with operator input, and incorporate excessive realism, i.e. operating environment and computer interface mimic real-world conditions. Nonetheless there are considerable similarities in these approaches that will be highlighted throughout this paper.

Of course, the use of simulations, e.g. in-vitro work samples, as measures and/or predictors of job performance is not new (Campbell, 1990). The advantages of "high fidelity" simulators however are in their ability to accurately simulate a diversity of problem solving and decision-making situations, typical of more complex occupations, under precisely controlled conditions. The electronic nature of these simulators also allows for the collection of immense amounts of data that would be impossible in work samples.

As with any assessment approach, there are drawbacks to the use of "high-fidelity" simulators as measures of performance. The dynamic and realistic nature with which these systems simulate tasks actually complicates the measurement of performance. As opposed to other measurement instruments, the content, i.e. stimulus-response sets, is not known apriori. Indeed each new simulation is unique consisting of a nearly infinite number of possible stimulus-response sets. Fortunately developments in Artificial Intelligence are allowing us to tackle this extremely difficult problem. With these issues in mind, this paper discusses the development of an Artificial Intelligence based performance measure utilized in the DC-TRAIN Naval Damage Control Simulator (Wilkins & Sniezek, 1997), and a methodology by which to validate it. In doing so we attempt to explicate some general guidelines for the use of this novel approach to performance measurement.

The DC-TRAIN 2.0 Simulator (DC-Train) is a multimedia training system currently in use at the US Navy's Surface Warfare Officer School in Newport, RI. It was designed to train and measure the performance of Damage Control Assistants (DCA) in the area of ship crisis management. During the course of a scenario the DCA makes decisions directing the repair and containment of damage during combat situations. DC-Train simulates damage control situations from physical principles. Both the spread of damage and the effectiveness of damage control actions are simulated online. Because of this, DC-Train can create an unlimited number of scenarios. In fact, every simulation run on DC-Train is a unique scenario. The focus of this research is on two systems of DC-Train, Minerva 5.0 automated reasoning system (Bulitko, 1997) and DCA Yardstick, that serve as the performance measurement tool, providing automated, quantitative scoring

data. Essentially, Yardstick compares the actions the DCA selects to the "best" actions, as determined by Minerva, at each of a number of time points during the scenario.

The use of an expert-based system, Minerva, as the criterion of expert performance is a novel approach to the measurement of decision making. Minerva solves the scenario in parallel with the student, recording the "best" action at several points, however only the user's actions affect the state of the ship (See Figure 1). Because the scoring key is generated as the user runs the scenario, the user's decisions are evaluated on the basis of the conditions under which they are made (See Figure 2). This is a significant advance compared to past efforts in which the key is developed a priori. A scoring key developed a priori would be based on conditions that become more and more divergent from the actual conditions the user faces as the scenario progresses. Another approach is to ignore the decision making process altogether and focus on final outcomes (Bolger & Wright, 1992; Keren, 1992). This is particularly problematic in complex domains such as damage control in which the relation between decisions and outcomes is substantially effected by extraneous factors outside of the DCA's control.

While Minerva provides the expert criterion against which DCA performance is compared, a separate scoring program (DCA Yardstick) is responsible for actually assigning the scores to the DCA's chosen actions. Among Yardstick's features, is its ability to give partial credit for decisions that are not "ideal", as determined by Minerva, but within acceptable limits. This is crucial in decision-making tasks in which the occurrence of fully correct or incorrect decisions is rare (Bolger & Wright, 1992; Keren, 1992). In this way, DC TRAIN resembles a situational judgment test, whereby responses to job-related situations are scored along a continuum of effectiveness (Hanson & Ramos,

1996; Weekly & Jones, 1999). Job experts typically determine the level of effectiveness ascribed to a given response (Hanson, et. al. 1999), however in the case of DC-Train the nearly infinite number of possible stimulus-responses sequences renders this approach infeasible. Instead we must rely upon algorithms to determine levels of effectiveness.

Ultimately the level of effectiveness ascribed to a given action is intended to reflect the difference between the state of the ship that would result from the partially correct action versus the action suggested by the key. The average level of effectiveness of responses is then utilized as the total score as opposed to number correct scoring. The situational judgment approach has been shown to increase reliability of scores as compared to more traditional methods thereby increasing the information contained in such scores (Hanson, et. al., 1999).

