Ergonomics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/terg20

Tying up loose ends: a method for constructing and evaluating decision aids that meet blunt and sharp-end goals

N. Keller[ab], U. Czienskowski[a] & M.A. Feufel[ab]

[a] Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Berlin, Germany

[b] Department of Anesthesiology and Intensive Care Medicine, Charité University Medicine, Berlin, Germany

Published online: 22 May 2014.

To cite this article: N. Keller, U. Czienskowski & M.A. Feufel (2014): Tying up loose ends: a method for constructing and evaluating decision aids that meet blunt and sharp-end goals, Ergonomics, DOI: 10.1080/00140139.2014.917204

To link to this article: http://dx.doi.org/10.1080/00140139.2014.917204

Ergonomics, 2014
http://dx.doi.org/10.1080/00140139.2014.917204

Taylor & Francis
Taylor & Francis Group

Tying up loose ends: a method for constructing and evaluating decision aids that meet blunt and sharp-end goals

N. Keller[a,b]*, U. Czienskowski[a] and M.A. Feufel[a,b]

[a]*Max Planck Institute for Human Development, Center for Adaptive Behavior and Cognition, Berlin, Germany;* [b]*Department of Anesthesiology and Intensive Care Medicine, Charité University Medicine, Berlin, Germany*

(*Received 10 June 2013; accepted 11 April 2014*)

We present a methodological framework for constructing and evaluating decision aids – fast and frugal trees (FFTs) – ideally suited to the front line of an organisation. Their performance can be analysed in signal detection theory, allowing for transparent selection of FFTs given managerial-level trade-offs among type I and II errors. We extend FFTs from binary classification to selection from multiple actions (FFT multiple) as well as performance analysis to organisational goal states beyond type I and II error reduction. Concepts and framework are introduced and a tutorial-style example application (threat assessment at military checkpoints) is provided. Throughout, we discuss ways to deal with missing or incomplete data and show that the performance of decision aids may be overestimated if the effectiveness of actions is not heeded. The methodology can be used to construct and evaluate decision aids in any area characterised by dichotomised cues and a one-to-many mapping between categorisation outcomes and actions.

Practitioner Summary: The paper presents a methodological framework for the construction of decision aids and their evaluation along multiple goal states across institutional levels. We then apply this framework to construct and evaluate decision aids for threat assessment in military operations. Ways to deal with missing and incomplete data are discussed.

Keywords: fast and frugal trees; signal detection theory; decision support; military personnel

1. Introduction

When categorising a patient as having a high or low risk of a disease or classifying an approaching vehicle as a high or low threat at a military checkpoint, many factors determine the goodness of that decision. One way of categorising performance criteria is captured by Reason's (1990, 1995) distinction between sources of error at the 'blunt' versus 'sharp' end of an organisation. At the blunt end, that is the managerial level, consequences of decisions may be valued differently from how they are valued at the sharp end – for instance 'misses' may be considered more costly than 'false alarms' or vice versa. At the sharp end, that is the front line of an organisation, decisions should be accurate with respect to strategic trade-offs set at the blunt end, but they must also be made with attention paid to practical constraints such as the accessibility of information, availability of equipment or time pressure. To date, the connection between decisions made at the blunt end and feasible actions taken at the sharp end has been rather loose, often requiring operators to invent workarounds to get the job done, despite well-intentioned regulations (Reason 1990).

Although taking up Reason's (1995) systems approach to human error, much of the work in applied cognitive psychology, human factors or ergonomics deals with the identification of error pathways at the operator level (sharp-end) without crossing system boundaries during analysis (Waterson 2009). To span the blunt-end – sharp-end divide, we present a methodological framework that rests on two well-established theoretical concepts – fast and frugal trees (FFTs) and signal detection theory (SDT) – for helping researchers and practitioners develop and evaluate decision support tools across a large range of contexts. FFTs are a type of decision tree for binary-choice diagnostic tasks that can be analysed in a signal detection framework within which the setting of the decision bias allows for consideration of blunt-end trade-offs between 'hits' and 'misses'. We extend current theories by coupling FFTs for categorisation with concrete plans of action allowing more realistic integration of sharp-end constraints. We also extend performance analysis beyond type I (false alarm) and type II (miss) errors to additional effectiveness and efficiency criteria that may be relevant to an organisation.

In Section 2, we outline FFTs, their relationship with the SDT framework, and how these concepts can be extended by integrating concrete plans of action. Section 3 is a tutorial-style example application of the methodological framework to threat assessment at military checkpoints. The example is intended to confront those wishing to apply this methodology with challenges they may face in the field and to present possible solutions.

*Corresponding author. Email: nkeller@mpib-berlin.mpg.de

2. FFTs, SDT and plans of action

In a Michigan hospital, physicians sent about 90% of patients with intense chest pain to the coronary care unit (CCU). The diagnostic task was to predict whether such patients would later suffer from a heart attack, and the strategy was 'defensive' (Gigerenzer and Engel 2006): given that physicians are more likely to be sued for under-treatment (type II error) as opposed to over-treatment (type I error), the decision to place most chest pain patients in the CCU made sense from the sharp-end perspective. From the blunt-end perspective, this strategy was undesirable due to increased costs and risks of secondary infections in the CCU. As a remedy, the Heart Disease Predictive Instrument (HDPI), consisting of a chart with about 50 probabilities and a pocket calculator with a logistic regression (LR) program, was developed (Green and Mehr 1997). After this decision aid was presented to the physicians – but not yet implemented – their diagnostic performance improved. The authors proposed that instead of using the HDPI, physicians focused on only the most important cues (pieces of diagnostic information) they could remember from the HDPI and integrated them in a process similar to the FFT presented in Figure 1. Using simulated reconstructions of Green and Mehr's (1997) data, it was found that this simple tree had a higher number of hits and a lower number of false alarms than the physicians' original performance and the LR-based HDPI across a number of threshold settings regarding the likelihood of heart attack.

2.1 FFTs

How can a simple categorisation tree outperform LR? FFTs are a family of categorisation trees that process one binary cue at each level (Figure 1). As with other trees, they consist of *question nodes* in which a question is asked about the value of a particular cue (e.g. present vs. not present), and *exit nodes* in which the situation is categorised and the process stops. A categorisation tree is fast and frugal only if it has at least one exit at each level (Martignon, Katsikopoulos, and Woike 2008). Any binary diagnostic procedure can have various measures of performance with respect to a binary criterion, for example:

- Accuracy = (hits + correct rejections)/(all objects)
- Positive predictive value = (hits)/(hits + false alarms)
- Negative predictive value = (correct rejections)/(misses + correct rejections)
- Sensitivity = (hits)/(hits + misses)
- Specificity = (correct rejections)/(false alarms + correct rejections)

FFTs are computational algorithms and can be analysed using formal mathematical methods and simulations, making it possible to test their performance against other formally defined decision aids. Comparisons have shown that FFTs are

