CHAPTER 4

Measuring the effectiveness of emergency management command-and-control teams: Scale development and validation[6]

Abstract

This paper is about the development and validity testing of a context-sensitive measure of the effectiveness of multidisciplinary emergency management teams that coordinate the multidisciplinary assistance on an incident scene. The scale can assist in future research, and serve as an instrument to evaluate team effectiveness during not only actual incidents but also emergency management exercises and training programmes. After developing the scale, we validated it in a study with a field sample of 50 teams executing realistic emergency management exercises. Results indicate that the scale is internally consistent. We showed construct validity by an assessment of both convergent and discriminant validity. The scale indicates participant-external rater invariance and can be aggregated to a team score. Suggestions are offered for improving the scale, future validity testing, and practical use of the measure.

* This chapter is forthcoming as: Van der Haar, S., Segers, M., & Jehn, K.A. (*in press*). Measuring the effectiveness of emergency management teams: scale development and validation. *International Journal of Emergency Management.*

1. Introduction

Emergency management plays a crucial role in society (Comfort, 2007; McLoughlin, 1985; Zhou, Huang, & Zhang, 2010). In case of incidents such as a car crash on a motorway, a fire in a shopping centre, or a flood threatening a village, emergency management teams are responsible for controlling the situation, preventing death and damage, reaffirming people's sense of safety, and restoring the normal order of society. Therefore, as argued by Zhou, Huang, and Zhang (2010), it is of utmost importance to evaluate the effectiveness of emergency management operations with an eye to further improvement. In the current literature, no instrument for such an evaluation is mentioned that can be applied to the different settings in which this type of team operates. With this research we aim to help optimize the evaluation of the effectiveness of emergency management command-and-control teams acting either in a real-life or in a training situation. On the basis of a series of realistic emergency management exercises we developed and validated an ecologically valid instrument by which to measure the effectiveness of emergency management command-and-control teams.

2. Command and control at the scene of an incident

Although the exact organisation of emergency management differs among countries in the world, the execution and coordination of this task is often done by emergency management teams. Different assistance units (e.g. fire department, police, and medical assistance unit) cooperate at the scene of an incident. In the Netherlands, which is the setting of this study, the On Scene Command Team (OSCT) is the command-and-control team responsible for the coordination of this on-scene assistance. The common goal of the OSCT is to reduce the source of the crisis and control its consequences, while retaining work place safety and preventing errors that may cause more damage and create more victims (Baker, Day and Salas, 2006; La Porte, 1996). Together the assistance units and the OSCT are a multi-team system (MTS), defined as 'two or more teams that interface directly and interdependently in response to environmental contingencies toward the accomplishment of collective goals' (Marks, Mathieu, & Zaccaro, 2001, p. 290).

The OSCT consists of individuals with high levels of skills and abilities, who are specialized in their respective duties and who come together for the duration of an emergency situation to work interdependently towards a common valued goal (Salas, Burke, & Samman, 2001). In complex cases these teams are chaired by a team leader, facilitated by an information manager, who makes situational reports, and a spokesperson, who communicates with the media and general public, and extended to include representatives from other relevant organisations such as the government or the railway service (Helsloot, Martens, & Scholtens, 2010).

To coordinate the operation at the scene the OSCT has meetings at set times, each lasting for approximately 15 minutes. The teams use these meetings to share facts, interpretations, possible scenarios, decisions on actions, and to monitor these actions and their consequences. After each OSCT meeting the members continue coordinating their own unit at the scene. In the next section we explore the character of OSCT effectiveness as the first step in developing and validating the measurement instrument.

3. Team effectiveness

Team effectiveness is a multidimensional concept (cf. Hackman, 1987; Guzzo & Dickson, 1996; Tannenbaum, Beard, & Salas, 1992). In their review study Cohen and Bailey (1997) argue that team effectiveness has three dimensions: performance effectiveness (quantity and quality of outputs, e.g. efficiency, productivity, response times, quality, customer satisfaction, innovation), attitudinal outcomes (e.g. team satisfaction, commitment, and trust in management), and behavioural outcomes (e.g. absenteeism, turnover, and safety). These dimensions reflect the output of a team as referred to in the integrated Input-Process-Output model for team effectiveness (Tannenbaum, et al., 1992).

