

The development and assessment of behavioural markers to support counter-IED training.

Abstract

This article describes the method used to develop and test a checklist of behavioural markers designed to support UK military forces during Counter-Improvised Explosive Device (C-IED) training. IEDs represent a significant threat to UK and allied forces. Effective C-IED procedures and techniques are central to reducing risk to life in this safety critical role. Behavioural markers have been developed to characterise and assess non-technical skills which have been shown to be important in maintaining high performance in other safety critical domains.

The aims of this study were twofold. Firstly to develop a method which could be used to capture and assess operationally relevant behavioural markers for use in C- IED training relating primarily to non-technical skills. Secondly, to test the user acceptance of the behavioural marker checklist during military training activities.

Through engagement with military subject matter experts, operationally relevant and observable behaviours seen in C-IED training have been identified and their links to stronger and weaker performance have been established. Using a card-sort technique, the content validity of each of the markers was assessed in addition to their detectability in an operational context. Following this assessment, a selection of the most operationally relevant and detectable behaviours were assimilated into a checklist and this checklist was tested in C-IED training activities.

The results of the study show that the method used was effective in generating and assessing the behavioural markers using military subject matter experts. The study also broadly supports the utility and user-acceptance of the use of behavioural markers during training activities.

The checklist developed using this methodology will provide those responsible for delivering instruction in C-IED techniques and procedures with a straightforward process for identifying good and poor performance with respect to non-technical skills. In addition it will provide a basis for the provision of focussed feedback to trainees during debrief.

Keywords: behavioural marker; non-technical skills; crew resource management; counter-IED; Improvised Explosive Device; military; infantry

1. Introduction

Improvised Explosive Devices (IEDs) have been a significant cause of fatalities for UK military forces during recent deployments in Afghanistan and Iraq and are likely to remain a threat to deployed personnel in future military operations. Since the invasion of Afghanistan in 2001, the International Security Assistance Force (ISAF) and Afghan ground forces have

been fighting a continuing insurgency by the Taliban and associated paramilitary groups. Although the insurgents have engaged coalition forces directly with effective use of small arms 'fire and manoeuvre', they rely heavily on asymmetric tactics including the use of IEDs, suicide attacks and ambushes to inflict casualties and influence events (Meyerle & Malkasian, 2009).

Out of the 453 UK casualties reported due to hostile action from 2001 to July 2014, 219 (48%) have been caused by IEDs. This figure does not include non-fatal casualties, of which there were 2,177 classed as Wounded in Action (Ministry of Defence, 2014). IEDs can be constructed with a low metal content to reduce the chances of detection and can be carried by vehicles, people and animals, or hidden within roads or walls. IEDs can be detonated by command wire to a hidden observer, time based or by the victim themselves using methods such as pressure plates. To counter the IED threat, the British Army set up a Counter-IED (C-IED) Task Force to study the Taliban's methods and improve infantry soldier's ability to identify IEDs using new procedures and equipment (Oliver, 2014). This has resulted in improved tools, techniques and procedures to reduce the risk to the soldier in this safety-critical role. Effective C-IED training is essential to prepare UK and allied forces in reducing risk from this significant threat.

This work describes an approach for the development of a checklist of behavioural markers to assess trainee performance during C-IED training. C-IED training includes IED threat awareness and instruction in the techniques and procedures required to improve protection against the threat. The behavioural marker checklist offers a simple, but effective method for identifying both good and poor performance through the characterisation of observable behaviour relating to non-technical skills which are critical for effective performance. The checklist offers those responsible for C-IED instruction an effective means to assess performance and to focus feedback to trainees during debrief.

Complex operational roles performed by professionals in safety critical domains demand appropriate knowledge of rules and procedures together with the required technical skills and techniques. The military domain shares similarities with these safety critical domains since highly skilled soldiers operate in complex, dynamic environments according to standard operating procedures. In order to maximise performance and reduce risk, these procedures demand effective Non-Technical Skills (NTS) such as communication, teamwork, situation awareness and leadership in addition to technical skills (Rutherford et al., 2012; Flin et al., 2008; Flin et al. 2007, Flin et al., 2003). Given these similarities, application of the same tools and techniques used in other safety critical environments, such as behavioural markers, have the potential to improve safety and reduce risk in the military domain.

Aviation recognised the importance of NTS early since several serious accidents had their roots in a lack of NTS. In these accidents a causal factor identified in subsequent accident reports was the behaviour and interaction of the crew on the flight deck as opposed to a specific skill based error. Examples of such accidents include the Tenerife airport disaster (McCreary, 1998; Weick, 1993) and Eastern Airlines Flight 401 (Chou et al., 1996). In aviation a broad range of NTS are addressed explicitly through crew-resource management (CRM) training. CRM is primarily concerned with non-technical skills and behaviours focusing on cognitive and interpersonal skills as opposed to technical 'stick and rudder' skills (Flin et al. 2002). CRM training is now mandated both at the national level (CAA, 2006) and

at the European level through the European Aviation Safety Agency (Commission Regulation (EU) No 965/2012, 2012). This reflects the importance of NTS in achieving and maintaining safety in operations in addition to effective technical skills. Indeed the concept of CRM has been explicitly applied to other safety critical domains (for example see Shields and Flin, 2013; O'Connor and Flin, 2003)

The application of more general NTS training and measurement has since been expanded to include other safety critical domains such as energy (for example Crichton and Flin, 2004; O'Connor and Flin, 2003) and medicine (for example Flin et al., 2010). Evaluation of the effect of NTS training and development has shown a positive impact on safety (Fisher et al., 2000) indicating that such non-technical skills should be given as much attention as technical skills in order to ensure high performance and safety.