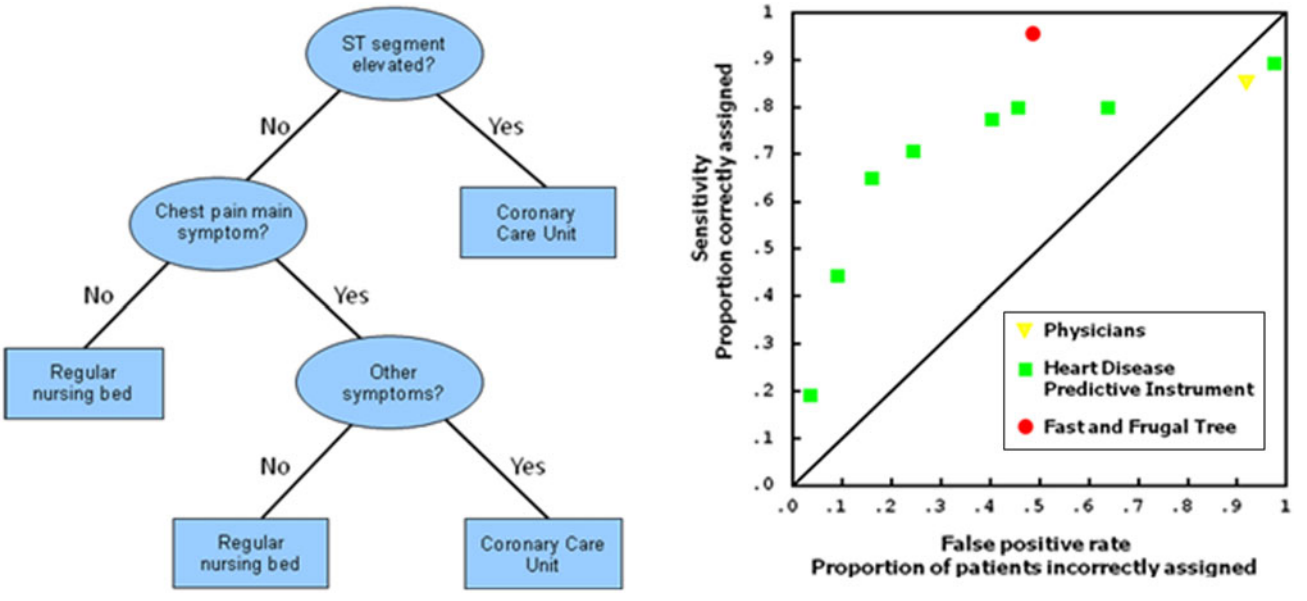

Figure 1. Left: An FFT for allocating patients (ST segment is a measure of heart-rate variability). Right: Receiver-operating-characteristic (ROC)-space plotting hits (*y*-axis) against false alarms (*x*-axis) for the FFT (red circle), physicians' original performance (yellow triangle), and the HDPI with varying threshold settings for classification as heart attack (green squares).

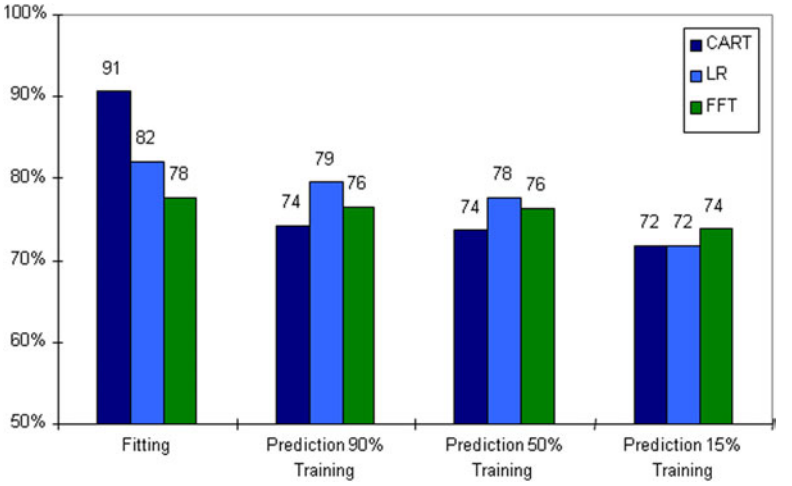

Figure 2. Average accuracy of three categorisation models across 11 medical problems in fitting and prediction with varying sizes of training sets (90%, 50% or 15% of the complete data-set) for parameter estimation (Martignon, Katsikopoulos, and Woike 2008).

robust. Unlike more complex decision aids, they incur relatively small losses in accuracy when making out-of-sample predictions under conditions of small sample sizes for parameter estimation (i.e. cue order, exit structure and so on). Martignon, Katsikopoulos, and Woike (2008) compared the predictive accuracy of FFTs to two 'gold standard' algorithms – LR and classification and regression trees (CART) – across 11 medical categorisation problems (e.g. diagnosis of heart disease, breast cancer and diabetes).

As shown in Figure 2, when all data can be used to estimate model parameters, LR and the CART outperform the FFT (fitting condition). When 90% of the entire sample is used for parameter estimation and the task is to classify the remaining unseen 10%, CART's performance drops by 17%. Both more complex models perform worse than the FFT when the training set is reduced to 15% and the remaining 85% have to be classified. In other words, FFTs are at least as effective as more complex tools when it comes to predicting new data based on small sample sizes. FFTs display this property because they rely on comparatively fewer parameters for how they process cues, leading them to ignore more of the accidental variation (noise), which impacts parameter estimation especially in small samples (for a theoretical account of this 'less-is-more' effect, see Gigerenzer and Brighton 2009).

2.2 Sharp-end features of FFTs

One advantage for application is that FFTs use few cues in sequential order and can therefore accommodate information-processing limitations. Also, they provide deterministic classification (either A or B) as opposed to likelihoods or other types of probabilistic information. Their simple structure makes them easy to communicate, teach and remember, and consequently facilitate their introduction into existing workflows. For example, in the medical domain, being able to utilise a decision aid at the point of care, without having to perform elaborate procedures, has been identified as a critical factor to successful implementation (Kawamoto et al. 2005). It has also been shown that use of computers for clinical judgment undermines perceived competence and the trust patients place in their physicians (Arkes, Shaffer, and Medow 2007).

2.3 Blunt-end features of FFTs and SDT

Luan, Schooler, and Gigerenzer (2011) worked out the conceptual mappings between FFTs and SDT. SDT separates the ability of a diagnostic process to differentiate between a true signal and background noise (diagnostic capacity or d') from its adaptive response bias under conditions of uncertainty (response criterion c; Green and Swets 1966). Luan, Schooler, and Gigerenzer (2011) showed that the positive and negative predictive values of cues and their ordering in an FFT correspond to the d' of that FFT. Moreover, the choice of a particular FFT exit structure corresponds to the setting of the decision criterion c.

Figure 3 shows four FFTs, all using the same d' (i.e. cues and cue order), but different response criteria c (i.e. exit structures). For blunt-end decision-making, this means that given a particular set and ordering of diagnostic cues, management can construct all possible sharp-end FFTs in advance and select the one representing the response criterion closest to achieving the desired organisational goals with respect to the costs of false alarms (type I errors) and misses (type II errors).