Cohen and Bailey (1997) argue that the importance of each of these dimensions depends on the teams' specific values and activities; the type of team is important for the team effectiveness determinants. The study by Delgado Piña, Romero Martínez, and Gómez Martinez (2007) reveals that the three dimensions of team effectiveness are not equally important for different team types. Behavioural outcomes are most important for work teams, self-managing teams and management teams. Attitudinal outcomes are most important for self-managing teams. Performance effectiveness matters for all team types. Moreover, the authors show that in many studies subjective measures of team effectiveness are used; objective measures (e.g. productivity, response time and return on capital) are applied only to work teams and management teams.

These reviews indicate that in order to define the dimensions that are important for a team's effectiveness it is important to characterize the type of team. This includes an understanding of the teams' organisational and situational characteristics, and the context that drives the difficulty, complexity and tempo of the team task (Cohen and Bailey, 1997; Kozlowski & Ilgen, 2006; Tannenbaum, et al., 1992). It also requires a characterization of the eventual team effectiveness.

We explored which team effectiveness measures were used in previous research into emergency management command-and-control teams, and found that team effectiveness is usually measured by the dimension performance effectiveness. We found different measures in the literature, including externally rated subjective overall performance effectiveness (e.g. 'Can you indicate how good you think the performance of

this team was during the crisis management simulation?', Uitdewilligen, 2011); the quality of decision making (e.g. 1 = very low, 6 = very high) in terms of the suitability of a decision as an action for dealing with each decision event (Crichton, McGeorge, & Flin, 2007); a performance measure indicating the proportion of total assets saved (McLennan, Omodei, Holgate, &Wearing, 2007); and the percentage of correctly and incorrectly chosen hypotheses during a decision-making task (Schraagen & Van de Ven, 2008).

These measures are all focused on performance effectiveness, and do not take into account behavioural and attitudinal outcomes. They vary in context specificity. The general team performance effectiveness measure as used by, for example, Uitdewilligen (2011), is not context related and reflects the subjective satisfaction of external raters rather than objective team performance effectiveness. Crichton et al.'s measure of the quality of decisions (2007) is also general in nature, but because they rated each decision on the suitability of its specific content this measure is in fact context specific. Measuring the total number of assets, as was done by McLennan et al. (2007), is very task and context specific, as is counting the percentage of correctly chosen hypotheses, and was therefore only applicable in the specific setting of that study.

The literature does not offer a measure for the team effectiveness (Cohen & Bailey, 1997) of emergency management command-and-control teams, such as the OSCT, which is valid in all different settings in which these teams operate, but the context-specific aspects of this type of teams are acknowledged. In our study we recognized this gap by developing an ecologically valid measure for command-and-control teams operating in any emergency situation.

4. Research strategy synopsis

In the next sections we will describe the development of the scale in accordance with research strategies recommended by Spector (1992) and DeVellis (2012). We executed the first step they advise, of defining the construct, and designing and reviewing an initial scale. The second step we followed is assessing the validity of the scale by administering the items to a large sample in order to obtain an internally consistent scale.

5. Developing an initial scale

5.1 Sample

Participants were required to have participated in an OSCT during actual incidents or exercises, or to have a role that gives them the opportunity to observe an OSCT closely or experience its outcomes. All 150 people participating in a Dutch network for the

professional development of emergency management exercises received an invitation to participate in this study. Thirty-two people agreed to participate. Analysis of the statements showed that with this number of participants data saturation was reached; the data that emerged repeated those we had already heard, and no new information was forthcoming. A number of 6 to 25 participants is considered to be necessary for data saturation in this kind of research (Guest, Bunce, & Johnson, 2006). Table 1 contains the description of the group of respondents.