In contrast to most situational judgment tests, simulations, such as DC Train, are fully dynamic thereby requiring a time component to the scoring system. Thus, the timing of DCA actions is also factored into effectiveness calculations via time windows and/or decreasing scoring coefficients. (See method section for a more detailed discussion of the scoring procedure).

While the logic employed in the development of the scoring system is sound, there are multiple threats to its validity. It is useful to distinguish between two components of the performance measure, the content, i.e. stimulus-response sets, and the method of assigning scores to those responses. As such, any reference to the validity of the performance measure necessarily involves both the content provided by the simulation, DC-Train itself, and the scoring system, Minerva and Yardstick.

With regard to content, difficulties arise when we attempt to partition the observed stream of behavior into measurable units of behavior (Vreuls & Obermayer, 1985). Ultimately, the measured units of behavior should reflect the underlying performance construct as determined in the task analysis process (Motowidlo, Borman, & Schmit 1997). Further, Guion (1978) posits that at least some degree of content relevance can be assumed from a properly conducted task analysis.

In examinations of naval doctrine and interactions with subject matter experts, it became apparent that ship-crisis management is essentially a time-critical decision-making task involving the allocation of resources to combat ship damage. As such, the content of the simulations primarily consists of damage reports presented to the user, stimuli, and choices amongst possible actions, responses. These actions, in light of the damage reports, serve as the unit of measurement.

When discussing the content relevance of such simulations it is also necessary to consider the level of fidelity, i.e. realism, with which the content is presented. It is reasonable to assume that more realistic simulations will provide better measures of performance (Hanson, et. al., 1999), but keeping in mind cost constraints, the question usually becomes what degree of fidelity is sufficient for the intended purposes. The level of fidelity has been viewed as made up of two components: realism of the simulated environment and the fidelity of the responses or behaviors collected. When comparing high-fidelity and low-fidelity simulations it appears that the main distinction is in response fidelity. In particular the dynamic nature with which high-fidelity simulations operate, mimics the user-environment feedback loop that is missing in most low-fidelity simulations.

As with the determination of our content, the level of fidelity incorporated in DC-Train ultimately reflected the functioning and features deemed necessary to invoke decision processes similar to those required in live damage control. In particular the highly stressful environment implicated in damage control was replicated by bombarding the user with multiple information reports, in both audio and visual format. In regard to responses, the selection of actions was accomplished via mouse and keyboard operations. Other behaviors that would be evident in a live ship crisis, e.g. the use of communications equipment, were determined to be peripheral to the damage control task and were not included.

Evidence of face validity has been demonstrated in previous studies of DC-Train in which students reported high levels of effort, anxiety, time pressure, and mental demand associated with simulator performance (Baumann, Sniezek, Donovan, & Wilkins, 1996). Even after four trials on DC-Train, students' anxiety levels did not decrease by a significant amount. Additionally, DCA students rated DC-Train as over a 6.5 on a 7-point scale (anchored at 1 with completely useless and 7 with extremely useful). Taken together, this suggests that DC-Train is psychologically realistic and useful. More precisely, it creates and maintains a psychologically stressful environment that is necessary to prepare DCAs for what they would experience in an actual crisis.

While content relevance is necessary conditions for the establishment of validity, it is certainly not sufficient. The method for assigning scores must also be valid. This requires the scoring system to accurately represent performance differences amongst individual, within individuals, and across groups. Because the scoring system depends on the automated reasoning system, Minerva, and Yardstick, a lack of validity could be

the result of one or both of these systems. With the design of DC-Train in mind, we turn next to the assessment of the validity of the performance measure.

As a first step in assessing the validity of the scoring system, we utilized what we refer to as "simulated user behavior" to ensure stable and proper functioning of both Minerva and Yardstick. Simulation of user behavior was accomplished by having an expert in damage control and the functioning of the simulator; mimic excellent, average, and poor performance defined along strict performance dimensions. In other words, the expert simulated different levels of performance by intentionally manipulating one aspect of his performance, e.g. precision of fire boundaries, while keeping constant all other input behaviors. This allowed us to investigate whether manipulations of critical dimensions of performance would have predictable effects on the total score. While data garnered from such a procedure would be meaningless for more traditional measurement instruments where a manipulation of responses would have clear cut implications for a total score, the complexity of our scoring procedure required such an approach. Without such a demonstration, data collection efforts would have been unwarranted.