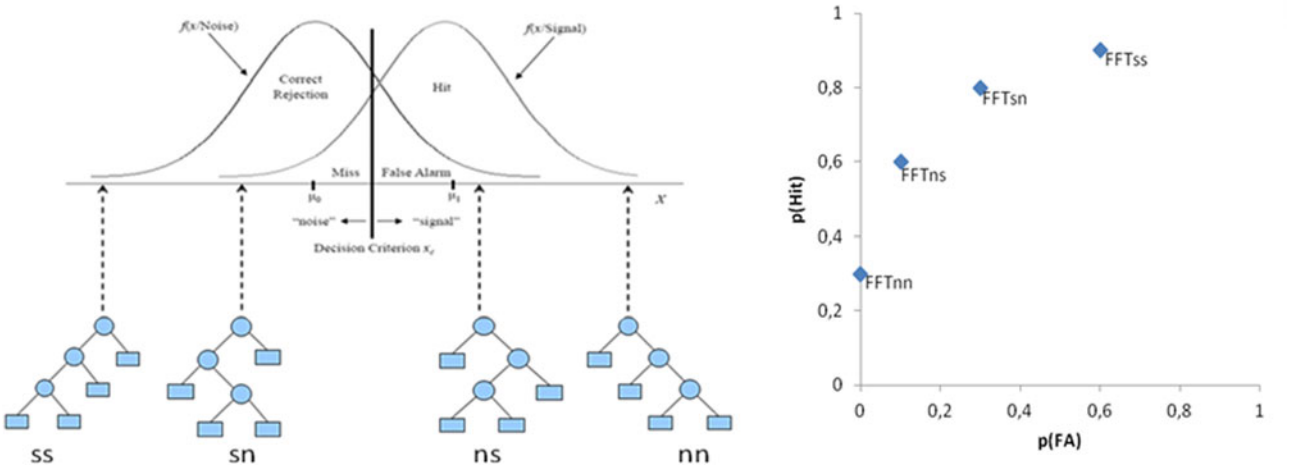

Figure 3. FFTs with the same cue order but different exit structures represent different settings of the response criterion *c* (left), which in turn correspond to performance differences within ROC space (right; FA, false alarm).

2.4 FFTs and plans of action

Decision aids are only useful if they suggest concrete plans of action or provide informational input for further analysis of a situation. So far in the application and theory of FFTs, there has been an implicit assumption of a one-to-one mapping between categorisation and the action to be performed. For example in Green and Mehr (1997), categorisation as high or low risk of a heart attack directly maps onto 'place in CCU' or 'place in regular nursing bed', respectively. However, in practice, often multiple actions or action sequences may be applied after categorisation (one-to-many mapping). The question is then which mapping of actions to exits is most effective. We call FFTs in which there is one-to-one mapping between exit nodes and actions FFT-binary, and FFTs in which there is a one-to-many mapping between exits and actions FFT-multiple. To evaluate these mappings, we differentiate between two action characteristics: action *effects* and action *constraints.*

Action effects are the success rate of an action *A* to attain a goal state *G*. Goal states are multidimensional (as many as there are goals) and present the identified goal-relevant measures across the different levels of the organisation:

$$\text{Action effect } A_{G_n} = \frac{\text{number of times } A \text{ achieves } G_n}{\text{number of times } A \text{ was used overall}}.$$

Each of the possible actions has a different impact on each of the goal states (ranging from 0 to 1). These may relate or be identical to reduction of type I or type II errors, but not necessarily. Taking the Green and Mehr (1997) study again, it is clear that the correct placement of patients is not an end goal but a means of reducing patient mortality and cost. If placement in the CCU saves only half of the patients who later suffer from a heart attack, this puts an upper bound on the institution's ultimate goal of reducing mortality that cannot be improved with higher classification accuracy. An analysis that focuses purely on categorisation accuracy without taking into account the effects of actions on the achievement of primary endpoints will likely arrive at over-optimistic estimates of a decision aid's performance in the field. Perhaps for this reason, systematic reviews by Heselmans et al. (2009) and Hunt et al. (1998) found that none of the decision support systems included in their reviews significantly improved patient health outcomes even though they significantly improved classification performance of healthcare practitioners. Careful *a priori* analysis and quantification of the effects of sharp-end actions on blunt-end goal states of an institution are therefore an important part of this framework.

Action constraints are sharp-end constraints on practitioners' abilities to implement and perform certain actions or action sequences in the field (e.g. time constraints, resource constraints, workflow constraints and so on). If a decision support system developed at the blunt end does not respect the constraints at the sharp end, end users will be forced to construct workarounds and implementation will be less efficient or fail altogether (Reason 1990). For example, a soldier may have time for only a certain number of actions before a contact breaches a checkpoint or may not have certain pieces of equipment available. Action constraints are central to ruling out infeasible actions and reducing the option set to be considered in designing and evaluating decision support systems.

3. Outline of the methodological framework

We propose a methodological framework for the construction of FFTs with multiple actions and their comparative evaluation across multiple goal states. The framework consists of three consecutive phases: (1) component analysis, (2) simulation and integration and (3) strategic selection (see Figure 4). During component analysis, the building blocks – actions and cues – of the FFT are identified and quantified. This breaks down to

(a) Identification of available cues and processing constraints.
(b) Identification of available actions and action constraints.
(c) Quantification of diagnostic cue measures and ordering of cues.
(d) Identification of relevant organisational goals and calculation of action effects.

In the integration and simulation phase, the identified building blocks are used to construct all FFTs that are feasible, given processing and action constraints at the sharp end. Then, computer simulation is used to comparatively evaluate their

