Objective oriented exercise evaluation with TARCK-it

Astrid Janssen
HKV Consultants
astrid.janssen@inbox.com

Hanneke Vreugdenhil
HKV Consultants
vreugdenhil@hkv.nl

ABSTRACT

Do we fully utilize the results of disaster management exercises? Do we miss valuable feedback? Many different types of disaster management exercises, command post exercises, tabletop exercises, or serious games have a specific purpose. Generally each exercise is designed to meet its own particular exercise goals. Evaluation of the exercises is achieved in many different ways. Not always guidelines for exercise evaluation are present. Generally the exercise participants' performance is assessed by experienced staff members. The main purpose of the evaluation is to see whether the exercise goals are met. In this publication the authors suggest that a valuable source of information about the participants' performance in exercises remains often undiscovered. A new level of information can be unlocked by evaluating the exercise using a structured, analytical method. The method TARCK-it directly compares measured participant or team performance with the exercise goals.

Keywords

Disaster management, evaluation, exercise, serious gaming

INTRODUCTION

In the world of disaster management an adequate skill level for command and control professionals cannot be solely obtained and maintained by training on the job. Therefore disaster management professionals are trained using exercises. Especially complex disaster management situations are infrequent, making exercises even more important for keeping skill levels up (Sinclair, et al., 2012).

Disaster management exercises can have several functions and objectives. They can be aimed at developing or validating plans, policies, agreements, and procedures. Or they can be aimed at familiarizing players with these plans and procedures, or at clarifying roles and responsibilities or at identifying resource gaps (Dept of Homeland Security, 2013). The objectives are specific for the exercise and usually depend on the actual circumstances, like the regulatory environment, the organizational environment and technical environment.

But how do crisis management professionals profit from exercises? Already by taking part in the exercise a learning effect is observed. The exercise brings a new experience to the participants and shapes their perceptions of the emergency management process (Perry, 2004). Evaluation teams observe the participants during the exercise and the findings are discussed in immediate post exercise debriefing ('hot wash') and debriefing after a couple of days or weeks. Evaluation topics are to what extent capability targets were met, critical tasks were executed and whether plans, policies and procedures support these tasks (Dept of Homeland Security, 2013). Debriefing is done to increase self-awareness by providing participants with feedback. A debrief is guided by an experienced facilitator; participants reflect on the scenario providing an opportunity to learn from their experience (Sinclair, et al., 2012).

Short Paper – Human Centered Design and Evaluation
Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24-27
Palen, Büscher, Comes & Hughes, eds.

In the meantime research suggests that organizations could benefit more from exercises if they conduct a more sophisticated training needs analysis, training design and evaluation. When creating new scenarios and exercises it is of great importance to use the results from the evaluation as feedback (Asproth, et al., 2013). (Schaafstal, et al., 2001) argues that technological developments in observer aids and intelligent automated cognitive performance diagnoses are of added value to events based training in emergency management. However, there is a lack of performance measurement tools and assessment methodology (Sinclair, et al., 2012).

Some performance criteria might be derived from serious gaming technology (De Kleermaeker, et al., 2011) to measure the fulfillment of exercise objectives, such as effectiveness, efficiency and timeliness of decisions. This type of assessment methodology is omnipresent in aviation safety. Evaluation efforts are targeted to observe the practice of crew resource management skills and the development of reliable, valid measures for assessing a crew's or a pilot's nontechnical skills (Flinn & Martin, 2001). A method commonly used in aviation safety for assessment of crew skills is the use of behavioral markers. The term 'behavioral markers' refers to a prescribed set of behaviors indicative of some aspect of performance. Typical behaviors are listed in relation to component skills and are used for competence assessment (Flinn & Martin, 2001).

This research paper describes a structured evaluation method to evaluate operations based exercises. The method TARCK-it uses behavioral markers and offers the possibility to use results of an exercise evaluation as input for the next step in training and exercising. TARCK-it is such a method, which is derived from serious gaming technology. The method TARCK-it can be used complementary to existing evaluation methods and does not replace unstructured feedback.

THE TARCK-IT FRAMEWORK

The TARCK-it framework enables a breakdown of a general set of exercise objectives into behavioral markers.

Two sets of aspects

Two sets of aspects are used to objectify the measuring of team or individual performance. These aspects are independent of the exercise itself. By measuring behavioral markers, both the fulfillment of exercise objectives and the performance against these aspects can be evaluated.