After insuring the proper functioning of the scoring system, we proceeded towards collecting actual data. With regard to reliability, we were restricted to utilizing the reliability coefficient, testing the correlation between parallel scenarios, because of the indeterminate number of actions made internal consistency measures impractical. It is important to note that while the exact stimulus-response sequences are unknown prior to a scenario, there are certain set damage events that are determined apriori. For example, a given scenario might include a mine hit five minutes into the scenario and an electrical fire ten minutes into the scenario. Of course the extent to which this initial damage

escalates is determined by the users actions, however these predetermined damage events allow us to develop scenarios of roughly similar content and difficulty. Additionally, parallelism between scenarios was established through expert ratings. Instructors of the DCA training course judged all of our scenarios to be of similar content and equivalent difficulty.

During the testing phase we were careful to consider issues related to method-specific variance resulting from the use of the simulator. Differences in scores attributable to computer and/or interface experience must be reduced. In order to reduce this effect, participants in this study went through an instructional session on how to operate the simulator prior to criterion performance.

To aid in the construct interpretation of our performance measure, we examined relationships between our performance measure and participants' domain knowledge, DCA course grades, and previous naval experience**.** The domain knowledge data included both overall knowledge and knowledge related to specific sub-domains, i.e. combating flooding damage. Overall we expected a generally positive relationship of these scores to our performance measure, however we also expected certain knowledge areas to be more important. More precisely we expect knowledge areas deemed critical to damage control performance, e.g. fire fighting, to demonstrate stronger relations with our performance measure as compared to more peripheral knowledge areas, e.g. administrative procedures. Additionally, we expected previous naval experience to have a positive effect on simulator performance.

Being that the DC-Train simulator also serves a training function we also investigated improvements in performance over time, over and above those resulting

from interface learning. By comparing a group of students who had additional practice prior to the criterion test to a control group of participants not receiving additional practice, we investigated the existence of a practice effect. We also utilized the practice group to investigate reliability of the scoring procedure and the existence of a within-subjects practice effect.

## Method

### Pilot Study: Simulating Excellent, Average, and Poor Performance

An expert in both the damage control domain and the simulator was the sole participant in this study. He simulated excellent, average and poor performance by manipulating performance dimensions that are presumed to influence the performance measure. The expert's best performance served as the performance baseline, i.e. excellent performance, while average and poor levels of performance were achieved by manipulating the precision of fire boundary settings and the number of fire boundaries actually set. Given that the expert had extensive experience with the simulator and apriori knowledge of the upcoming damage events, we presume that he is able to achieve a superior level of performance. Normally when the user performs the scenario they have no idea as to the "when and what" of damage occurrence.

All other actions by the expert were kept consistent across the scenarios. Only the particular performance dimension was manipulated. To allow for comparisons, the same scenario was used throughout. Additionally each simulated performance, e.g. poor performance, was repeated a number of times to demonstrate the stability of the performance measure.

Participants

The participants in this study were 20 Damage Control Assistant candidates. These 20 students (14 male, 6 female) were enrolled in a six-week DCA training course at the US Navy's Surface Warfare Officer School. Participant age ranged from 22 to 43 years old, with a mean age of 29. Previous naval experience ranged from 1/2 to 23 years, with a mean of 7. All data were collected during the last 4 days of the course.

It is important to note that while the participants were trainees, they had acquired considerable knowledge and practical experiences via the six-week training course. Additionally, many of the participants had extensive at-sea experience. Given the infrequent nature of real-world damage control experiences, we can presume that these trainees are extremely similar to DCAs currently stationed onboard naval vessels. Indeed, these same trainees were given assignments as DCAs approximately one week after our testing session.

Task

Participants attempted to solve damage control scenarios presented by a multimedia ship-crisis simulator. The simulator bombards participants with auditory, visual, and textual information to simulate the information overload common in damage control situations. It is the participants' task to analyze the information and decide which actions, if any to take, by inputting commands to the interface via the mouse and keyboard. Following the execution of an action, the system verbally reports the request and the scenario develops accordingly. Scenarios are finished when either the ship sinks or all the inflicted damage has been successfully combated.