Table 1. Description of respondents phase 1

Category	%	Description
Role	9.4	OSCT team leaders
	3.1	Team leader, other emergency management team
	18.8	Team leader, OSCT and other emergency management team
	21.9	OSCT officers
	12.5	OSCT trainers
	34.4	Other role (e.g. developer of OSCT exercises)
Gender	6.25	Women
Age		*M* = 35.55; *SD* = 7.79
Organisation	46.9	Fire department
	9.4	Police
	15.6	Medical assistance unit
	9.4	General function safety office safety region
	3.1	Water management
	15.6	Other organisation (e.g. consultancy agency)
Tenure	59.4	> 8 years
Experience	21.9	Participation OSCT actual incident > 20 times
	37.5	Participation OSCT exercise > 20 times

5.2 Data collection method

We collected data using a questionnaire either handed in person or sent by email, which was filled in by each participant in private. The questionnaire contained three open-ended questions: *1) What is the goal of an OSCT? 2) Think of an OSCT you think has performed well: On what results do you base this assessment? What did you see the team do that you approved of? 3) Think of an OSCT you think has not performed well: On what results do you base this assessment? What did you see the team do that you disapproved of?*

5.3 Analyses and results

For the analyses of the open-ended questions we used a thematic analysis approach (Guest, MacQueen, & Namey, 2012) in which we inductively identified themes emerging from the data. To this end we listed all 372 statements we found in the responses

(i.e. 'the team made clear decisions' or 'the team formulated a clear goal') and conducted an iterative content analysis to have a first rough indication of the subjects mentioned in the statements (Cope, 2004). We asked a researcher without any earlier experience with either this study or emergency management to do the same (Cope, 2004). The results included topics such as decision making, task division, leadership, cooperation and emergency response. We discussed these topics to come to an agreement about what categories of statements could be distinguished, and the relations between these categories (Cope, 2004).

We concluded that most statements reflected performance effectiveness. They did not reflect behavioural outcomes such as absenteeism and turnover, nor did they refer to attitudinal outcomes such as team member satisfaction, commitment, or trust in the team leader. Some of the statements referred to process characteristics (e.g. 'fast decision making') and others to actual outcomes (e.g. 'there were no unnecessary victims'). Furthermore, some related to the team meetings (e.g. 'team members listened to each other'), and others to the emergency response at the scene (e.g. 'the emergency management approach caused more damage'). We therefore divided the statements into four categories: team meeting process, team meeting outcome, emergency management process at the scene, and results of the emergency response at the scene (see Table 2). Since the statements describing 'team meeting process' and 'emergency management process at the scene' refer to the process instead of the outcome component in the Input-Process-Output model (Tannenbaum, et al., 1992), we excluded them from the scale.

Because the data did not show statements reflecting the behavioral and attitudinal dimensions of team effectiveness, there is some evidence that when the scale developed on the basis of these statements is used, it is actually performance effectiveness that is measured. Therefore, when this scale is used as a single method to measure performance effectiveness, the validity of the research will probably not be threatened by mono-method bias. However, to avoid common method bias, the scale should be rated by different rater groups.

Table 2. Number of statements per category after discussion and moderation

No. of items	Team meeting process	Team meeting outcome	Emergency management process at the scene	Results of the emergency response at the scene	Total
Researcher 1	207	89	56	20	372
Researcher 2	280	58	19	15	372
Categorized the same	196	41	10	11	258
Categorized the same by researchers after discussion and moderation	271 (73.4%) (2 items removed: 138 en 365)	68 (18.4%)	15 (4.1%) (1 item removed: 73)	15 (4.1%)	369 (100%) (3 items removed in total)

5.4 Scale development

In order to develop items that can be scored on a Likert scale we transformed the statements of the two categories 'team meeting outcome' and 'results of the emergency response at the scene' into items. We integrated statements in order to get a workable number of items, and ended up with a list of 68 items about the team meeting outcome, and 15 about the results of the emergency response at the scene.

Since this long list of 83 items is not feasible for research (DeVellis, 2012), and we wanted to improve the communicative validity of the eventual scale (Kvale, 1989; Sandberg, 2000), we discussed the relevance of the items with field practitioners during two workshops at a congress of Netherlands Fire Organization (the Dutch national organisation for fire departments). For the first workshop there were ten participants (four Commanding Officers of fire departments, one of them also an instructor and observer of OSCT trainings, two OSCT team leaders, and four policy makers), and for the second there were seven (four OSCT leaders, two Commanding Officers of fire departments, and one person did not state his/her profession). The participants individually checked the list of items, and marked those relevant for the measurement of the OSCT team outcome and for the results of the emergency management response at the scene. For each category they prioritized the five most important items. The participants then discussed their lists in small groups in order to come to an agreement about the five most important items for each category.