The first set of aspects that is used is the well-known OODA-loop, developed by military strategist John Boyd for the U.S. Army Command and General Staff College. The loop describes the phases of a feedback loop that was first described in the field of combat operations at the strategic military level. The OODA-loop consists of four phases which, when applied to disaster management exercises, read:

O - Observation – Information about the situation, whether received from others or personally acquired needs to be perceived as information by the team. In terms of observable behavior this relates to sensory activities, gathering information.

O - Orientation –Information is valued by the exercise participant or team, based upon a process of many-sided implicit or explicit cross-referencing projections empathies, correlations and rejections. In terms of observable behavior this can be discussions, interpretations, looking up additional information.

D - Decision making – Decisions or hypotheses are made upon the evaluated information. In terms of observable behavior this means searching for agreement or expressing decisions that are made.

A - Acting – The hypothesis is tested or the decision is carried out. In terms of observable behavior this can be a whole variety of actions, all linked.

The second set of aspects is applicable to each of these OODA phases. These aspects refer to the quality of performance in each phase and have been defined through experience:

T - Timeliness – whether activities are completed timely enough to be successful

A - Accuracy - whether activities are completed correctly to be successful

R - Relevance - whether the activities are relevant for completing one's task

C- Completeness - whether activities are completed to a sufficient extent

K[1] - Cost effectiveness – whether the cost of activities that are carried out are in proportion to the gain

When combining these two sets of aspects, twenty different questions can be posed. Each of these questions can be applied to the (sub) goals of the exercise, leading to observable behavior from participants.

Example:
When the aspects 'observation' and timeliness' are combined, one can pose the question: is the information perceived in time? When applied to an imaginary exercise goal "Communication with citizens", this can lead to observable behaviour like: Participant has noticed the need for information by citizen group X within Y time.

Using TARCK-it in 3 steps

In the exercise preparation phase (step 1) the exercise goals have to be broken down into subgoals. The subgoals should be described in terms of observable behaviour. To facilitate this process a wizard tool 'TARCK-it' is available that guides evaluation teams through the process. TARCK-it assists the evaluation team during the preparation phase of the exercise. In the preparation phase it facilitates the process of deriving exercise sub goals, due to its structure consisting of OODA phases and TARCK aspects. Through this preparation method TARCK-it assists the evaluation team in defining and describing observable behaviour, which is direct input for the evaluation form. At the end of the preparation phase the evaluation forms are ready for distribution. The forms are sorted according to the OODA-phases. During the exercise (step 2) the behavior is being observed or not and when so, on the form a tick is placed by the observer. After the exercise (step 3) the evaluation scores are gathered and analyzed. The wizard tool TARCK-it calculates and presents scores both for individual questions and performance on main exercise goals.

Scaling

Depending on the exercise goals and the setup of the exercise it is possible to 'scale up' or 'scale down', by predefining the combinations or markers that will and will not be measuring. In the tool scoring is relative: the scores will be expressed as a percentage of the total maximum possible score. If certain aspects are not measured at all, there will be no scoring or result on this aspect. A combination of aspects that is present multiple times will have a more fine and detailed score than a combination of aspects that can only be measured once or twice.

Analyzing the results

By using the TARCK-it method, the overall score of the exercise goals can be represented per goal and per subgoal by adding up the scores and dividing these by the potential score. In addition to the overall scores, scores on specific aspects can be calculated. This helps in identifying the root or cause of problems. Just like the unstructured feedback by observers, TARCK-it points out what aspects need extra attention in future training. The difference is that the TARCK-it method is less dependent on the observers and their skills. And also it makes exercise scores comparable when looking at the scores per phase and per aspect.

Looking into scores per phase and per aspect enables the organization to enhance procedures and helps to define training needs. Independent of future exercise goals like the familiarization of a new set of procedures, the team can be trained to pay more attention to timeliness or accuracy. Or the information gathering phase may require further procedure development.

[1] K stands for 'Kosteneffectiviteit', which is the Dutch word for Cost effectiveness. In order not to duplicate the C that already covers Completeness, the researchers maintain the K in the abbreviation.

Example:
When we look back at the imaginary exercise goal "Communication with citizens", the total score for this goal may look satisfactory. For further procedure development it may be interesting to know that the need for information was noticed and understood correctly and in time. The exercise shows at the same time that although the information given to citizens was correct and complete, warnings were distributed too late and in a difficult format.

CASE STUDY: FLOOD WARNING EXERCISE

TARCK-it has been used in an operational forecasting and warning exercise of the Dutch Water Management Centre (WMCN) together with the Meteorological Institute (KNMI) in 2011 (De Kleermaeker & Arentz, 2012). In this three day exercise seven different teams worked together in two shifts per day. Team 7 (Figure 1) advises the Ministry on the actual national flood situation. The other teams produce regional situational information and forecasts on their behalf.