Experimental manipulation

Participants were randomly assigned to one of two groups: practice or no practice. In the practice condition, participants were given an orientation trial on how to use the simulator; two practice trials with the simulator; and the criterion trial. In the no practice condition, participants received only an orientation trial followed by the criterion trial.

Measures

Demographics & previous experience. Participants were asked their age, rank, years of naval experience, and amount of experience with other computer simulation training in damage control.

Course Grades. During the six-week training course, students were assessed based on their knowledge of sub-domains related to the DCA occupation. In all there were 9 separate grades that jointly determined the final grade. All grades were determined prior to simulator performance.

Performance Measure. After a scoring key is generated by Minerva (See Figure 1), Yardstick compares the actions in the key to the actions taken by the student. Actions can be scored as completely incorrect, partially correct, or completely correct. Completely incorrect actions receive an action score of 0.0; completely correct actions receive an action score of 1.0.

For some actions, the best action is the only acceptable action. For such actions, yardstick uses a binary scoring coefficient -- full action points or no action points. However, there are some situations in which there are several acceptable actions in addition to the best action. We call these acceptable actions “partially correct.” Partially correct actions are not optimal, and therefore not worth full points (1.0). However, they

are still acceptable, and therefore worth at least some points (greater than 0.0). Overall the points allocated to a particular action is meant to reflect the difference between the state of the ship that would result from the partially correct action as compared to the action suggested by the key, with greater differences resulting in lower point values. In other words it reflects the effectiveness of the actions taken. The actual point determination is more complex involving the consideration of multiple performance dimensions. For example, the setting of fire boundaries is assessed with regard to the spatial accuracy as well as the timeliness of the action.

Once all of the actions have been assigned scores, the average action score was calculated. This served as the performance score. Hence the performance score reflects the average level of effectiveness of user actions. It is important to note that errors of omission, the absence of a necessary action, are scored as zero and are included in the calculation of this performance score.

Procedure

Upon arrival at the criterion trial, all 20 participants were asked to fill out a questionnaire that assessed their background experiences and demographics. To ensure that all participants had a sufficient knowledge of the system, a 20-minute training module proceeded the criterion trial. In addition to the DCA, who had sole responsibility for making decisions, there was a "plotter" present to keep track of damage on dry-erase ship diagrams. The use of a plotter is standard operating procedure for damage control. . To insure independent assessment of the DCA, groups were discouraged from sharing decision-making power. All 20 participants completed one criterion trial as DCA that served as the basis of comparison between groups.

Prior to the criterion trials, participants in the practice group received additional training on the simulator. Participants completed 3 scenarios as DCA, again with the assistance of a plotter. All practice scenarios were judged to be of equal difficulty and similar content, i.e. damage incurred by the ship, by a panel of expert judges serving as instructors at the training school. The first trial was utilized to acquaint the participants with the interface and functioning of the simulator, while the latter two trials were used to establish reliability of the scoring system.

## Results

### Simulation of User-Behavior

Results show that the scoring system was able to distinguish between excellent, average and poor performance for each of the performance dimensions that were manipulated. Averages across the different performance manipulations were 61.77, 51.93, 32.34, respectfully. Further the extent to which the performance manipulation affected scores varied depending upon which performance dimension was manipulated, per our expectations.

### Reliability

A reliability coefficient was computed by correlating the last two trials, judged as parallel, for the ten participants in the practice condition. Results indicated a high level of reliability ($R = .824$), indicating a stable performance score.

### Validity

Investigation of correlations between criterion performance and the DCAs' demographic data revealed a relationship between performance and years of naval experience ($R=.35$, $p<.05$). No other demographic characteristics had any observable

relationship with performance. To investigate the relationship between domain knowledge and performance scores, correlations between criterion performance and the course grades were computed. The overall grade correlated .304 with criterion performance. Correlations ranged from .007 to .387 across the various sub-domains, but none were significant due to the small sample size. As was expected, knowledge areas that assessed trainees' relevant procedural knowledge generally showed a stronger relationship with simulator performance score than did declarative knowledge areas. (See Table 1 for correlations).