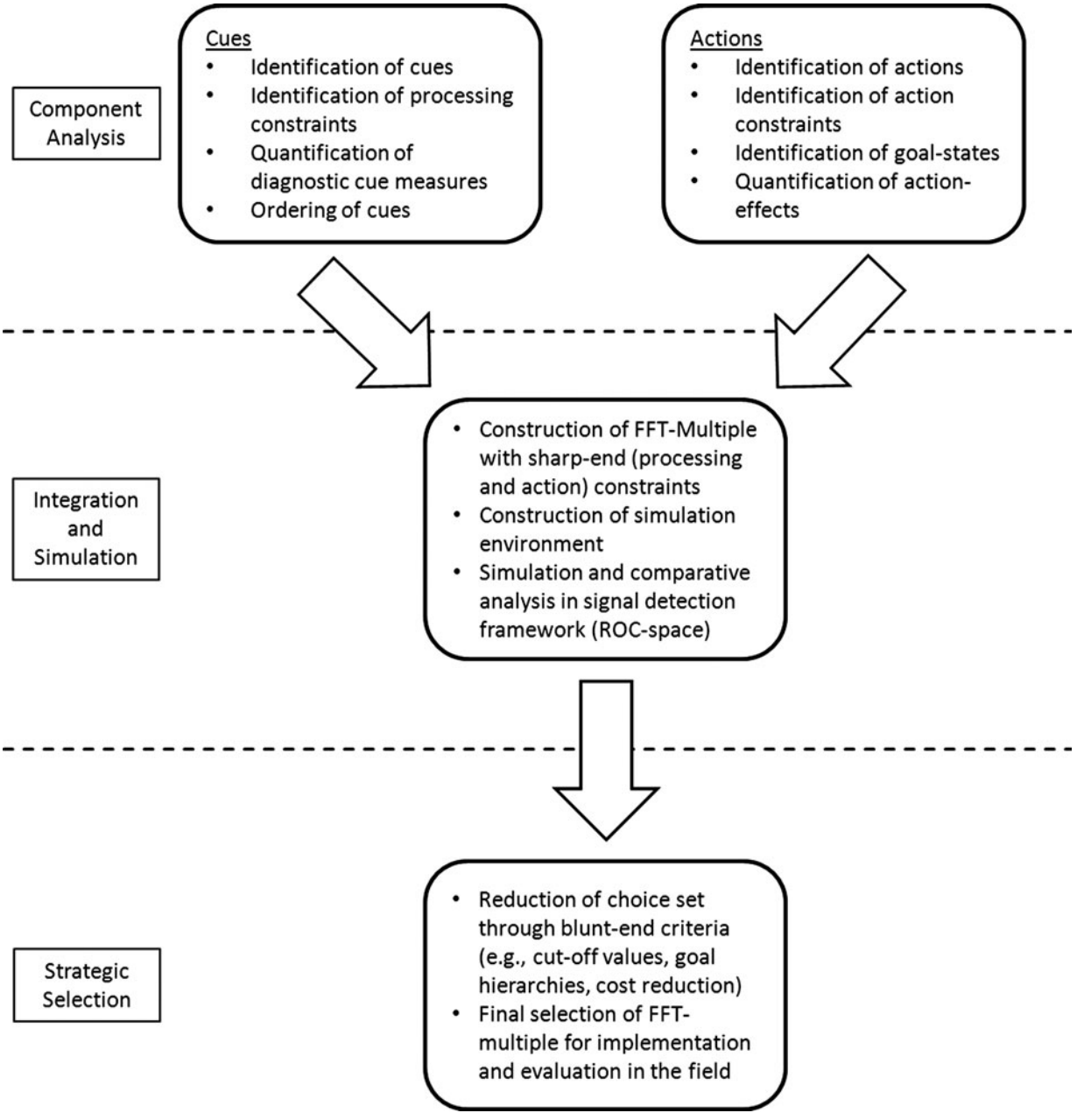

Figure 4. Consecutive phases of the methodological framework.

performance with respect to trade-offs between type I and type II errors, as well as additional identified primary endpoints. In the strategic selection phase, blunt-end performance criteria are used to constrain the number of FFTs to those most useful to the organisation. The FFT that best satisfies relevant blunt-end trade-offs *and* sharp-end constraints can then be taught to end users and evaluated in the field.

4. Example application of the methodological framework

We now illustrate the proposed stepwise methodological framework. The example describes the development and evaluation of a decision aid designed to help North Atlantic Treaty Organization (NATO) military personnel in Afghanistan predict whether an approaching vehicle is a suicide attacker (SA) or a civilian and react in line with both sharp-end requirements and blunt-end strategic goals. Initial analyses suggest that soldiers currently rely almost exclusively on whether the approaching traffic slows down or stops in response to soldiers' signalling, with soldiers escalating the force level until compliance has been achieved or the vehicle destroyed. Between January 2004 and December 2009, this strategy led to 204 civilian casualties and 7 suicide attacks, none of which could be repelled, across a total of recorded 1060 incidents (see Keller and Katsikopoulos, forthcoming).

This analysis is intended as a tutorial for practitioners wishing to use the methodological framework. It is not our aim to promote the use of specific information gathering or analytic methods or propose a particular FFT for this situation. Our choice of domain is meant to highlight the informational requirements of the methodology, the challenges in meeting these requirements in the field and to point to some of the remedies practitioners can use to deal with these challenges.

4.1 Phase 1: component analysis

4.1.1 Identification of cues and processing constraints

Goals:

(1) Identification of available cues
(2) Identification of processing constraints

Method used: Goal-directed task analysis (Endsley, Bolte, and Jones 2003; Hoffman 2005; for a general discussion on combining FFT theory with naturalistic decision-making methods, see Keller et al. 2010).

The analysis consisted of literature reviews, observations of checkpoint exercises and semi-structured interviews with German Federal Armed Forces instructors ($N = 5$). In preparation for the interviews, standard operating procedures, rules of engagement, best practice and other relevant documents were studied. The results are shown in Table 1.

Table 1. Identified cues and processing constraints for soldiers manning a NATO checkpoint for classifying oncoming traffic as hostile or not.

Available cues	Processing constraints
•Single occupant in vehicle •Vehicle approaches at high speed •Matching intelligence information •Occupants carry visible weapons •Soldiers take fire •Vehicle 'lies low', indicating heavy load	• Time constraints: as the time window for these situations is between a minute and only a few seconds, the experts suggested limiting the maximum number of cues to be identified and processed to three

Challenge: too many cues. As was the case here, it may be that the number of cues available exceeds the number of cues that can be identified and processed.

Solution: cue ranking or use of catch-all cues. The goal-directed task analysis indicated that the cues 'intelligence information', 'weapons visible' and 'taking fire' occurred very rarely, and 'lying low' was considered to be difficult to ascertain and not very diagnostic.

We decided to introduce a catch-all cue 'other threat information' to account for rare and difficult-to-ascertain cues, leaving only three dichotomised cues to be considered:

- Number of occupants ('one' or 'more than one').
- Speed of approaching vehicle ('high' or 'not high').
- Other threat information ('present' or 'not present').

Creation of a catch-all cue comes at the price of introducing subjectivity in judgment, reducing overall reliability. Alternatively, one may simply rank the cues according to a preferred performance measure and take the most diagnostic cues, ignoring the rest. As discussed in Section 2.1, if information is scarce or sample size low, limiting the number of cues considered (or ignoring other parameters such as weighting) may improve accuracy in out-of-sample prediction (see Gigerenzer and Brighton 2009; Martignon, Katsikopoulos, and Woike 2008).

Challenge: missing cues. While a task analysis is often critical in providing a framework and guiding subsequent data collections efforts (Shryane 1998), it is limited to eliciting what is already known (consciously or unconsciously) by the experts or in the literature. There is therefore no guarantee that all the diagnostic cues have been identified.

Solution: performance improvement as benchmark and use of statistical methods. Principally, it is not necessary to identify all cues as long as improved performance can be demonstrated. Also, there will be domains in which the information requirements for statistical analyses (e.g. cluster analysis and regression analysis) are met such that diagnostic cues may be identified using these methodologies.

4.1.2 *Identification of actions and action constraints*

Goals:

(1) Identification of available actions
(2) Identification of action constraints

Method used: Goal-directed task analysis.

The available actions were taken from official documents on training doctrine and rules of engagement. They are the escalation of force (EOF) sequence, a series of six actions that gradually increase in lethality. The action constraints were identified primarily during the interviews with subject-matter experts. The results are shown in Table 2.

Table 2. Available EOF actions and action constraints for soldiers manning a NATO checkpoint, where they have to react to oncoming vehicles of unknown status.

Available actions	Action constraints
•EOF1: waving/visual warning •EOF2: shouting/auditory warning •EOF3: signal flare/dazzling laser •EOF4: warning shot into air or ground •EOF5: disabling shot into motor or tyres •EOF6: lethal shot – aim to kill occupants	• Workflow (in line with current training doctrine): at least one action has to be performed; actions can only be employed in ascending EOF order; no repetition • Time constraints: maximum of four actions allowed; reduced to two if the vehicle is approaching at high speed

Challenge: missing information. When identifying action constraints, some relevant information may not be available. In this example, resource constraints involving the availability of certain pieces of equipment (EOF3: signal flare/dazzling laser) are important, but we were unable to ascertain whether such equipment is available across troops and checkpoints.