In all, 32 different items were marked as relevant. Twenty were copied directly to the final list of items. Several of the remaining 12 items were redundant (e.g. 'Decisions are clearly marked and assigned', 'The decisions are translated into an assignment for a specific/certain professional or team', and 'The assignments are given to the relevant team or person') and were therefore combined into one or two new items. We split up other items (e.g. 'Effects were prevented' was further specified as 'There are no unnecessary victims' and 'There is no unnecessary damage'). Even though the respondents marked as relevant neither the items that described the collectively developed image of the situation during team meetings (i.e., 'The image reflected the situation and its dilemma's', 'The image of the situation was unambiguous', and 'The image of the situation was realistic') , nor the items referring to the quality of actions at the scene (i.e. 'The actions on scene are justified', 'The actions on scene are adequate', 'The actions on scene are coordinated'), we decided to include these items in the final scale in any case. Both are part of the OSCT responsibility (Helsloot, et al., 2010) and relate to performance effectiveness (Cohen & Bailey, 1997).

The final OSCT performance effectiveness scale contains 21 items. Fifteen of these items refer to the 'Team meeting outcome', and thirteen items describe the 'Results of the emergency response at the scene'. On the basis of content we clustered the items into eight components (see Table 3). In the next phase of our research we tested the validity of this scale.

Table 3. OSCT team effectiveness: Initial components and factor structure

				Factor				
Initial components phase 1	**Factor structure phase 2**		Items	1	2	3	4	5
Team meeting outcome								
Image of the situation	Factor 1. Image building	1	The team managed to create a shared image of the situation in a short time.	**.63**	-.05	.01	-.02	.11
		2	The image reflected the situation and its dilemma's.	**.87**	.05	.07	.09	-.03
		3	The image of the situation was realistic.	**.78**	.07	.03	.04	.01
		4	The image of the situation was unambiguous.	**.69**	-.03	.08	-.19	.01
Decisions		5	The decisions, advices and assignments for the units are based on the actual own image and the overall image of the situation.	**.73**	.12	.01	-.19	-.09
		6	*The decisions are based on priorities.*	*.55*	*.13*	*-.08*	*-.33*	*-.00*
	Factor 2. Wrapping up the meeting	7	The decisions are translated into an assignment for a specific professional or a team.	.07	.02	.09	**-.90**	-.06
		8	The assignments are given to the relevant team or person.	.15	.04	.04	**-.81**	-.03
		9	*Decisions are taken on time.*	*.36*	*-.01*	*-.11*	*-.40*	*.26*
Information management		10	The information in the plot of the team at the end of the meetings was relevant: the development and effects of the incident; risks for the rescue services and others on scene; the approach; the people and material needed.	-.02	.10	.04	**-.39**	.13
		11	*The plot of the team at the end of the meetings was complete: an image, possible scenarios and matching frictions, decisions, actions.*	*.19*	*.06*	*-.02*	*-.11*	*.28*
Cooperation with the tactical team	Cooperation with the tactical team*	*12*	*The team takes the operational decisions, not the tactical team*	-	-	-	-	-
		13	*The team formulates a clear question to the tactical team about dilemma's and frictions.*	-	-	-	-	-
		14	*Tuning with the tactical team works well.*	-	-	-	-	-
		15	*The plot of the team is available for the tactical team and all other relevant teams.*					