Behavioural markers have been successfully used in a variety of safety critical applications to assess and improve performance. Applications include aviation, medicine, energy and the military. Aviation has been a major area in which behavioural markers have been applied. Behavioural markers to assess non-technical skills have been developed in order to assess the quality of Crew Resource Management (CRM) in the cockpit (Flin and Martin, 2001). Effective CRM is reliant on good communication and attitudes, the outputs of which can often result in specifiable behaviours (Maurino and Murray, 2010; Kanki et al., 2010). As such these behaviours can be traced from a CRM training programme, codified and measured to provide an assessment of the quality of CRM in the cockpit. Flin and Martin review a number of behavioural marker systems in use by airlines to assess pilots' CRM skills. They conclude that both UK and international airlines intend to use behavioural markers in the future to integrate the assessment of technical and non-technical skills.

The medical domain has also attracted research into behavioural markers, especially relating to non-technical skills in surgery (Mitchell et al., 2013; Shields and Flin, 2013; Mitchell et al., 2012; Yule et al., 2006; Yule et al., 2006; Fletcher et al., 2004; Carthey et al., 2003), emergency medicine and intensive care (Haerkens et al., 2012; Thomas, Sexton and Helmreich, 2004) and anaesthesia (Rutherford et al., 2012; Flin et al., 2010; Fletcher et al., 2004). These studies have shown that behavioural markers are an effective way of evaluating non-technical behaviours which relate to task performance. In the military domain, Fautua et al. (2010) successfully used behavioural markers to measure performance of US border patrol personnel as part of a larger study.

Behavioural markers are descriptions of observable behaviours of teams or individuals, not attitudes or personality traits (Flin and Martin, 2001). Effective behavioural markers are clear concepts which are described simply and relate to task performance. The behaviour can be measured as a frequency (the absence or presence of the marker) or on a scale. Simple three-point scales (for example observed, not observed, not applicable) are often used on behavioural marker checklists in order to improve the clarity of the concept and to ensure reliability between different assessors (for example see Fletcher et al., 2001).

Well-designed behavioural markers have good reliability since they are based on directly observable behaviour. Good reliability has been empirically demonstrated following the development of such checklists in aviation (Klampfer et al., 2001) and medicine (Mitchel et al., 2012; Yule et al., 2008). Well-designed behavioural markers also have strong face-validity since the behaviours identified are demonstrably related to task performance. This is

an advantage since it allows effective, specific feedback to be given in a timely manner. Behavioural markers can also be used as a feed-forward tool to provide guidance to participants regarding the behaviours they are expected to exhibit in order to perform well. The most effective checklists are short, often less than one page and since the markers are defined in domain-specific language, they are straightforward to learn and have been shown to achieve good user acceptance (Flin and Martin, 2001). Long questionnaire based methods of assessment or note taking can suffer from poor user acceptance due to the high workload required to make judgements and fill in the questionnaires in a training environment (Rowley, 2014). In addition, questionnaire based methods can also require extensive post-processing of data delaying the provision of feedback. Inter-rater reliability can also be problematic unless very comprehensive training is given to instructors.

The aims of this study were two-fold. Firstly to develop a method which could be used to capture and assess operationally relevant behavioural markers for use in C-IED training. Secondly, to test the user acceptance of the behavioural marker checklist assessment method during military training activities.

To address these aims a methodology which has been successfully used to generate and assess behavioural markers associated with C-IED task performance is reported. Two checklists have been constructed as a result of these activities: a long-form checklist consisting of 21 items for detailed coverage of behaviours during training and a short-form checklist comprised of 15 items consisting of only the most relevant behavioural markers for C-IED task performance. This short-form checklist could be adopted as an addition to a Tactical Aide-Memoire (TAM) allowing a soldier quick reference to critical behaviours. Tactical Aide Memoires (TAM) are short documents, typically a single or small number of A5-size pages, to provide useful military information concerning the most up to date experience and best practice available for personnel to use in highly proceduralised operations and training. Individual Aide Memoires are generally bound together in ruggedized ring binders to form customisable reference documents, usually carried in a pocket for easy and rapid access when required.

The checklists have been tested during C-IED training activities to assess the user acceptance. Due to the sensitive nature of C-IED operations, the specific behavioural markers identified cannot be published. However, the methodology used to develop the checklists is flexible enough to be applied to a variety of domains and this application of behavioural markers in a military context is a first in the UK.