Solution: quantitative investigation during the simulation or strategic selection. A major advantage of quantitative analyses using simulations is that one can investigate the impact of the presence or absence of particular factors prior to implementation. Also, simulations can help in prioritising information gathering about such factors. In our case, we dealt with missing information about the availability of equipment in Section 4.3 after running the simulations.

4.1.3 *Quantification of diagnostic cue measures and ordering of cues*

Goals:

(1) Quantification of cue–criterion relationships and base rates
(2) Identification of constraints that impact the ordering of cues

Methods used: Goal-directed task analysis and database analysis of 1087 after action reports.

To compute cue performance measures, we conducted interviews with subject-matter experts in a frequency format to elicit the ratio of SAs to civilian contacts given a particular cue profile (e.g. 'Of 1,000 contacts with < more than one occupant, high speed, and no other threat information > how many would be civilian/suicide attackers?'). In the same way, we also ascertained the overall frequency of a cue profile (e.g. 'How often did you encounter contacts with < more than one occupant, high speed, and no other threat information >?'). The results are shown in Table 3.

Table 3. Frequency distributions of cue profiles and their probabilities of being a SA *p*(SA) versus civilian *p*(civ) based on interviews.

Cue profile (no. of occupants, speed of vehicle, other threats)	Frequency (How often did the cue profile occur?) (%)	*p*(civ) (%)	*p*(SA) (%)
(0, 0, 0)	57.6	100	0
(0, 0, 1)	14.4	100	0
(0, 1, 0)	6.4	100	0
(0, 1, 1)	1.6	100	0
(1, 0, 0)	14.4	99.3	0.7
(1, 0, 1)	3.6	75	25
(1, 1, 0)	1.6	50	50
(1, 1, 1)	0.4	25	75

Note: Cue profiles represent the presence or absence of threat information, e.g. the cue profile (0, 1, 0) represents the situation of (more than one occupant, high speed no other threat information).

The task analysis also served to construct the coding categories for a database of 1087 after-action reports of vehicles approaching NATO positions in Afghanistan between January 2004 and December 2009 (Wikileaks 2010). Two raters coded these incidents with respect to the presence or absence of threat information; inter-rater agreement in the first round was 86.2%. The raters then discussed points of divergence and resolved most disagreements. The 27 cases for which there was no agreement were eliminated from the data-set, leaving 1060 cases. While the database did not include the true number of correct rejections and thus could not be used to compute cue performance measures, it did help confirm some of the interviewees' intuitions and ordering the cues (see below).

To order the cues, we relied on an 'informed-common-sense' approach. The interviews, existing literature and database analysis of after-action reports all indicated that cars with 'more than one occupant' had, up to that point, never been involved in a suicide attack. Combined with the high occurrence of cars with multiple occupants (estimated 80%), we placed this cue first, enhancing the 'speed' of the tree (defined as the average number of cues looked up before a categorisation can be made) – an important sharp-end aspect. We placed 'other threat information' last as it is a subjective cue, reducing its reliability. 'Speed' was consequently placed as the middle cue.

Challenge: missing data and/or low-quality data. The required frequency and base-rate data have never been systematically collected by the German Federal Armed Forces, forcing us to rely on the interview responses given by the subject matter experts. These interviewees had themselves only secondary sources available on which to base their estimates (they had all only served in Kosovo, not in Afghanistan). It is therefore unlikely that their estimates accurately reflect the true statistical distribution.

Solution: 'informed common sense', use of statistical methods or simulation. Apart from the 'informed-common-sense' approach we applied for ordering the cues, alternative solutions rely on statistical analysis or simulations. When sufficient data are available, statistical analysis may be used to calculate cue performance measures and order the cues accordingly. Even in situations of information scarcity, Katsikopoulos, Schooler, and Hertwig (2010) have shown across 19 real-world environments that cue order and direction can be inferred from small samples or even people's intuitions about causal relationships between cue and criterion. They have furthermore shown that having the correct cue directions correlates highly with a models' overall predictive accuracy and in fact makes up a lion share of its predictive power. Alternatively, the simulation phase of the framework can be used to identify regions of cue values in which performance remains relatively stable or beyond which it collapses. This, in turn, can guide targeted information-gathering efforts.

4.1.4 *Identification of goals and quantification of action characteristics*

Goals:

(1) Identification of goals across institutional levels
(2) Quantification of the impact of each action on these goals

Methods used: Goal-directed task analysis and database analysis.

The task analysis identified three primary goals (G1–3) of checkpoint operations across NATO missions. For the soldiers operating the checkpoint, there is the sharp-end goal of self-protection (G1). At the blunt end, there is the additional strategic goal of minimising civilian casualties (G2; Petraeus 2010). And there is the operational purpose of the checkpoint of stopping civilian traffic (G3). We used the information contained in the database of 1087 after-action reports to calculate the following action effects with respect to the EOF action sequence identified in Table 2 (for results see Table 4):

1. *Lethality*. Relates to 'self-protection': number of times an action killed or injured a contact/number of times it was used overall.
2. *Compliance*. Relates to 'reducing civilian casualties': number of times an action did not kill or injure a contact and induced compliant behaviour (i.e. stopping or slowing down)/number of times the action was used overall.
3. *Drive through*. Relates to 'stopping civilian traffic': number of times an action neither killed nor injured a civilian contact and did not induce compliant behaviour/number of times the action was used overall.

For instance, a warning shot (EOF4) is highly effective in inducing compliance in civilians but ineffective in eliminating a hostile threat (i.e. a SA). A lethal shot (EOF6), on the other hand, has a comparatively high lethality of 0.72. It is, thus, better suited to eliminating likely threats (i.e. a SA) than to stopping civilian traffic.

Challenge: missing data. Databases often suffer from missing data. In our case, the database did not contain information on EOF1 (waving/visual warning) and EOF2 (shouting/auditory warning), as after-action reports are written only once, EOF3 (signal flare/dazzling laser) is reached.

Solution: combining data sources and/or quantitative investigations during simulation phase. We addressed this issue by combining the database of after-action reports with estimates of action effects for EOF1 and EOF2 elicited from subject-matter experts during the task analysis. As is the case in evaluating cue performance measures, simulations may allow for the evaluation of action effects.

4.2 Phase 2: integration and simulation

With the cue and action components identified, selected, quantified and ordered, it is now possible to conduct simulations to evaluate FFT performance. Due to computational resource constraints, we first investigated the performance of the four possible FFT exit structures given the cue ordering identified in Section 4.1.3. Based on the results, we then selected one exit structure for further investigation of the mapping of action sequences to exits (see Figure 5). This initial analysis suggests that the identified exit structure yields improvements over current soldier performance.