Goals

The main exercise goal was to test the new procedures on providing concerted information about floods on the lakes, coast and rivers. Another main goal was to test the joint advisory role to the Ministry. The warning was based on a set of realistic events. The participants were supposed to use this information for their messages. The accuracy of their forecasts was not part of the evaluation.

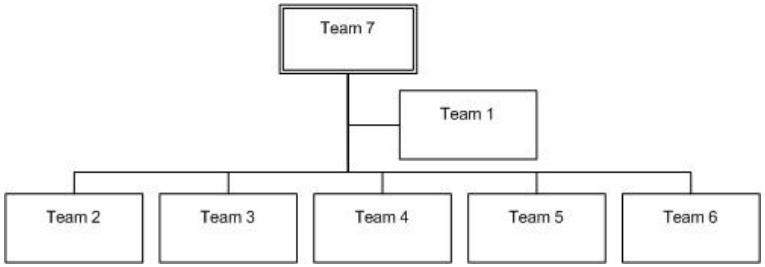

Figure 1: During the flood warning exercise many teams worked together to provide team 7 with sufficient information and data.

Apart from the main goals, the following sub-goals were defined:

- Tuning information between regional and national teams within the WMCN
- The process of scaling-up in and between the meteorological team, the flood warning teams and ministerial flood advisory team.
- The internal tuning of information between different meteorologists of the meteorological institute
- Shift handover between teams
- Dealing with peak work load

As a part of the exercise preparation per team a list of behavioral markers was deduced from the exercise goals using the TARCK-it method. Each list has been approved by the team leader. All markers are a combination of OODA phase and TARCK aspect (**Error! Reference source not found.**). The aspect 'cost effectiveness' could not be evaluated. Whether this is solely an effect of the markers reflecting the exercise goals or this is due to the complexity of translation into a marker is not researched.

	Timely	Accurate	Relevant	Complete	Cost effective	*Total*
Observe	0	0	9	8	0	*17*
Orient	0	1	20	5	0	*26*
Decide	7	13	20	16	0	*56*
Act	34	32	0	7	0	*73*
Total	***41***	***46***	***49***	***36***	***0***	***172***

Table 2: Markers that were used in the evaluation forms

During the exercise the activities of the participating teams were scored against a list of behavioral markers. The teams were observed by their direct colleagues as well as by external observers. The observers used 7 different observation forms, one for each team. All forms contained multiple pages and had a front page in which the instructions for use were given. Marker lists varied from 23 to 37 markers. The warnings and messages that were created by team 1 during the exercise enabled us to evaluate the message afterwards against the TARCK-

aspects, resulting in seven additional scoring markers.

Results

After the exercise the scoring was uploaded from the observation forms and added up. Of all 41 forms handed out, two were returned completely empty. All other forms contained entries. Of all possible entries on the filled out forms 6% of the entries were filled with "?" or "not applicable" (Table 2).

Score analysis

Score analysis showed that the direct colleagues as a group of observers scored the behavior differently from the external observers. The reason behind this seemed to be that direct colleagues made implicit assessments on aspects that the external observers did not observe explicitly.

	Forms		Markers	Scored "not applicable" or "?"
	handed out	returned	per form	
Team 1	3	2	23	4%
Team 2	6	6	37	10%
Team 3	3	2	33	11%
Team 4	5	5	30	3%
Team 5	6	6	35	5%
Team 6	6	6	29	10%
Team 7	12	12	35	3%

Table 3: Forms handed out and scored

The scores of the teams are shown in Table 3. The scores show that almost all teams paid enough attention to gathering information (observe), although two teams do not show more than half of all potential behavioral markers. Interpreting information (orient) produces a more stable score for all teams. So lack of a good quality information phase does not keep the teams from making sense of this information like they were supposed to. Decision making shows a similar pattern. Execution of the decision (Act) is in most team scored lowest, mostly around 60%, which means that about a third of all expected behavior in the acting phase was not shown by the teams. The lower score of the act-phase could be adhered to lack of quality assurance of the final product, the message. Only team 4 and team 6 show most of the expected behavior consistently.