2. Method

A structured approach was taken to the development and validation of the checklist. Three stages were conducted: item generation, item assessment and item validation. In the item generation stage general behaviours associated with C-IED were elicited from Subject-Matter Experts (SMEs) engaged in a training role. Specific behavioural markers were then developed to reflect the behaviours identified. Assessment activities with SMEs provided quantitative measures of three axiomatic properties of the markers: their detectability, their relevance to performance of C-IED tasks and the frequency of their occurrence during training. This provided data upon which to base a marker's inclusion or exclusion from the

checklists. Finally, a validation activity was conducted to explore user acceptance of the checklists and the behavioural markers' ability to discriminate stronger and weaker performance.

2.1 Item generation

The objective of the item generation phase was to generate a comprehensive list of candidate behavioural markers that could be used in subsequent assessment and validation activities.

2.1.1 Participants

Participants comprised of ten SMEs who were instructors either at the UK military Defence Explosives Munitions and Search School (DEMSS), or from the Patrol Search Awareness Instructors Course (PSAIC). All participants were engaged in C-IED training activities. Participants included a range of ranks including Officers and Non-Commissioned Officers. Sampling of participants was purposive. A sample was requested of C-IED trainers who were current instructors and had operational experience of C-IED to provide the required depth of expertise. The sample size was constrained by the current operational demands made on the C-IED training service.

2.1.2 Design

Semi-structured interviews were conducted in order to elicit from SMEs as many behaviours associated with the performance C-IED tasks as possible. Participants were encouraged to discuss behaviours freely during interviews. The interviewer facilitated the identification of behaviours by highlighting, when necessary, specific areas of interest, such as the impact of fatigue on performance. It was important to generate descriptions of behaviours which could represent the unique lexicon of the user group and engage individuals with a broad range of educational experiences. Participants were always asked to consider ways in which the behaviours that they suggested could be observed by a third party. The interview continued until no new behaviours were forthcoming from the participants. Six of the ten participants were interviewed individually and two interviews were conducted with pairs of participants due to time constraints. Two interviewers were used to conduct the interviews with participants: the first author participated in all interviews to ensure consistency in procedure. A consolidation activity was then conducted on this list of behaviours by all authors. This activity had two aims:

1. To remove duplication of items.
2. To ensure that items proposed for assessment fulfilled the requirements of a behavioural marker (i.e. behaviour that is visible to an observer).

Where markers were combined or excluded this was recorded to ensure that each emerging behavioural marker had full traceability from its originating behaviour or behaviours. Using the random number function in Microsoft Excel, the revised list of behavioural markers were assigned a pseudo-random number. Markers were then ordered by this number to reduce the likelihood of clusters of similar markers producing order effects in subsequent activities.

2.1.3 Materials

The interviewer recorded participant responses on large sheets of A1 paper in full view of the participants themselves. A number of example behavioural marker checklists from other domains (aviation and medical) were made available to the participant to demonstrate the nature of the final product. The option to make audio recordings of the interviews was not available to the research team and so participant responses were recorded as self-contained sentences on the sheets of paper. In this way, any misunderstanding could be immediately identified and addressed. A single output from participants interviewed in pairs was recorded. Participants were encouraged to make links between concepts and behaviours and this was recorded on the sheets of paper during the interviews. This method encouraged the discussion to expand further on areas of interest, and provided a record of the discussion for later analysis.

2.1.4 Procedure

Each participant was interviewed individually or in pairs of similar rank and experience in order to reduce effects from rank or experience 'gradient' between individuals. At the start of each interview the purpose of the project was explained to the participant(s) and examples of behavioural marker checklists from other domains were shown to provide context.

Interviews lasted up to one hour and were flexible, allowing exploration of specific areas of interest raised by the participants. Participants who required further encouragement were referred to five general areas as a guide to discussion. These five areas adopted descriptive terms suggested by the SMEs, which are commonly used within the military.

1. Behaviour indicating good performance;
2. Behaviour indicating poor performance;
3. Behaviours associated with fatigue;
4. Behaviours associated with effective or less effective teamwork;
5. Behaviours associated with effective or less effective attitude.

2.2 Item Assessment

2.2.1 Objective

The objective of the item assessment phase was to inform selection of the most effective behavioural markers for inclusion in the final checklists. The set of 53 markers was too large for a usable checklist and so a structured method of selecting markers for inclusion was developed. Two axiomatic dimensions were used to inform selection: the detectability of the behaviour to the C-IED trainer and the relevance of the behaviour to C-IED performance. In addition, two other measures were taken: the valence of the measure (either positive or negative) and how often the behaviour was observed in C-IED training.

2.2.2 Participants

Assessment of the candidate behavioural markers was conducted in two tranches. Sampling was purposive. Participants were requested who are required to perform C-IED operations as part of their role. Both samples consisted of a mix of ranks from Private to Captain. All participants had received pre-deployment Mission Specific Training for operations in Afghanistan including C-IED instruction. All participants were male and ages ranged from 18

to 35 years in both samples. Samples were again selected on the basis of availability given their Regiments' continuing operational commitments.