Based on the selected FFT exit structure, we constructed FFT-multiples by adding the action sequences and action constraints identified in Table 2 and simulated their performance across about 200,000 contacts. These were generated based on the frequency of cue profiles and probabilities of SAs and civilians shown in Table 3. Within the simulation, SAs differed from civilians in that they did not show compliant behaviour. Thus, the 'compliance' of actions for SAs was set zero and the difference added to the 'drive-through' value (see Table 4). After constructing 200,000 contacts based on Table 3 and all possible mappings between the EOF sequence, action constraints and the FFT exits based on Table 2, the simulation proceeded by:

(1) Selecting an FFT.
(2) Testing that FFT on all contacts:
    (a) The cue profile of the contact determined which FFT exit was applied.
    (b) The first action within that exit was performed on the contact and the action effect (outcome) was stochastically determined based on Table 4.
    (c) If the contact was killed or injured (G1) or stopped (G2), this was recorded. If the contact drove through (G3) and it was the last action in the FFT exit, this was recorded.
    (d) If the contact drove through and there were more actions, the next action was applied to the contact.
    (e) This was repeated until there were no further actions left in the FFT exit.
(3) Repeating this process for all FFT-multiples.

Challenge: computational resource constraints. Depending on the number of identified cues and actions, the number of possible FFTs can be very large. In our case, three cues and six actions result in 15 trillion possible FFTs per exit structure. The possible substitution of missing empirical data with investigations in the simulation phase (e.g. to determine regions of performance for cue performance measures, action effects, etc.) will impose an additional computational burden.

Solution: action constraints and stepwise investigation. In our example, the imposition of action constraints drastically reduced the number of viable FFTs to ca. 187,000 per exit structure. As we did here, one may additionally simulate the FFT-binary's performance first and then select one promising decision tree for further investigation.

Challenge: artificiality of simulations. Simulations are a reduced model of the real decision environment and thus incur artificialities. Here, the simulation assumes immediate placement of a contact into an FFT exit, determined by its cue profile, without considering the differential ability to perceive these cues.

Table 4. Action effects for the six EOF action sequences.

Escalation level	Lethality	Compliance	Drive through
EOF1	0.00	0.5 (0)	0.5 (1.0)
EOF2	0.00	0.5 (0)	0.5 (1.0)
EOF3	0.01	0.39 (0)	0.6 (0.99)
EOF4	0.03	0.89 (0)	0.08 (0.97)
EOF5	0.29	0.65 (0)	0.06 (0.71)
EOF6	0.72	0.28 (0)	0.00 (0.28)

Notes: Numbers indicate action effects for civilians, those in brackets indicate values for SAs. The analysis assumed that SAs do not show compliant behaviour.

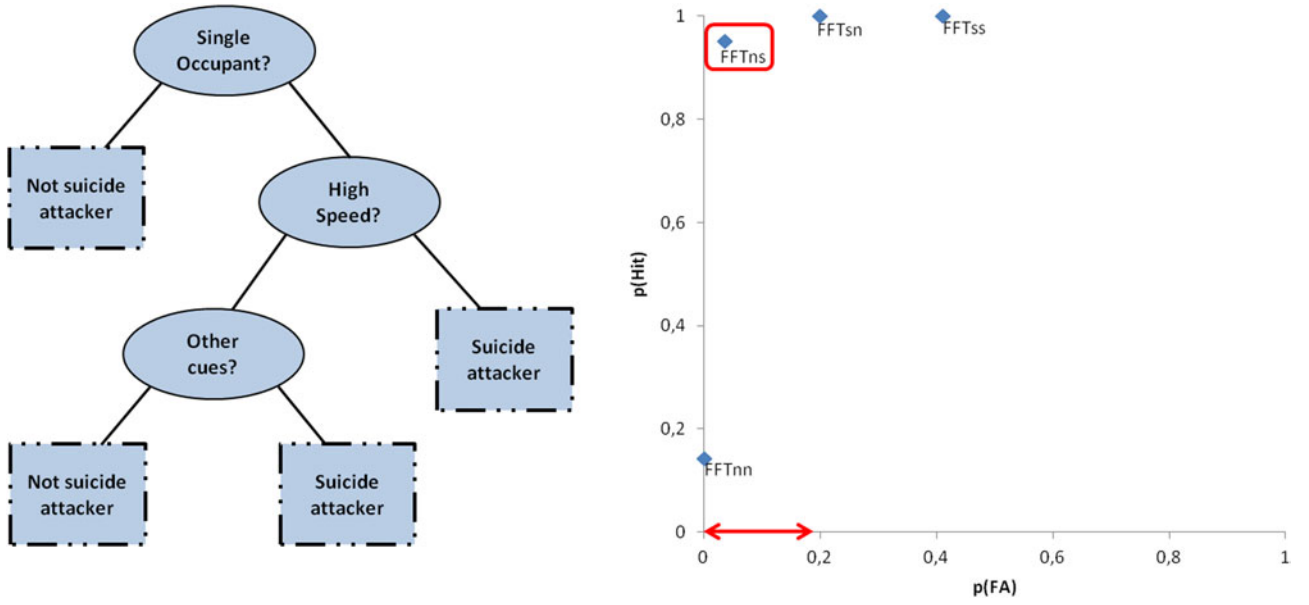

Figure 5. Left: The FFT exit structure selected for further simulation. Right: ROC curve with the performance of the chosen FFT (red circle) compared with the other FFTs (blue diamonds) and possible range of soldiers' current performance (red arrow; see Keller and Katsikopoulos, forthcoming). This range reflects the data uncertainty about the true number of correct rejections and the true effectiveness of EOF1 and EOF2 in inducing compliant behaviour.

Solution: experimental verification or task analysis. While time-critical perception may not always be an issue, if it is, perceptual aspects should be experimentally verified and either implemented in the simulation or heeded during Phase 3 of this framework. A thorough task analysis or use of other cognitive field research methods can increase the chance that the simulation reflects real-life workflows as closely as possible.

4.3 Phase 3: strategic selection

With the performance of about 187,000 FFTs across 200,000 contacts, desired performance criteria at the managerial blunt-end can be imposed for final selection. In most situations, the hit and false alarm rates will be of importance. Here, *p*(hit) and *p*(false alarm) correspond to the primary goals of 'checkpoint personnel self-protection' and 'minimisation of civilian casualties', with the goal of maximising *p*(hit) and minimising *p*(false alarm). Each red dot in Figure 6 represents one FFT-multiple.

As can be seen, performance of the FFT-multiple varies widely as a result of the impact of the action effects on hit and false alarm rates. Most importantly, the FFT-multiple shows reduced sensitivity and false alarms compared with the FFT-binary (see star in Figure 6). This effect is due to the differences in lethality of the various action patterns. The FFT-binary assumed that classification as SA results in certain elimination (i.e. lethality = 1.0). If actions are considered, however, even the highest EOF level (EOF6: lethal shot) has a lethality of only 0.72, hence the inability of FFT-multiples to attain hit rates much higher than 0.8 (the joint effects of EOF5 and EOF6). Conversely, this has a positive effect on the false alarm rate as civilians whom the FFT-binary would have mistakenly identified as SAs have a 0.2 probability of not being killed/