	Observe	Orient	Decide	Act
Team 1	0%	n/a	90%	40%
Team 2	69%	59%	65%	60%
Team 3	50%	67%	55%	33%
Team 4	100%	74%	86%	62%
Team 5	36%	86%	63%	65%
Team 6	73%	93%	81%	72%
Team 7	70%	75%	71%	53%

Table 4: TARCK score in OODA phases

Value of TARCK-it in the exercise

The TARCK-it observations forms were well used by observers. The use of TARCK-it for preparation and observation resulted in more detailed evaluation information above the usually unstructured observations made by the observers. The use of TARCK-it helped the evaluation team to work more thoroughly in their approach. The team was triggered to think about what it actually wanted to see happening. This resulted in a valuable process to go back and forth between goals and observable behavior several times.

The TARCK-it results have given the preparation team practical handles to focus on improvements in the phases Observe and Act in future training.

DISCUSSION

Evaluating each exercise separately and against its specific exercise goals hampers the opportunity to combine the outcomes of exercise evaluation. Although the exercise evaluation shows what was learnt from this specific exercise, comparison with year to year results is almost impossible. TARCK-it assists in composing evaluation forms. With these evaluation aids the specific exercise goals are still measured, but the exercise can also be compared to other exercises.

The TARCK-it method guides the evaluation process in a structured way. It also enables the evaluation team to contemplate thoroughly on which explicit observable behavior is needed in crisis teams. TARCK-it cannot replace unstructured feedback and should be used in combination with other evaluation methods. Unstructured feedback can lead to gathering unexpected feedback. TARCK-it evaluation questions and results should be bound to the usual rules and regulations that apply to any other type of exercise evaluation feedback to avoid undesirable social and legal impacts.

The TARCK-it method enables organizations to compare results of different exercises with different objectives by using the two sets of aspects. This also provides long term information for competence development.

Because of the automated scoring and integrating the scores the TARCK-it method works especially well with many observers in a complex exercise, even if the observers do not know exactly the procedures that should be followed by teams or team members. The personal background of observers still influences the scoring, even though they use objective observation forms. These evaluation aids do not overcome completely the problem of the subjective observers. Proper instructions and working with different scoring scales should help to reduce these influences.

TARCK-it evaluation forms work with binary scoring: specific behaviour is observed or not. Several observers declared that this type of scoring in some situations give rise to difficulties in checking the markers, when observable behaviour shows ‘something in between’. A four point scale (without a ‘neutral’) could solve this scoring problem for the observers, but also introduces extra subjectivity and supports the tendency to mediate.

The five quality aspects (TARCK) have been based on the researchers’ experiences with crisis exercises and evaluations. While working with the system it becomes clear that for example scoring observable behaviour cost-effectiveness is difficult, although crisis managers report the relevance of these aspects in decision making. The next step in developing and using TARCK-it therefore will be to evaluate the aspects thoroughly and adjust them if needed.

ACKNOWLEDGMENTS

We would like to thank Rijkswaterstaat for enabling us to use and test TARCK-it in their exercise and to publish the results in this publication.

REFERENCES

1. Asproth, V., Borglund, E. & Öberg, L. (2013) - Exercises for crisis management training in intraorganizational settings. In: *Proceedings of the 10th International ISCRAM Conference.* Baden-Baden: ISCRAM.

2. De Kleermaeker, S. & Arentz, L.(2012) - Serious gaming in training for crisis response. In: J. R. a. Z. F. e. L. Rothkrantz, ed. *Proceedings of the 9th International ISCRAM Conference.* Vancouver, Canada: ISCRAM.

3. De Kleermaeker, S., Zijderveld, A. & Thonus, B.(2011). Training for Crisis Response with Serious Games Based on Early Warning Systems. In: *Proceedings of the 8th International ISCRAM Conference.* Lisbon: s.n.

4. Dept of Homeland Security (2013) - *Homeland Security Exercise and Evaluation Program (HSEEP).* s.l.:Govt of the USA.

5. Flinn, R. & Martin, L.(2001) - Behavioral markers for crew resource management: a review of current practice. *THE INTERNATIONAL JOURNAL OF AVIATION PSYCHOLOGY,* 11(1), pp. 95-118.

6. Perry, R. W.(2004) - Disaster exercise outcomes for professional emergency personnel and citizen volunteers. *Journal of Contingencies and Crisis Management,* June, 12(2), pp. 64-75.

7. Schaafstal, A., Johnston, J. & Oser, R.(2001) - Training teams for emergency management. *Computers in Human Behavior,* Volume 17, pp. 615-626.

8. Sinclair, H., Doyle, E., Johnston, D. & Patton, D.(2012) - Assessing emergency management training and exercises. *Disaster Prevention and Management,* 21(4), pp. 507-521.

Short Paper – Human Centered Design and Evaluation
Proceedings of the ISCRAM 2015 Conference - Kristiansand, May 24-27
Palen, Büscher, Comes & Hughes, eds.