USING COGNITIVE WORK ANALYSIS TO DEVELOP A CAPABILITY FOR THE EVALUATION OF FUTURE SYSTEMS *

[1]David J. Crone, [2]Penelope M. Sanderson, and [1]Neelam Naikar

[1]Defence Science and Technology Organisation (DSTO), Melbourne, Australia
[2]ARC Key Centre for Human Factors, The University of Queensland, Brisbane, Australia

Evaluating complex human-machine systems is a multidimensional problem that requires the analyst to specify the characteristics of the system to be evaluated and to specify the method of evaluation to be used in determining its effectiveness. Whereas current approaches using task analysis techniques are used widely for evolutionary systems, other approaches, such as those from within the Cognitive Work Analysis (CWA) framework, may be better suited to revolutionary systems. This paper presents ongoing work directed at developing a human-in-the-loop simulation capability for revolutionary, or novel, systems that aims to evaluate the impact of a subsystem's performance (e.g. radar detection range) on the overall system's purpose (e.g. the safe transportation of troops). Two aspects of the development of this capability are presented: 1) the use of Work Domain Analysis and Control Task Analysis to represent systems for the purpose of providing measures of performance for human-in-the-loop experiments performed in support of system evaluation and 2) criteria for judging the utility of a CWA approach for system evaluation compared to other more traditional techniques.

INTRODUCTION

Determining the effectiveness of complex human-machine systems is a multidimensional task that encompasses specifying the characteristics of the system itself as well as specifying the method, or process, of evaluation used to determine its effectiveness or performance.

Within the Defence Science and Technology Organisation (DSTO) in Australia there is a general requirement to provide advice to our customers for the procurement of revolutionary (novel) systems. In order to provide this advice a system evaluation capability is being developed. The capability involves developing an evaluation process that includes a human-in-the-loop simulation (HILS) facility, the selection of *meaningful* performance measures (measures of performance are quantitative measures of low level physical performance, for example range and velocity; MOP) and the development of a means to track the impact of a subsystem's performance on the overall system purpose. In addition, if the capability is to be adopted the process of selecting the performance measure must be "useful" or "effective". For example, if the resulting evaluation capability provides "answers" that are accurate but too labour intensive to generate (compared to current approaches) then the adoption of the capability may be questioned.

In addition, we wish to determine whether the CWA approach may be more or less effective compared to current techniques when it is applied to evolutionary systems (systems that have been developed and are based on the functionality of a previous system) and revolutionary systems (systems with functionality that is novel -- not based on previous systems). The CWA approach may yield greater benefits in the evaluation of revolutionary systems rather than evolutionary systems when compared to other processes, such as those based in mission, function, and task analysis methods.

The work reported in this paper presents an extension of the use of Cognitive Work Analysis (CWA; Vicente 1999) for system evaluation. In particular, we will extend the use of Work Domain Analysis (WDA) and Control Task Analysis (CTA) to the problem of selecting performance measures for future (revolutionary) intentional systems and also determine criteria for assessing CWA against other approaches.

We show that the traditional process of selecting performance measures for system evaluation is influenced by a number of factors, including the use of task-based procedures. For future systems this includes predicting (or synthesizing) tasks or activities based on an understanding of current tasks with similar systems. However, although the process has been used widely, it may not be suitable for systems where there is little or no knowledge about how the system may be used (Naikar & Pearce, 2003).

Two main aspects of our work are reported. The first aspect concerns how we might use WDA and CTA to model our candidate system (Australian Army helicopter) and extract useful performance measures. By using WDA and CTA, not only are performance measures produced that relate to system purpose and values, but also it is made clear how low-level subsystem properties affect high-level mission purposes. The second aspect of our work concerns how we might develop criteria to judge whether making decisions based on CWA-based performance measures leads to more or less effective decisions when compared to current task-analytic based techniques.

SELECTING PERFORMANCE MEASURES

Complex systems have human, software and hardware components that are purposefuly integrated to perform specific functions in an organisational and physical environment. Hence, the concept of determining a system's effectiveness has been approached from many different domains of practice: for example, from human factors and systems engineering practice. Practitioners within these domains, however, have tended to take a discipline-centric view, with relatively little account taken of other significant factors. For example, within Human Factors Engineering (HFE) the choice of what measure of performance to use is commonly seen to be a function of experimenter experience (Muckler & Seven, 1992) and a reflection of the resources that are available to the experimenter. This is illustrated clearly within the controlled experiment paradigm, used widely in basic and applied research, in which the experimenter chooses a metric (for example, reaction time or error rate) that they feel will adequately reflect the impact of an independent variable during a trial of some sort in the particular research environment. What processes there have been in the past for selecting measures have been driven by ensuring that measures meet a series of criteria: for example validity, reliability and sensitivity.