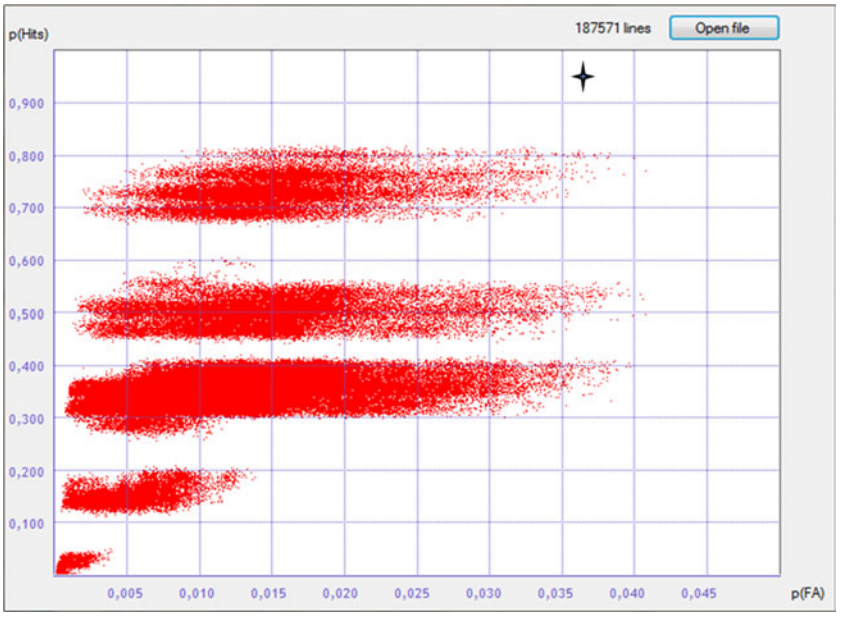

Figure 6. The performance of the 187,571 FFT-multiples in ROC space. The star in the upper left corner shows classification performance of the best performing FFT-binary for comparison. Graphical representation of the 'drive-through' dimension would have reduced transparency.

injured. Thus, most FFT-multiples have a lower false alarm rate than the FFT-binary. This analysis also shows that relying on classification performance alone can be a poor guide to performance in the field if one does not factor in the effectiveness of action effects on achieving certain goal states. Given a particular classification system, what action or action pattern the practitioner selects in response incurs performance variability, arguably in many cases greater than the performance variability that exists between different classes of classification models.

Table 5 shows the 12 top-performing FFT-multiples based on cut-off values for *p*(hit) 0.8, *p*(false alarm) 0.02, and *p* (drive through) 0.001. These values are the closest rounded values to the upper bounds of performance. We applied these cut-off values in a hierarchical fashion, in accordance with the goal hierarchy elicited during the task analysis: first, the FFTs with the highest hit rates, from this subset the FFTs with the lowest false alarm rates and finally the FFTs with the lowest drive-through rates.

Challenge: high similarity of resultant FFTs. As can be seen in Table 5, the 12 remaining FFTs are very similar in performance on the three primary goals.

Solution: reintroduction of processing and action constraints. Action constraints such as the availability of equipment that we were unable to implement earlier can now enter the selection process. For example, in the absence of data on the

Table 5. Twelve FFT-multiples remain after applying cut-off values to the three goal states.

FFT #	AP 1	AP 2	AP 3	AP 4	*p*(hit)	*p*(false alarm)	*p*(drive through)
1	1-2-4-5	5-6	1-4-5-6	1-3-4-6	0.809	0.019	0.0009
2	1-2-4-5	5-6	1-4-5-6	1-4-5-6	0.808	0.019	0.0009
3	1-2-4-5	5-6	2-4-5-6	1-4-5-6	0.808	0.018	0.0010
4	1-2-4-5	5-6	4-5-6	2-4-5-6	0.807	0.019	0.0010
5	1-2-4-5	5-6	2-4-5-6	2-4-5-6	0.806	0.019	0.0011
6	1-2-4-5	5-6	1-2-5-6	1-4-5-6	0.806	0.020	0.0011
7	1-2-4-5	5-6	4-5-6	2-3-4-6	0.806	0.020	0.0011
8	1-2-4-5	5-6	1-4-5-6	2-4-5-6	0.804	0.020	0.0009
9	1-2-4-5	5-6	3-4-5-6	1-4-5-6	0.804	0.019	0.0009
10	1-2-4-5	5-6	3-4-5-6	1-2-4-6	0.802	0.018	0.0011
11	1-2-4-5	5-6	3-4-5-6	2-3-4-6	0.801	0.019	0.0009
12	1-2-4-5	5-6	1-2-5-6	1-2-4-6	0.8	0.020	0.0010

Notes: AP 1 to 4 refer to the order of actions applied (APs) at Exits 1 through 4 (top left to bottom right) of the FFT. Numbers within the AP columns refer to Actions 1 through 6 of the EOF sequence.

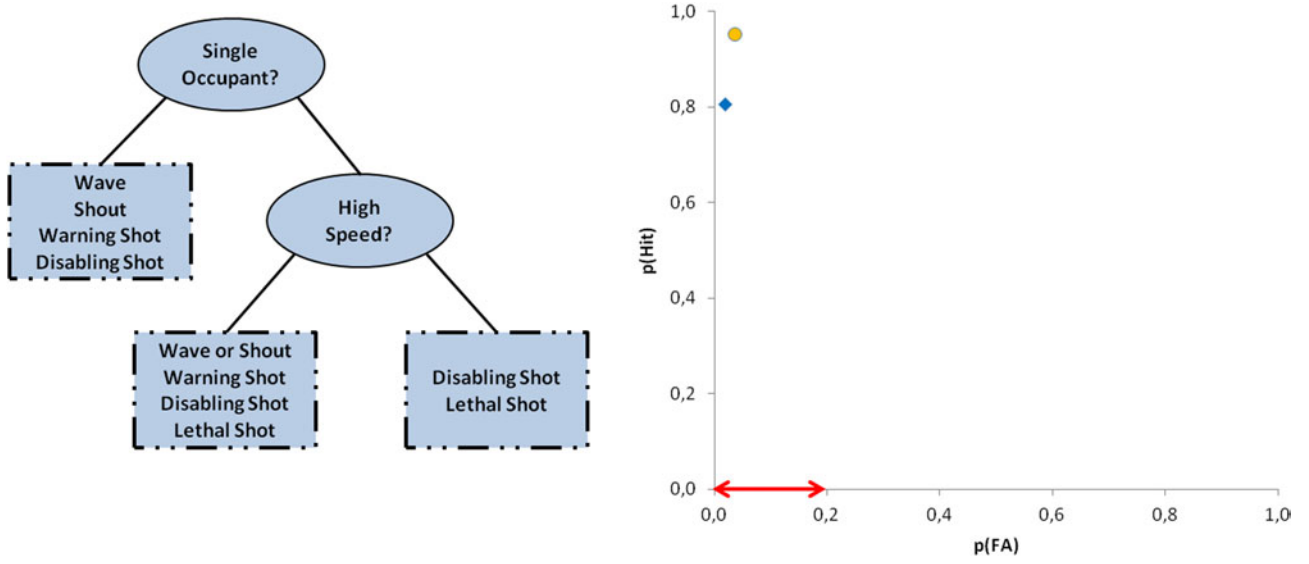

Figure 7. Visualisation of merged FFT-multiples 2 and 5 (left) after eliminating redundancies. The ROC space shows the performance of the FFT-multiple (blue diamond), the FFT-binary (cross in yellow square) and the possible range of soldiers' current performance (red arrow).

distribution of signal flares or dazzling lasers (EOF3) across NATO troops, exclusion of FFTs utilising such equipment could be a possible step. This would remove FFT Nos 1, 7, 9, 10 and 11 from the choice set. When looking at the remaining FFT Nos 2, 3, 4, 5, 6, 8 and 12, one notes that FFT Nos 2 and 5 suggest the same action sequence for the last two exits, independent of the value of the last cue. Thus, we can 'prune' these decision trees by one level without loss of performance, increasing the speed with which they can be executed. None of the other FFTs allow for similar simplification. Pruning of one or more cues further simplifies the decision aid, making it more robust to noisy samples or low sample sizes (see Geman, Bienenstock, and Doursat 1992; Gigerenzer and Brighton 2009) and more suitable for use in situations of time pressure or when the decision aid needs to be memorised to be effectively employed. Lastly, these FFTs differ on only one single action ('Wave' vs. 'Shout' as the first action in Exit 2) and can therefore be easily integrated (i.e. 'if a car has only one occupant and is approaching slowly, use either audio or visual signals to attract initial attention, whichever is more feasible'). Figure 7 shows the performance of this merged FFT.