The first sample consisted of 11 personnel from 1 Royal Anglian Regiment (1ANG). 1ANG deployed to Afghanistan under 12 Mechanised Brigade in 2012 providing recent operational experience in addition to their pre-deployment training. All participants in 1ANG reported recent operational experience in C-IED operations.

The second sample consisted of 11 personnel from 1 Royal Regiment of Fusiliers (1RRF). At the time 1RRF were serving as 1 Mechanised Brigade's Battlefield Casualty Replacement capability for operations in Afghanistan. Five participants in 1RRF reported recent operational experience in C-IED operations.

2.2.3 Design

The selection of behavioural markers to include in the checklist was achieved using four card-sort activities associated with four assessment criteria. Criteria to inform the selection of items to include in the checklist are described in Table 1. Card-sorts were performed by participants in sessions lasting between 30 and 45 minutes. Valence was scored either positive or negative. All other dimensions were scored on five-point scales allowing quantitative analysis of the resulting data.

Table 1. Assessment dimensions for candidate behavioural markers

Dimension	Explanation	Purpose in Analysis
Valence	Assessment of whether the behaviour indicates positive or negative performance.	Any disagreement between participants as to whether behaviour is positive or negative may indicate poor wording or ambiguity. Behavioural markers where disagreement existed were excluded.
Detectability	Assessment of how observable the behavioural marker is to a C-IED instructor.	Items assessed as highly detectable were considered for inclusion in the checklist.
Relevance	Assessment of how relevant the behaviour is to C-IED performance.	Items assessed as highly relevant were considered for inclusion in the checklist.
Frequency	Assessment of how often behaviour is seen in C-IED training.	Since the checklist is to be used in C-IED training, there was a risk that certain behavioural markers may only be seen on operations, limiting their utility in the training context.

A design was generated which mitigates against order effects in the assessment data. Firstly, the packs of cards containing the candidate behavioural makers were presented

alternately in ascending and descending card order. Secondly, the order in which the three main card sorting activities (visibility, relevancy and frequency) were performed was balanced (the valence card sort was always conducted first to act as a familiarisation activity). Each of the twelve possible combinations of activity and pack order was written on a separate piece of paper. At the start of each interview, the SME selected a piece of paper at random which defined the order in which the interview would then proceed.

2.2.4 Materials

All candidate behavioural markers were printed onto individual, 148mm x 105mm white cards. Each behavioural marker was written in the centre of the card and the corresponding identification number was written in the top-right corner. Scales were constructed from five laminated sheets of coloured paper. The sheets of paper had a 5-point scale and anchor points marked on them. Participants were requested to place the behavioural marker cards on the sheet of paper they considered most appropriate. The question associated with the card-sort was printed on another sheet of paper of the same colour, viewable at all times by the participant. An example layout is shown in Figure 1.

(Figure 1 about here)

Figure 1. Example of card-sort layout presented to participants

Care was applied to the selection of the scale anchors and questions for the three dimensions which employed five-point scales. Table 2 shows the questions relating to the dimensions and the scales used. The emphasis was on simplicity. The scales for detectability and relevance employed scale end-point anchors. The scale for training frequency employed a standard, frequency-based Likert scale.

Table 2. Questions and scales used in the card sort activities

Dimension	Question	Scale	
Detectability	“If you were looking for this behaviour, how easy is it to spot?”	1	Less easy
		2	
		3	
		4	
		5	More easy
Relevance	“How much does this behaviour matter?”	1	Matters less
		2	
		3	
		4	
		5	Matters more

Frequency	“How often do you see this behaviour during training?”	1	Never
		2	Rarely
		3	Occasionally
		4	Frequently
		5	Very frequently

2.2.5 Selection Criteria

Success criteria for each dimension listed in Table 2 were developed in order to select a smaller pool of behavioural markers from those resulting from the consolidation activity.

Relevance

Markers which achieved an average score below the mean were excluded from consideration in the final checklist. Borderline items are included.

Detectability

Markers which achieved an average score below 1-sd from the mean were excluded from consideration in the final checklist. A less stringent criterion was used for this dimension to avoid excluding markers with high relevance which are more difficult to detect. Such markers may have high performance discriminability and this will be assessed in the validation stage of the work.

Frequency

Markers which achieve an average score below 2-sd from the mean were excluded from consideration in the final checklist. A less stringent criterion was selected for this dimension due to issues with this scale observed during the card-sorts; ranks below Corporal tended to score negative behaviours as never occurring in training. When probed, such participants explained that since the behaviours were instructed against, they would not be displayed. This is an important issue; however the explanation indicates digression from the meaning of the question.

Valence

The valence data differs from the other scales since it is dichotomous, composed of negative and positive categories coded 0 and 1 respectively. An ideal scenario is that markers would present means of 0 or 1 and standard deviations of 0, showing full agreement as to the valence of the behaviour. In reality, no marker fulfilled such stringent criteria. To address this issue, a sign-test was conducted for each marker using a symmetrical binomial distribution for the hypothesis test. A one-tailed hypothesis-test was conducted for each marker using $p(1)$ or $p(0)$ as the variable of interest depending on the hypothesised valence of the marker. Items which achieved a one tailed significance level of >0.95 were considered significantly positive or significantly negative. Markers which scored probabilities below 0.95 were excluded since their valence could not be agreed upon at levels above chance. Variance was found in whether behaviours were rated as positive or negative.