In addition to valuing these criteria, some researchers have specific preferences for the types of metrics they employ. These preferences reflect the theoretical perspective that researchers prefer and also may reflect requirements from customers. For example, Meister (1999) advocates using objective measures over subjective ones, taking non-physiological measures in preference to physiological measures, embedding measures in operator tasks, and taking easily collected metrics over ones that may prove difficult to collect. In contrast to Meister, Hennessey (1990) takes the view that observational techniques, rather than more objective measures (for example, automatically recorded data), should be used for human performance testing. Irrespective of the type of data collected both authors argue that measures should be made during field-testing of systems (i.e. in an operational environment during operational tasking).

During operational testing, measures of performance are based on tasks performed during typical missions or scenarios. A number of mission scenarios are distilled into a small number of representative missions that include critical tasks. Performance measures are then associated with the critical tasks. If the human or system achieves the required performance during the critical tasks it is assumed that the mission can be performed. Determining the performance of future systems essentially follows the same process except that subject matter experts and analysts synthesize the scenarios and tasks used. In other words, the tasks used are considered representative of how the future system *may* operate.

Although Hennessey (1990) advocates performance testing under operational conditions, he argues that methods used during operational test and evaluation often fail to separate the machine and human components of design shortcomings in a way that will aid system design. For example, a human performance measure in a military context could be how closely a weapon hits a target. However, this measure clearly is affected by both human and system characteristics. The "rolling-up" of both the human and the machine components to produce a measure that is considered a measure solely of human performance when in fact it is not, limits the utility of the measure—one cannot partition out performance attributable to the human or system. Therefore, the use of metrics such as this one for system design in a diagnostic sense, is limited.

Clearly, selecting performance measures is a function of experimenter experience, measurement theory and operational concerns. In addition, selecting performance measures is currently based on a task-analytic approach. Even though analysts recognise the importance of testing a system in operational conditions there still remains problems of identifying what tasks to use in experiments, where a change in system performance originates, and how a low-level subsystem change influences high-level system purpose. What we need, therefore, is a way of developing a system representation that allows us to see the relationships between low-level system properties and high-level system purposes in a way that is task independent. In addition, the system representation must allow us to quantify these relationships.

Evaluating Future Systems with CWA

So far we have looked mainly at performance-measure selection for existing systems or for future systems where some knowledge of likely tasks exists. But what about future systems where there is no real understanding of how the system may be used?

System evaluation is more complicated when one considers the design and evaluation of revolutionary new systems. Traditionally, new systems are modifications of older systems. This means that generally there is some knowledge about operational factors, based on the old system, which can be used for design and evaluation purposes. For example, a new system may be designed to perform the same operational task as the one it is replacing; hence, the same task-based performance measure may be used to assess the effectiveness of the new system. However, what performance measure should be employed to evaluate a system if it offers new functionality and there is no knowledge about the functionality and its implications for human-system integration? We are currently faced with this challenge and are developing an evaluation capability that meets it.

Because of the particularly challenging nature of evaluating unbuilt, future systems, and the limitations in traditional approaches to system evaluation (especially performance measure selection), the CWA framework is being investigated as a key component for the evaluation capability. The CWA phases have been applied in many contexts over the past fifteen years including laboratory contexts and, more recently, operational systems contexts, such as command and control (Burns, Bryant & Chalmers, 2000), and tender evaluation, (Naikar & Sanderson, 2001). CWA has also been shown to be applicable throughout the whole system life-cycle (Sanderson, Naikar, Lintern & Goss, 1999) and has been used to determine human performance measures within a *causal* micro-world system (Yu, Lau, Vicente & Carter, 2002). Here we extend the use of CWA for performance measure selection for *intentional* systems evaluation. In particular, CWA is being used to model the candidate system (Australian Army helicopter) and provide a means to select diagnostic measures of performance and track the impact of a subsystem on the overall system's purpose.

Performance Measures for System Evaluation

WDA presents and indicates the relationship of the human values and priorities with software and hardware components of the work domain and their purpose. The integration of these different components and the ability to reflect on how they function together to achieve system purposes is a unique strength of WDA. The WDA is made up of a series of nodes. Each node has a number of properties associated with it.