The existence of a practice effect was investigated by comparing the criterion performance of the practice and no practice groups. The mean for the practice group was 66.16 and for the no practice group it was 31.32. An ANCOVA with naval experience as a covariate was utilized to test for a between-subjects practice effect. Results revealed that after controlling for military experience, the group receiving practice had significantly higher scores on the criterion scenario than the group who did not receive practice ($F=11.77$, $p<.05$). Additionally the ANCOVA results revealed a significant effect of naval experience on performance ($F=12.71$, $p<.05$). Within-subject analyses of the practice effect proved insignificant primarily due to the restricted sample size ($n=10$).

## Discussion

Overall the scoring system represents a significant advance in decision-making assessment. In most cases of decision analysis only the final outcome and not the process itself is scored (Bolger & Wright, 1992; Keren, 1992). Because in many cases the final outcome is affected by multiple extraneous factors outside the decision-makers control, it

is a poor indicator of performance. For example, many scenarios that DC-Train simulates will inevitably result in the loss of the ship thus reducing the variance of an outcome-based criterion to zero. As Campbell (1990) has pointed out, the performance behavior itself and not its effectiveness should be the preferred criterion. By comparing operator actions to a dynamic expert system on-line, we focus on the actual behaviors and eliminate the influence of extraneous factors (such as the effectiveness of fire-fighting teams) on performance.

The preliminary results from our validation study, although based on small sample sizes, are promising. The pattern of correlations between the performance measure, past experience, and domain knowledge provide evidence of construct validity of our performance measure. Further the scoring system was discriminating in terms of performance differences between individuals receiving different amounts of practice.

To further establish the validity of the scoring system several modifications to the system and future validation studies are being planned. The performance score is being decomposed into multiple independent sub-scores reflecting salient distinctions in content. As such we will be able to compute internal consistency measures. Being that grades are susceptible to multiple biases, we are currently constructing our own domain knowledge test that will also have sub-scores that overlap with the sub-scores of the performance measure. By comparing the sub-scores from the simulator with those garnered from our domain knowledge measure we will more thoroughly investigate the construct interpretation of our performance measure. Finally, we are also constructing an instructor-rating instrument that will serve as an independent assessment of simulator performance allowing us to test for concurrent validity.

Methodology for the design and testing of simulator performance measures will continue to grow in importance as more and more domains are simulated with increasing sophistication. This level of sophistication however brings with it particularly complex measurement issues, which are just now being addressed. In discussing our work with DC-Train it was our attempt to provide a framework for thinking about such issues.

# References

Ackerman, P.L., & Humphreys, L.G. (1990). Individual differences theory in industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of Industrial Organization Psychology: Vol. 1. (pp. 225-282). Palo Alto, CA: Consulting Psychology Press.

Adelman, L. (1992). Evaluating Decision Support and Expert Systems. New York: John Wiley & Sons, Inc.

Baumann, M. R., Sniezek, J. A., Donovan, M. A. and Wilkins, D. C. (1996). Training Effectiveness of an Immersive Multimedia Trainer for Acute Stress Domains: Ship Damage Control. University of Illinois Technical Report UIUC-BI-KBS-96008, Urbana, IL: Beckman Institute.

Bolger, F. & Wright, G. (1992). Reliability and validity in expert judgement. In G. Wright & F. Bolger (Eds.), Expertise and decision support. (pp. 47-71). New York: Plenum Press.

Borman, W.C. (1991). Job behavior, performance, and effectiveness. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of Industrial Organization Psychology: Vol. 2. (pp. 271-327). Palo Alto, CA: Consulting Psychology Press.

Bulitko, V. V. (1997). "Expert Systems in Damage Control Domains" for the Knowledge Based Systems Research Group, April 1997. http://www-kbs.ai.uiuc.edu/bulitko/pub/minerva/ES_in_DC_domains.ppt

Campbell, J.P. (1990). Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of

Industrial Organization Psychology: Vol. 1. (pp. 687-727). Palo Alto, CA: Consulting Psychology Press.

Cream, B.W., Eggenmeier, F.T., & Klein, G.A. (1978). A strategy for the development of training devices. Human Factors, 20, 145-158.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart, & Winston.

Flexman, R. P. & Stark, E.A. (1987). Training simulators. In G. Salvendy (Ed.), Handbook of Human Factors (pp 1012-1037). New York: John Wiley & Sons.