2.2.6 Procedure

At the start of each interview, participants were told the purpose of the project and the structure of the session. Each SME was also told not to think too deeply about each candidate behavioural marker, and to go with their 'gut feeling'. Participants were also informed that they were free to change their minds about the placement of behavioural marker cards at any time. The participant was then asked to select a piece of folded paper at random to determine the order of the following activities and the order of the initial pack of cards.

Firstly, participants were asked to sort a new pack of cards into positive and negative behaviours, placing each card in turn on either the positive or negative laminated coloured sheet. This gave participants an opportunity to practice sorting the cards and to gain familiarity with the behavioural markers written on them.

Participants were then asked the three questions described in Table 2 in the order shown on the piece of paper selected by the participant. Participants were given a new set of cards presented in the opposite order to the previous pack for each card-sort activity. The participant then worked through the cards placing each on one of the five laminated sheets representing the five graduated responses.

Finally, the participant was given a new set of cards and asked to select their two most preferred and two least preferred behavioural markers. This activity was used to inform selection of borderline items for inclusion.

2.3 User Testing of the Checklist

2.3.1 Objective

The objective of the user testing phase was to assess the external validity of the checklist its user acceptance in a training activity. Three observers used the checklist to assess two different patrols engaged in a C-IED training scenario.

2.3.2 Participants

Validation of the behavioural marker checklist was conducted during a representative training activity with participants from 1 Royal Regiment of Fusiliers (1RRF) located at Tidworth, Wiltshire, UK. The C-IED instruction was provided by the Platoon Sergeant and facilitated by the Platoon Commander (Lieutenant), who along with an experienced Corporal, made up the three observers using the checklist. Instruction was received by approximately sixteen personnel (Lance Corporals and Privates) split into two patrols. Although the majority had undergone complete Mission Specific Training in preparation for deployment to Afghanistan, new members had only recently completed basic training and had limited previous C-IED experience.

2.3.3 Materials

A typical C-IED patrol training activity was used in order to validate the checklist in a realistic setting. It was stressed that no activity should happen during the training session that did not occur normally, and no additional equipment or documentation was to be used, apart from the behavioural marker checklist. The training scenario took place on the Salisbury Plain

Training Area close to Tidworth camp. The activity consisted of an eight-man patrol progressing along a dirt track before approaching a Vulnerable Point (VP). Short of the VP, the patrol practiced C-IED drills. Once complete, the patrol approached the VP, detected a buried practice IED and carried out further drills.

The checklist consisted of a single page of A4 with each of the behaviours listed. Behaviours were separated into positive and negative. Observers were asked to record whether behaviour was observed during the training activity. A more complex rating scale using frequency of the behaviour or a qualitative judgement relating to the behaviour was intentionally avoided to ensure that the checklist was as straightforward to use as possible.

2.3.4 Procedure

The observers were briefed on how to use the checklist and given time to familiarise themselves with the checklist items. Following familiarisation, the Platoon Sergeant completed the normal pre-training briefing for the patrol members. The patrol completed the exercise which lasted approximately 30 minutes. The checklist was used to record whether the behaviours listed were detected by the observers as the exercise progressed. Trainees were then debriefed by the assessors. Following trainee debriefing, the observers were asked to comment on the usability of the checklist.

3 Results

3.1 Item Generation

Participants responded well to the interviews often commenting on how interesting the experience had been in stimulating detailed thinking about behaviours that were frequently taken for granted or tacitly assessed. The raw results from the interviews were transcribed to form a list of 120 items describing behaviours that have an effect on C-IED task performance.

Following the consolidation activity, a revised list of 53 behavioural markers was agreed by all authors. A range of NTS were represented in the 53 behavioural markers including markers which indicate positive and negative aspects of communication, situation awareness, teamwork, attitude and leadership skills.

3.2 Item Assessment

Eight behaviours were excluded following the sign-tests since their valence was not rated sufficiently consistently by the SMEs. These behaviours frequently had a context sensitive valence and as such would not be reliable indicators if included in a checklist.

The training frequency data is weakly correlated with the detectability data ($r^2=0.04$, $p<0.01$). Behaviours seen more frequently in training are associated with behaviours which are more easily detected. It may be the case that if not explicitly instructed to look for behaviours, individuals may not be able to generate an informed judgement as to whether they are seen in training or not. These issues indicate that the training frequency data should not be relied upon as a basis for excluding potentially valuable markers from the final checklist. The

significant (albeit weak) correlation also supports the decision to assign the less stringent selection criteria, rejecting items scored as less than 2-sd from the mean training frequency.

No significant correlation was found between the relevance and detectability dimensions ($r^2=0.01$, $p>0.05$). This suggests that these dimensions are orthogonal, measuring independent aspects of the behavioural marker. As with other research that relies on ordinal scales for measurement, the data are treated as interval for the purposes of analysis (Oppenheim, 1992). Specifically, levels of precision employed by the selection criteria are considered in the light of the scales from which the data originate.