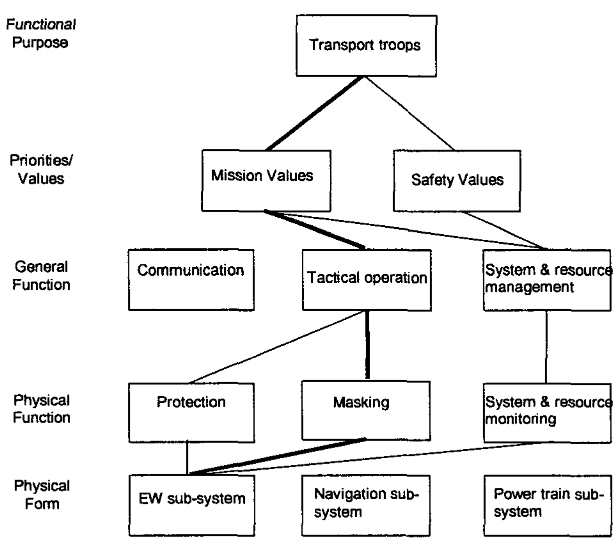

Figure 1 Simplified extract from WDA for Australian Army helicopter

Figure 1 shows a portion of the WDA produced as part of the Australian Army helicopter project. Table 1 lists a number of properties associated with each of the highlighted nodes and associated measures. Of interest to us is the way measures may be mapped onto the nodes and how the WDA may be used to make predictions about the effect of changing low-level system properties on the ability of the system to achieve its purpose.

Table 1 Simplified extract of properties and measures from Australian Army helicopter WDA

Abstraction Layer	Work domain node	Node properties and measures
Functional Purpose	Transport troops	– Can the troops be transported? (Yes'/ No)
Priorities and Values	Mission values	- Maximise distance from enemy (may be measured in terms of absolute distance between enemy assets and own assets). - Maximise time available to react (may be measured in terms from detection of enemy to possible response). - Minimise probability of damage (may be measured in terms of height above terrain).
General Function	Tactical operation	- Timeliness of helicopter arrival (may be measured in terms of difference in time between planned time of arrival and actual time of arrival) - Surprise of enemy (may be measured in terms of closeness of enemy assets to landing point)
Physical Function	Masking	- Reflection of RF energy (may be measured in terms of difference between amount of RF energy produced by enemy radar and the amount of RF energy reflected by the chaff strip or cloud minus the amount of RF energy lost through other processes)
Physical Form	EW sub-system: Chaff Radar Warning Receiver	- Chaff dimensions (length, breadth, width, mass) - Material used in construction (chemical composition) - Signal processing speed (may be measured as time taken to detect, process, classify and present information) - Number of contacts processed (may be measured as the absolute number or as a percentage of total contact available)

At the physical form layer *measures of performance* for each of the properties of the object may be derived. Similarly, measures may be derived for the properties of the nodes at the physical function, general function and priority/ values layers. *Measures of effectiveness* (a measure of effectiveness is a quantitative or qualitative measure that indicates how well the system meets its purpose; MOE) may be derived at the functional purpose layer.

Clearly, not only are there many possible MOPs and MOEs associated with a work domain, but there are also many possible unique relationships or links between the properties at each abstraction layer. For example, the object "chaff", at the physical form layer may have five properties (length, breadth, width, mass and chemical composition); the physical function node it is linked to, "masking", may have one property (reflection of RF energy); the general function node "tactical operation" may have two properties (timeliness and surprise); the priority/ value node, "mission values", may have three properties (maximize distance, maximize time and minimize probability); and the functional purpose node may have one property. This equates to 30 unique alternatives, or 30 possible predictions linking object properties (and their respective MOPs) to the system's purpose (and its respective MOEs). This is without considering the relationships between nodes on the same abstraction layer. Because of the sheer amount of possible predictions a method is needed to make the problem manageable.

CTA presents the limitations on human action as a result of the system control limitations and may be represented as a Temporal Coordination CTA (TC-CTA: Sanderson & Naikar, 2000; see also Naikar & Pearce, 2003). Measures associated with TC-CTA indicate how an activity is constrained, for example by time or the action of another activity.

Figure 2 presents an extract of the TC-CTA produced for the Australian Army helicopter. Control tasks are represented as a movable bead that may occur within a particular time span. Each task has a number of properties associated with it. For example, each has a completion time and an error rate. In addition to these task-specific properties, there are several properties related to all the tasks. Table 2 provides an example of the properties associated with the control task of "Manage EW system: operate in response to threat". These include task-specific properties (for example, "time task takes" and "start time") and general properties (for example, task priority and task sequence). Hence, there are many possible interactions within and between task properties that need to be captured. For example, a change in the "start time" of a task may cause a change in the sequencing of all the tasks.