Goldstein, I.L. (1991). Training in work organizations. In M.D. Dunnette & L.M. Hough (Eds.), Handbook of Industrial Organization Psychology: Vol. 2. (pp. 507-621). Palo Alto, CA: Consulting Psychology Press.

Guion, R.M. (1978). Scoring of content domain samples. Journal of Applied Psychology, 63, 499-506.

Guion, R.M. (1977). Content validity- the source of my discontent. Applied Psychological Measurement, 1, 1-10.

Hanson, M.A., Borman, W.C., Mogilka, H.J., Manning, C., & Hedge, J.W. (1999). Computerized assessment of skill for a highly technical job. In F. Drasgow (Eds.) Innovations in computerized assessment. Mahwah, NJ, USA: Lawrence Erlbaum Associates.

Hanson, M.A., & Ramos, R.A. (1996). Situational judgment tests. In R.S. Barret (Eds.) Fair employment strategies in human resource management. Westport, CT, USA: Greenwood Publishing Group.

Hunter, J.E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Clevelent (Eds.), Performance measurement and theory (pp.257-266). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Keren, G. (1992). Improving decisions and judgments: The desirable versus the feasible. In G. Wright & F. Bolger (Eds.), Expertise and decision support. (pp. 25-44). New York: Plenum Press.

Madni, A.M. (1988). The role of human factors in experts system design and acceptance. Human Factors, 30, 395-414.

Motowidlo, S.J., Borman, W.C., & Schmit, M.J. (1997). A theory of individual differences in task and contextual performance. Human Performance, 10, 71-83.

Payne, J.W. (1982). Contingent decision behavior. Psychological Bulletin, 92, 382-401.

Salas, E., Bowers, C.A. and Rhodenizer, L. (1998). It is not how much you have but how you use it: Towards a rational use of simulation to support aviation training. The International Journal of Aviation Psychology, 8 (3), 197-208.

Steadham, S.V. (1980). Learning to select a needs assessment strategy. Training and Development Journal, 30, 55-61.

Uhlaner, J.E. & Drucker, A.J. (1980). Military research on performance criteria: A change of emphasis. Human Factors, 22, 131-139.

Vreuls, D., & Obermayer, R.W. (1985). Human-system performance measurement in training simulators. Human Factors, 27, 241-250.

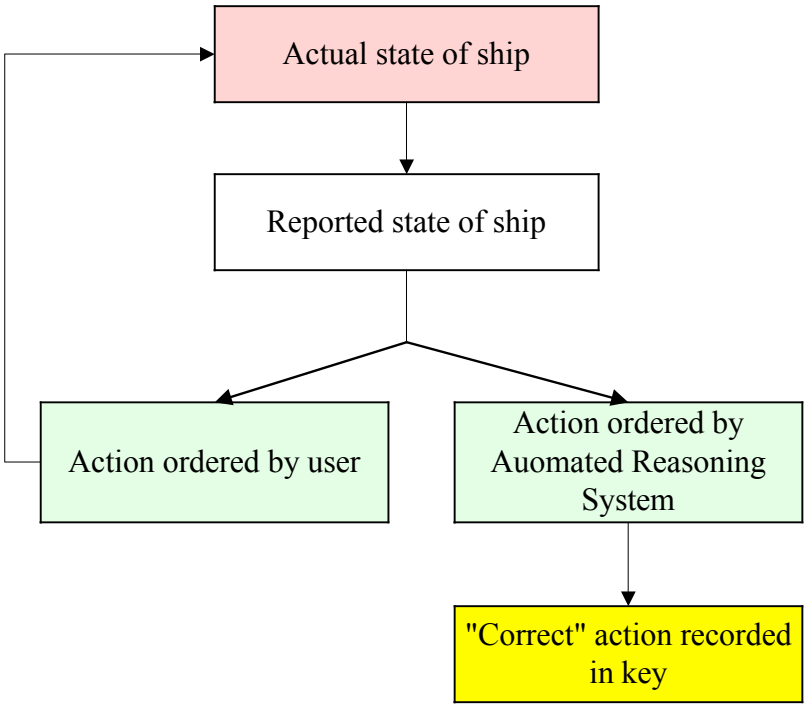
Weekly, J.A., & Jones, C. (1999). Further studies of situational tests. Personnel Psychology, 52, 679-700.