As expected when using a five-point scale, scores are normally distributed. All dimensions show negative skew (Figure 2) indicating that, in general, markers sampled are relevant, detectable and seen in training.

(Figure 2 about here)

Figure 2. Distribution of scores for the relevance, detectability and training dimensions

Relevance and detectability were the key dimensions used to select markers for inclusion into the final checklist. Scores were standardised to z-scores (mean = 0, sd =1) to aid clarity when comparing markers to their selection criteria. Selection criteria were marked as reference lines on the plot (Figure 3). Jitter of 0.025 (Standardised detectability) was added to some groups of points to ensure readability of the behavioural marker IDs. This did not move any marker over the success threshold. Twenty-one items met the success criteria identified and were included in the checklist deployed in the user-trial.

(Figure 3 about here)

Figure 3. Standardised relevance and detectability scores with selection criteria boundaries for detectability and operational relevance

3.3 User Testing of the Checklist

Overall, the checklist showed validity as a training and debriefing tool. Evidence for face- and content-validity was found during the validation activities. The criterion validity and hence the predictive validity of the checklist were not fully established as part of this work. As a result, the checklist is not ready to use as a formal assessment tool since it has yet to be formally assessed when used in this way. Furthermore, it is likely that the behaviours on the checklist vary in importance in regard to performance and therefore require weightings. Until these weightings have been empirically established and validated, summing behaviours to derive a single overall score of performance is not advisable.

3.3.1 Face Validity

The validation activity indicates that the checklist has acceptable face validity since all observers agreed that items on the checklist measure aspects of C-IED performance.

3.3.2 Content Validity

Content validity is the property of the checklist whereby all important areas of C-IED performance are represented. Results of the validation activity suggest that the checklist has sufficient content validity. All aspects addressed by the C-IED instructors during the debrief of the two observed training sessions were included in the checklist indicating that the most important areas have been covered. Where possible, omissions were addressed and one new behavioural marker was proposed.

3.3.3 Criterion Validity

Criterion validity is shown when a test score relates to another measure of success. No formal measures of performance were taken during the validation scenarios. As a 'quick and dirty' measure, observers were asked to mark each patrol in the training session using a 100mm visual analogue scale anchored from worst-ever performance to best-ever performance

All observers agreed that the performance of Patrol 1 was worse than the performance of Patrol 2. The first observer judged the performance difference between the two groups to be greater than the other two observers. Comparison between observers is not possible since observers did not all observe the same aspects of the scenario. Table 3 shows the results from this activity in addition to the number of positive and negative behaviours recorded. The data shows weak evidence for criterion validity. For observer 1 more negative behaviours were associated with Patrol 1 and fewer negative behaviours are associated with Patrol 2. For observer 2, more positive behaviours are associated with Patrol 2. No variation in the number of negative and positive behaviours between patrols was recorded by observer 3, although different behaviours were recorded for each patrol. It is likely, and assumed, that different behaviours have different weightings on performance and so simple addition of the number of positive and negative behaviours is not an appropriate way in which to interpret the data from the checklist. Inter-rater reliabilities have not been assessed. As the exercise progressed, observers started to observe different aspects of the exercise. As such, behaviours marked as seen do not represent independent and repeated observations of the same part of an exercise.

Table 3 Number of positive and negative behaviours associated with each patrol with overall observer rating

Observer	Patrol 1		Patrol 2	
	Positive Behaviours Observed	Negative Behaviours Observed	Positive Behaviours Observed	Negative Behaviours Observed
Observer 1	12	8	12	4
Observer 2	5	3	8	4

Observer	9	4	9	4
3				

3.3.4 Feedback from Observers

Overall, the response to the behavioural marker checklist was varied. The checklist had simple observed/not observed check-boxes next to each behaviour. Observers stated that on many occasions they would have expected to see a positive behaviour and did not. In these cases, observers actually adjusted the checklist to support the recording of such instances. The observers made the following suggestions for improvements:

1. Include 'observed' and 'not observed' check boxes for positive behavioural markers.
2. Allow familiarisation of checklist items prior to delivery.
3. Adjust size of checklist to A5, consistent with the observer's notebook.
4. Group more similar items together on the checklist.

4. Concluding Remarks

The work presented in this article provides an empirical basis for the generation and assessment of a behavioural marker checklist. The validation activity has provided evidence to support the use of the checklist as a training aid and as a method to debrief patrols following exercises. The checklist has the potential to fast-track the expertise of those with more limited experience of C-IED assessment and to define the structure and content of an exercise debrief. Use of the checklist would allow more consistent feedback to be given to trainees between training organisations. Regular use would also generate valuable data with which to refine and improve the checklist assessment scales and general usability.

Military participants were responsive to the method employed and found the numerous judgements required straightforward to perform. These expert judgements have provided a strong basis for the selection of behavioural markers for inclusion into a useful and usable checklist. The subsequent validation exercise has provided valuable insight into the usability of the checklist overall.