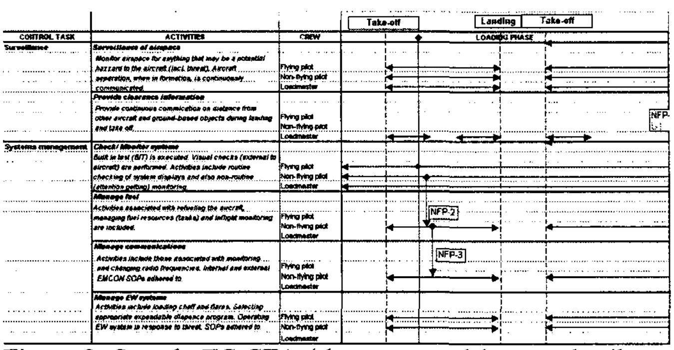

Figure 2. Sample TC-CTA (shows general layout: details not intended to be legible).

The relationship between WDA and CTA is important to our work. The system properties and a system's purpose, as measured by MOPs and MOEs and represented in the WDA, *afford* various activities, seen in the TC-CTA. This means that a change in a system property may change a control task property. For example, we may predict that changing the chaff property, "material", will affect the total system's purpose (transport troops) *and* that change will affect the control task property "the time it takes to operate system in response to threat".

In addition, the control tasks themselves will *act on* the work domain of interest. Therefore, by clearly identifying control tasks that have changed as a result of an object change we have an objective mechanism for identifying tasks that should be used for human-in-the-loop simulation trials and we have a method by which to target data collection.

It is clear, however, that because of the sheer amount of possible predictions associated with the WDA and CTA and their interaction a method is needed to make the problem manageable. To this end a Microsoft Access™ database has been developed to provide a means to extract objective logical predictions linking object properties to the system's purpose and control tasks, rather than just all-possible predictions.

Table 2 Example of properties derived from a control task

Control Task	Properties
Manage EW system **- Operate systems in response to threat**	**- Time task takes** **- Start time of task** **- End time of task** **- Number of errors made during task** **- Type of errors made during task** **- Can occur after take-off** **- Must end before landing** **- Higher priority than "Manage fuel"** **- Only occurs in response to threat** **- Occurs with "Control aircraft"**

COMPARING CWA TO CURRENT TECHNIQUES

Although CWA offers a way to model systems and select performance measures for evaluating new systems, does it provide a better or more useful way of doing so than other existing techniques? Figure 3 formalises the problem.

On one hand we have a universe of possible predictions: {A, B, C, D, E, G} about the relationship between object properties (and MOPs) and mission purpose (MOE) and activity. Predictions may emerge from CWA-based techniques only {D, E}, from current task-analytic techniques only {G}, or from each technique equally {A, B, C}. These predictions are therefore an *output of the technique* used and may be examined in the human-in-the-loop simulation.

On the other hand the techniques used to derive the predictions can be evaluated for their relative effectiveness. To answer this question we need a set of criteria with which to assess the merits of the evaluation techniques. Table 3 presents a sample of criteria that have been developed from an understanding of what constitutes the "traditional" system evaluation process and the requirements of the various system evaluation stakeholders. Once the criteria are validated they will be used to compare the CWA process against traditional process for both evolutionary and revolutionary systems. For example, we may have a set of criteria {time, diagnosticity} where time is the speed with which predictions can be made with the different techniques and diagnosticity is the validity of the prediction for real helicopter operations. We may find that predictions {A, B, C} are slower to emerge with CWA than with a conventional technique. However, predictions {D, E} are so powerfully diagnostic that it is worth the extra time to find them.

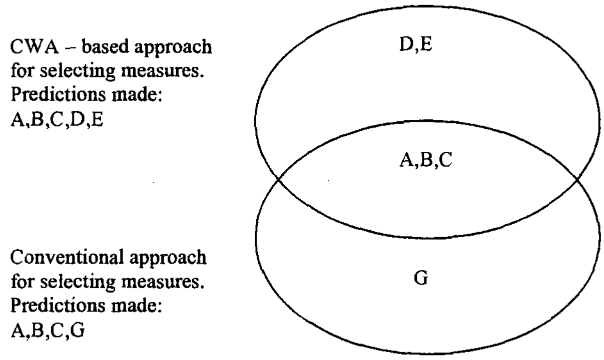

Figure 3. Comparing performance measure selection across techniques.

Future work comparing approaches

The WDA of an Australian Army helicopter conducting a class of missions has been constructed using subject matter experts and Australian Army documents. It has been validated using a scenario-based process advocated by Burns, Bryant and Chalmers (2001). A TC-CTA has also been produced. Finally, a Microsoft Access™ database represents the WDA and the CTA in a way that predictions can be made. An analysis has been produced of a typical helicopter mission.