Wilkins, D. C., Baumann, M. R., Sniezek, J. A. and Donovan, M. A. (in revision). Decision-making performance on an immersive training simulator for ship damage control.

# Figure Captions

Figure 1. A graphical representation of online key generation.

Figure 2. The relationship between the environment, decision making performance, and outcomes.


Actual state of ship
Reported state of ship
Action ordered by user
Action ordered by Auomated Reasoning System
"Correct" action recorded in key

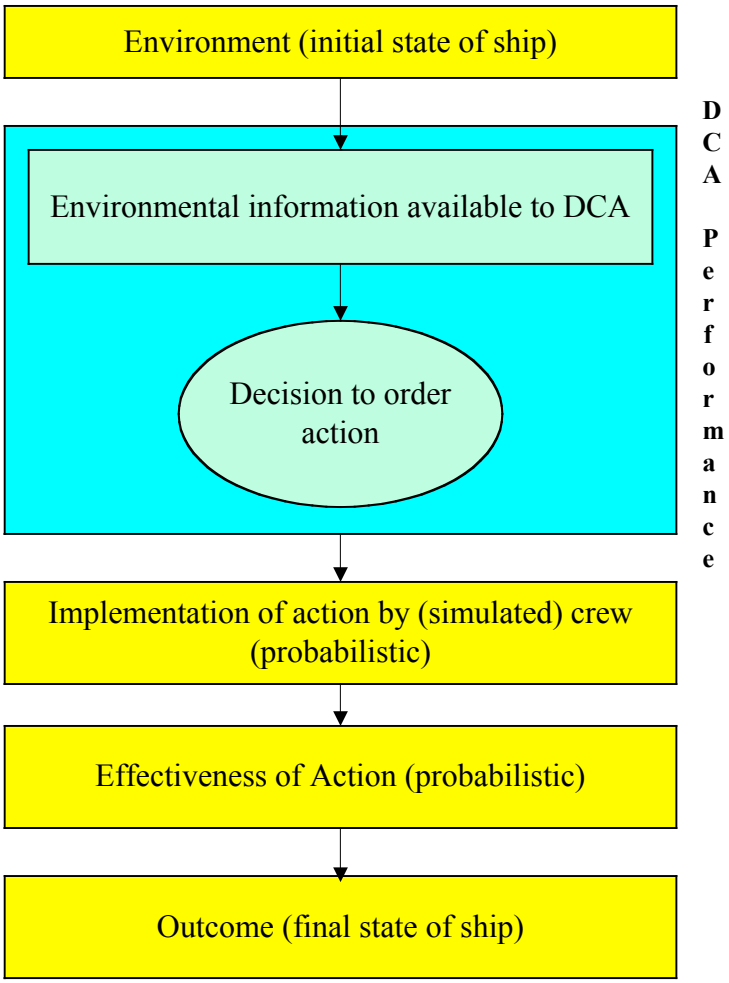

Environment (initial state of ship)
Environmental information available to DCA
Decision to order action
DCA Performance
Implementation of action by (simulated) crew (probabilistic)
Effectiveness of Action (probabilistic)
Outcome (final state of ship)

Table 1: Correlations amongst domain knowledge scores and criterion performance

| | Criterion score | Overall grade | *Damage Control procedure | Admin | *Fire-fighting procedure | *Stability procedure | Engineering principles |
|---|---|---|---|---|---|---|---|
| Criterion score | 1 | 0.30 | 0.39 | 0.01 | 0.36 | 0.24 | 0.02 |
| Overall grade | | 1 | 0.87 | 0.38 | 0.77 | 0.83 | 0.64 |
| *Damage Control procedure | | | 1 | 0.33 | 0.66 | 0.62 | 0.59 |
| Admin | | | | 1 | 0.15 | 0.00 | 0.51 |
| *Fire-fighting procedures | | | | | 1 | 0.53 | 0.65 |
| *Stability procedures | | | | | | 1 | 0.24 |
| Engineering principles | | | | | | | 1 |
| n=20<br>*domain areas hypothesized to correlate strongly with criterion performance<br>correlation between years of naval experience and criterion score=.35 | | | | | | | |