We are now left with a simple decision aid that can be expected to yield better performance than current checkpoint procedures across goals at various levels of the institution. It provides concrete guidance on which of several courses of action to pursue. And it is easy to teach and practice and requires no special equipment.

Challenge: multiple FFTs. It may be that after applying all criteria, several FFTs remain.

Solution: final selection by practitioners or experimentation. We suggest the final choice be left with the practitioners using the aid. Depending on the quality of the data used to arrive at the FFT, it is also important that their ease of use and performance is experimentally verified and validated in a field setting.

5. Conclusion

Reason (1995) identified the failure pathway that originates at top-level decisions and proceeds via error-producing and violation-promoting conditions at the sharp end as the primary source of organisational error. One effective counter-measure is to improve the transparency and predictability of the effects of blunt-end decisions as they trickle through an organisation to the sharp end. This is precisely the goal of the presented methodological framework. It has been suggested that Human Factors and Ergonomics as a science 'is still largely atheoretical' and 'could benefit from a theoretical infusion' (Salas 2008, 352). The framework is firmly grounded in well-established theoretical approaches such as SDT and FFTs. It is widely applicable and can be used in *any* domain characterised by dichotomised cues and a one-to-many mapping between categorisation outcomes and available actions. At the same time, it is flexible with regard to the methods used to acquire the necessary information to identify and quantify the cue and action components of FFT-multiples, and to construct the simulation. The transparency gained by the ability to assess performance across levels of the institutions as well as the ability to simulate variations makes this methodology particularly attractive to decision-makers at the managerial levels of an organisation. This, in turn, will help strengthen the utility and impact of Human Factors and Ergonomics as a

profession (Dul et al. 2012). In conclusion, we believe that our framework strikes the right balance between strong foundations in theory, the methodological flexibility required for application in naturalistic settings and attractiveness to those that have the power to change organisational structures.

References

Arkes, H. R., V. A. Shaffer, and M. A. Medow. 2007. "Patients Derogate Physicians Who Use a Computer-Assisted Diagnostic Aid." *Medical Decision Making* 27: 189–202.

Dul, J., R. Bruder, P. Buckle, P. Carayon, P. Falzon, W. S. Marras, J. R. Wilson, and B. van der Doelen. 2012. "A Strategy for Human Factors/Ergonomics: Developing the Discipline and Profession." *Ergonomics* 55: 377–395.

Endsley, M. R., B. Bolte, and D. G. Jones. 2003. *Designing for Situation Awareness: An Approach to Human-Centered Design*. London: Taylor & Francis.

Geman, S., E. Bienenstock, and R. Doursat. 1992. "Neural Networks and the Bias/Variance Dilemma." *Neural Computation* 4: 1–58.

Gigerenzer, G., and H. Brighton. 2009. "Homo Heuristicus: Why Biased Minds Make Better Inferences." *Topics in Cognitive Science* 1: 107–143.

Gigerenzer, G., and C. Engel. 2006. *Heuristics and the Law*. Cambridge, MA: MIT Press.

Green, L., and D. R. Mehr. 1997. "What Alters Physicians' Decisions to Admit to the Coronary Care Unit?" *Journal of Family Practice* 45: 219–226.

Green, D. M., and J. A. Swets. 1966. *Signal Detection Theory and Psychophysics*. New York: Wiley.

Heselmans, A., S. Van de Velde, P. Donceel, B. Aertgeerts, and D. Ramaekers. 2009. "Effectiveness of Electronic Guidelines-Based Implementation Systems in Ambulatory Care Settings – A Systematic Review." *Implementation Science* 4: 82. doi:10.1186/1748-5908-4-82.

Hoffman, R. R. 2005. "Protocols for Cognitive Task Analysis." http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA475456

Hunt, D. L., B. R. Haynes, S. E. Hanna, and S. Kristina. 1998. "Effects of Computer-Based Clinical Decision Support Systems on Physician Performance and Patient Outcomes." *Journal of the American Medical Association* 280 (15): 1339–1346.

Katsikopoulos, V. K., L. L. Schooler, and R. Hertwig. 2010. "The Robust Beauty of Ordinary Information." *Psychological Review* 117: 1259–1266.

Kawamoto, K., C. A. Houlihan, E. A. Balas, and D. F. Lobach. 2005. "Improving Clinical Practice Using Clinical Decision Support Systems: A Systematic Review of Trials to Identify Features Critical to Success." *British Medical Journal* 330 (7494): 765. doi:10.1136/bmj.38398.500764.8F.

Keller, N., E. Cokely, K. V. Katsikopoulos, and O. Wegwarth. 2010. "Naturalistic Heuristics for Decision Making." *Journal of Cognitive Engineering and Decision Making* 4: 256–274.

Keller, N., and K. V. Katsikopoulos. Forthcoming. "To Shoot or Not To Shoot? Simple Heuristics for Threat Assessment and Action Taking." Submitted to *Military Operations Research*.

Luan, S., L. Schooler, and G. Gigerenzer. 2011. "A Signal Detection Analysis of Fast and Frugal Trees." *Psychological Review* 118: 316–338.

Martignon, L., K. V. Katsikopoulos, and J. K. Woike. 2008. "Categorization with Limited Resources: A Family of Simple Heuristics." *Journal of Mathematical Psychology* 52: 352–361.

Petraeus, D. 2010. "Tactical Directive 100804, Headquarters of the International Security Assistance Force, Kabul, Afghanistan." http://www.isaf.nato.int/article/isaf-releases/general-petraeus-issues-updated-tactical-directive-emphasizes-disciplined-use-of-force.html

Reason, J. 1990. *Human Error*. New York: Cambridge University Press.

Reason, J. 1995. "A Systems Approach to Organizational Error." *Ergonomics* 38: 1708–1721.

Salas, E. 2008. "At the Turn of the 21st Century: Reflections on Our Science." *Human Factors* 50: 351–353.

Shryane, N. M. 1998. "Task Analysis for the Investigation of Human Error in Safety-Critical Software Design: A Convergent Methods Approach." *Ergonomics* 41: 1719–1736.

Waterson, P. 2009. "A Critical Review of the Systems Approach Within Patient Safety Research." *Ergonomics* 52: 1185–1195.

Wikileaks. 2010. "Afghanistan War Diaries." Accessed July 29. http://wikileaks.org/wiki/Afghan_War_Diary,_2004-2010