Decisions as to how the behaviours are recorded remain unresolved. Version 1 of the checklist used simple single judgements: behaviour present or behaviour not present. The decision to simplify the judgement was at variance with the wider literature which tends to suggest scales measuring frequency or quality of the results (for example see Fletcher et al., 2004). However, the requirement to reduce the training or familiarisation with the checklist prior to use drove this decision. Users indicated that this did not support their requirements and some went so far as to develop their own ad-hoc rating scales. As a result, a more sophisticated scale was proposed in version 2 to accommodate the recording of both frequency and the absence of a good behaviour. Future work should retain and test more sophisticated scales which deliver the required level of granularity but are straightforward to use, needing minimal training or familiarisation prior to use.

The criterion validity of the checklist remains empirically unsupported. The validation activity provided limited evidence indicating potential criterion validity. Establishing criterion validity would allow the checklist to be considered as a formal assessment tool. Further work is

required to specify measures of performance which could be associated with changes in behaviours seen in training. An example of further work would be to employ objective measures taken on C-IED range facilities. C-IED range facilities offer high-fidelity training in C-IED skills and drills. Facilities can offer objective measures of performance such as time to detection and in addition to quantitative measures which characterises the individuals' use of equipment. These objective measures could be compared against the behaviours observed by instructors using the checklist.

Through engagement with military SMEs, key behaviours associated with effective NTS have been identified. These behaviours have been codified to form a checklist that can be used to assess performance during training. This work has demonstrated that behavioural markers are a suitable approach to the assessment of performance in a high hazard military environment. Overall, the work supports the continued use of behavioural markers to measure, train and reinforce non-technical skills in safety critical domains. As with other research in the medical (Carthey et al., 2003; Fletcher et al., 2001), aviation (Flin et al., 2003) and energy (O'Connor and Flin, 2003) domains we have found good user acceptance to the behavioural marker approach in the military domain. The behavioural markers generated for the C-IED application also correspond to NTS found across other safety critical domains including teamwork, communication and situation awareness. This finding supports the application of techniques found in other high-hazard industries to the military domain since there is clearly overlap in the NTS required for effective performance even when there are marked differences in the skills, procedures and ultimate aims of the tasks. Performance data from the checklist can be immediately communicated to trainees to enhance and reinforce the application and development of NTS in the military domain.

The development of the behavioural marker checklist represents the start of research in this area and not its completion. Further research is required to establish the criterion validity of the checklist. If criterion validity is empirically demonstrated, the checklist could be used to assess trainees more formally. Many behavioural markers have been developed which have not been included in the checklist. A limitation of the assessment method used is that a marker which discriminates performance very effectively may not meet the relevance or detectability criteria. Although this scenario is unlikely, it is possible and the scope of the work conducted did not allow user testing of all markers developed.

Our results suggest that our approach and method could be used to support other high hazard tasks undertaken by military personnel where high levels of performance are required. It is anticipated that by continuing to apply techniques already used in other high-hazard domains, we can assist the UK Armed Forces assess C-IED performance and so reduce the risk to the soldiers carrying out this difficult and hazardous task.

Acknowledgements

The authors would like to thank Steve Harmer, Hannah Blackford and Robert Hutton of BAE Systems Advanced Technology Centre and Sarah Smith, Ben Grindley, Mike Boardman and Alison Rogers at Dstl for their valuable comment when reviewing this manuscript. The contribution of the C-IED trainers at Defence Explosives, Munitions and Search School, and the Patrol Search Awareness Instructors Course, the soldiers from 1 Royal Anglian Regiment and 1 Royal Regiment of Fusiliers is gratefully acknowledged. The authors would also like to thank two anonymous reviewers for their insightful comments and suggestions to

improve the quality of the paper. This work was funded by the Centre for Defence Enterprise (DSTLX1000073776).

References

CAA, 2006 Crew Resource Management (CRM) Training Guidance for Flight Crew, CRM Instructors (CRMIs) and CRM Instructor-Examiners (CRMIEs) (Civil Aviation Authority, Gatwick) (CAP 737), available from www.caa.co.uk. Civil Aviation Authority Safety Regulation Group

Carthey, J., de Levala, M.R., Wright D.J., Farewelld, V.T., Reason J.T., 2003. Behavioural markers of surgical excellence. *Safety Science* 41, 409 – 425.

Chou, C., Madhavan, D., Funk, K., 1996 Studies of Cockpit Task Management Errors The *International Journal of Aviation Psychology* 6(4) pages 307-320

Commission Regulation (EU) No 965/2012, Official Journal of the European Union 296, 25.10.2012, p.1-148

Crichton, M.T., Flin, R., 2004. Identifying and training non-technical skills of nuclear emergency response teams *Annals of Nuclear Energy* Volume 31, Issue 12, Pages 1317–1330

Fautua, D., Schatz, S., Kobus, D., Spiker, V A., Ross, W., Johnston, J H., Nicholson, D., Reitz, E A., 2010. Border Hunter Research Technical Report (IST-TR-10-01). University of Central Florida.