In the short term, a series of trials is planned to test the predictions made using CWA against predictions using the task-analytic techniques. The comparison will be made using a human-in-the-loop simulation environment. Initially the trials will be based on a current avionic subsystem (a evolutionary system). Later trials will test predictions for a future avionic subsystem (a revolutionary system).

CONCLUSIONS

This paper shows that CWA may have a significant role to play in the selection of objective performance measures for future, revolutionary systems. By using WDA and CTA, the effect on the mission purpose and human activity of changing an object's properties may be followed in a way that is objective and quantifiable. In addition, because of the relationship between WDA and CTA tasks may be selected for human-in-the-loop simulation trials that are known *a priori* to be ones that are sensitive to a change. Hence data collection becomes targeted.

By comparing the processes taken to generate performance measures against a wide ranging set of customer focused criteria, we shall be able to identify whether the process taken is useful and practical or not.

Table 3. Sample assessment criteria for different techniques

Stakeholder community	Process evaluation criteria	Method of assessing criteria
Systems Engineering	Is the process complementary to known systems engineering test and evaluation standards and processes?	- Qualitative. Can the process be integrated with IEEE 1220? - Qualitative. Do the outputs of the process support the aims of the IEEE 1220 phases?
Technology Owners	Does the process aid the selection of alternative systems at various levels of abstraction?	- Qualitative. Are assessments of operator workload and situational awareness supported using the process? - Qualitative and quantitative. Can system properties (hardware and software) be clearly identified? Can their relative importance be quantified?
	Does the process produce "timely" answers?	- Quantitative. How long does it take to produce outputs to the process?
Human Factors Engineering	Does the process support analytic and empirical investigations within a human-in–the loop simulation environment?	- Qualitative. Can relevant dependant and independent variables be identified from the process products?
	Are the process outputs (models) valid?	- Qualitative and quantitative. Do the models contain elements that support scenarios other than the scenarios used in their development?
	Are the process outputs (model) reliable?	Qualitative and quantitative. Do the models produce the same information on different occasions? If both experts and novices use the model, are the results the same?

REFERENCES

Burns, C. M., Bryant, D., & Chalmers, B (2000). A work domain model to support naval command and control. *Proceedings of the 2000 IEEE International Conference on Systems, Man, and Cybernetics,* 2228-2233.

Burns, C. M., Bryant, D., & Chalmers, B (2001). Scenario mapping with work domain analysis. *Proceedings of the 45th Annual Meeting Human factors and Ergonomics Society, Minneapolis, MN.*

Hennessy, R. T. (1990). Practical human performance testing and evaluation. In H. A. Booher (Ed.), *Manprint: An Approach to Systems Integration.* Van Nostrand Reinhold, New York.

Meister, D. (1999). Measurement in aviation systems. In D. Garland, J. Wine and V. Hopkin (Eds), *Handbook of Human Factors.* Lawrence Erlbaum.

Muckler, F. A., & Seven, S. A. (1992). Selecting performance measures: "Objective" versus "subjective" measurement. *Human Factors, 34 (4),* 441-455.

Naikar, N., & Pearce, B. (2003). Analysing activity for future systems. Proceedings of the 47th Annual Meeting of the Human Factors and Ergonomics Society. Denver, CO: 13-17 October.

Naikar, N., & Sanderson, P. (2001). Evaluating system design proposals with work domain analysis. *Human Factors, 43(4)*, 529-542

Sanderson, P., & Naikar, N. (2000). Temporal coordination control task analysis for analysing human system integration. *Proceedings of the Joint Meeting of the Human Factors and Ergonomics Society and the International Ergonomics Association (IEA2000/HFES2000).* Santa Monica, CA: HFES. Vol 1, Pp 206-209.

Sanderson, P. M., Naikar, N., Lintern, G., & Goss, S. (1999). Use of cognitive work analysis across the system life cycle: Requirements to decommissioning. In *Proceedings of the 43rd Annual Meeting of the Human Factors and Ergonomics Society.* Santa Monica, CA: HFES, 318-322.

Vicente, K. (1999). *Cognitive work analysis: Towards safe, productive, and healthy computer-based work.* LEA. Mahweh NJ.

Yu, X., Lau, E., Vicente, K. J. & Carter., M. W. (2002). Toward theory-driven, quantitative performance measurement in ergonomics science: the abstraction hierarchy as a framework for data analysis. *TIES, 3 (2),*124-142.