Fisher, J., Phillips, E., Mather, J., 2000 Does crew resource management training work? *Air Medical Journal* Volume 19, Issue 4, Pages 137–139

Fletcher, G., Flin, R., McGeorge, P., 2001. Anaesthetists' non-technical skills: Development of behavioural marker taxonomy. Technical Report (97/SCR/1) to the Scottish Council for Postgraduate Medical and Dental Education. University of Aberdeen.

Fletcher, G., Flin, R., McGeorge, P., Glavin, R., Maran, N., 2004. Rating non-technical skills: developing a behavioural marker system for use in anaesthesia. *Cognition, Technology & Work* 6, 165 -171

Flin, R., Martin, L., 2001. Behavioural Markers for Crew Resource Management: A survey of current practice. *International Journal of Aviation Psychology* 11, 95 -118.

Flin, R., Martin L., Goeters K., Hörmann, H., Amalberti, R., Valot, C., Nijhuis, H. 2003 The development of the NOTECHS system for evaluating pilots' CRM skills. *Human Factors Aerospace Safety* 3:95–117.

Flin, R., O'Connor, P., Crichton, M., 2008. *Safety at the Sharp End. A Guide to Non-Technical Skills*. Ashgate: Aldershot

Flin, R., Patey, R., Glavin, R., Maran, N. 2010. Anaesthetists' non-technical skills. (ANTS). *British Journal of Anaesthesia* 105, 38-44

Flin, R., Yule, S., Paterson-Brown, S., Maran, N., Rowley, D. R., Youngson, G, G. 2007. Teaching surgeons about non-technical skills. *The Surgeon*, 5, 86-89.

Haerkens, M., Jenkins, D. H., van der Hoeven, J. G. 2012. Crew resource management in the ICU: the need for culture change *Annals of Intensive Care* , 2:39

Kanki, B., Helmreich, R., Anca, J., 2010. *Crew Resource Management*. Elsevier Academic Press, London.

Klampfer, B., Flin, R. Helmreich, R, Husler, R., Sexton, B., Fletcher, G., Field, P., Staender, S., Lauche, K., Dieckmann, P. & Amacher, A. 2001. Enhancing performance in high risk environments: Recommendations for the use of behavioural markers. Report from the behavioural markers workshop, Zurich, June. Berlin: Damler Benz Foundation.

McCreary, J., Pollard, M., Stevenson, K., Wilson, M. B. 1998. Human Factors: Tenerife Revisited. *Journal of Air Transportation World Wide* 3 (1) –

Meyerle, J., Malkasian, C. (2009). *Insurgent Tactics in Southern Afghanistan 2005-2008*. CNA Analysis & Solutions.

Ministry of Defence 2014, UK forces: operations in Afghanistan. Available from: <https://www.gov.uk/uk-forces-operations-in-afghanistan>. [Accessed 16 July 2014].

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., Youngson, G., 2013. Development of a behavioural marker system for scrub practitioners' non-technical skills. *Journal of Evaluation in Clinical Practice* 19, 317-323.

Mitchell, L., Flin, R., Yule, S., Mitchell, J., Coutts, K., Youngson, G. 2012. Evaluation of the Scrub Practitioners' List of Intraoperative Non-Technical Skills (SPLINTS) system. *International Journal of Nursing Studies* 49, 201-211.

Murray, P. S., Maurino, D.E., 2010. Crew Resource Management. In Wise, J.A., Hopkin, V. D., Garland, D. J., (Eds.) *Handbook of Aviation Human Factors*. Taylor & Francis: United States

O'Connor, P., Flin, R. (2003) Crew Resource Management training for offshore oil production teams *Safety Science* 41, Issue 7 Pages 591–609

Oliver, D. (2014). Countering the IED threat. Available from : <http://www.cbrneportal.com/countering-the-ied-threat/>. [Accessed 16 July 2014].

Oppenheim, A., 1992. *Questionnaire Design, Interviewing and Attitude Measurement*. Pinter Publishing Group, London.

Rowley, J., 2014 Designing and using research questionnaires. *Management Research Review* 37 (3), 308-330

Rutherford, J., Flin, R., Mitchell, L. 2012. The non-technical skills of anaesthetic assistants in the peri-operative period: a literature review. *British Journal of Anaesthesia* 109, 27-31.

Shields, W., Flin, R. 2013. Paramedics' non-technical skills: a literature review. *Emergency Medicine Journal* 30, 350-354

Thomas, E. J., Sexton J.B., Helmreich, R. L., 2004. Translating teamwork behaviours from aviation to healthcare: development of behavioural markers for neonatal resuscitation *Quality and Safety in Healthcare* 13.

Weick, K. 1993. The vulnerable system: An analysis of the Tenerife air disaster. In Roberts, K. (Ed.) *New challenges to understanding organizations* 173-198. New York: Macmillan.

Yule, S. Flin, R., Paterson-Brown, S., Maran, N., 2006. Non-technical skills for surgeons in the operating room: A review of the literature. *Surgery* 139,140 -149

Yule, S., Flin, R., Maran, N., Rowley, D. R., Youngson, G. G., Paterson-Brown, S. 2008. Surgeons' non-technical skills in the operating room: Reliability testing of the NOTSS behaviour rating system. *World Journal of Surgery* 32,548-556