Initial components phase 1	Factor structure phase 2		Items	Factor 1	2	3	4	5
Result of emergency response								
Quality of actions	Factor 3. Quality of actions	16	The coordination of each unit is adjusted to the coordination of the other units.	.02	**.53**	-.03	-.12	.06
		17	The actions on scene are justified.	-.10	**.95**	.06	-.09	-.07
		18	The actions on scene are adequate.	.12	**.91**	.09	.06	-.08
		19	The actions on scene are coordinated.	.06	**.88**	.06	-.02	-.04
Workplace safety		20	On scene safety of professionals is taken into account.	.06	**.54**	-.09	.04	.29
		21	On scene safety of civilians / companies is taken into account.	.07	**.47**	.01	.03	.27
Goal achievement	Factor 4. Goal achievement	22	The source is reduced efficiently and effectively.	.03	.08	.16	-.07	**.64**
		23	The crisis is controlled.	.07	.03	.10	-.01	**.76**
		24	There is a fast stabilization.	.05	-.01	.15	.05	**.65**
		25	Stabilization happened safely.	-.10	.21	.09	-.15	**.62**
Error rate	Factor 5. Error rate	26	There are no unnecessary victims.	.10	.02	**.81**	.02	.06
		27	There is no unnecessary damage.	-.05	.02	**.92**	-.05	-.01
		28	Based on what is happening and has happened, the media can report positively.	.07	.03	**.45**	-.05	.22

Notes. 1. This English translation is based on the Dutch original validated in this study.

2. * This component was left out from the validation study, since no tactical teams participated.

3. Factor structure: N = 224. Extraction method Exploratory Factor Analysis: Maximum Likelihood. Rotation method: Oblimin with Kaiser Normalization. Rotation converged in 14 iterations. Items in *italics* have double loadings or a loading below .3.

4. The items in *italics* were excluded from the final scale.

6. Assessment of the scale

6.1 Setting and sample

We collected data during realistic OSCT exercises (simulations) sampled from five different safety regions in the Netherlands, including 50 OSCTs with 224 team members in total (Table 4). The teams involved showed the same diversity as is usual for an OSCT (Helsloot, et al., 2010). Some teams had a team leader in meetings two and three (n = 26), some did not (n = 24). Teams differed as to number of members (three n = 20; four n = 2, five n = 15, six n = 11 and seven n = 2) with an average of five members. The teams had two (n = 22) or three (n = 28) meetings during the exercise, and participated in different exercises, of which there were nine in total.

Table 4. Description of respondents phase 2

Category	Description
Gender	18% Women
Age	M = 45; SD = 7.8
Education	64% Bachelor's degree or higher
Average experience in an OSCT	Actual incident: 13 times (*SD* = 17.1; range 0 – 150) Exercise: 15 times (*SD* = 14.5; range 0 – 75)

6.2 Task and procedure

The assignment for each team was to coordinate collectively the incident at hand, as they would in actual emergency situations. The team members filled in a demographic questionnaire before the exercise started and the newly developed team effectiveness scale after the exercise. Because there was no tactical team participating, the component about the cooperation with the tactical team was left out from the scale we used in this validation study.

6.3 Data analysis

Factor analysis. To investigate whether we could confirm the seven components underlying the performance effectiveness construct developed in phase 1, we first conducted an exploratory factor analysis (maximum likelihood, direct oblimin) (n = 224). The initial solution showed five underlying factors (Table 3), with an eigenvalue above 1 (Kaiser Criterion, in Costello and Osborne, 2005) and factor loadings above .3. Three items had a double loading or a loading below .3.

In addition, we tested three models involving four, five, and six factors (see Table 5) with a Confirmatory Factor Analysis (Table 5). Comparison of their goodness-of-fit indi-

ces confirmed that the five-factor model has the best fit (χ^2 = 387.939, p = .00, NNFI = .93, CFI = .94, GFI = .85, SRMR = .06, RMSEA = .08) with all indices falling within acceptable ranges (Tabachnick & Fidell, 2007).

Internal consistency - Item analysis and reliability. We examined the item-total correlations and coefficient alpha to test the internal consistency and reliability of the scale (DeVellis, 2012; Spector, 1992). All item-total correlations were above .47 for each subscale, showing strong relationships between the items and the scale. The Cronbach's Alpha for each subscale provided evidence for the reliability of the scale (α = .90, .82, .91, .88, .84, respectively). These results indicated the scale is internally consistent (DeVellis, 2012), and suggested no further deletions than the items already deleted because of double loadings or loadings lower than .3 in the EFA (items 6, 9, and 11, Table 3). Table 6 shows the alphas and the scale characteristics.

Table 5. Confirmatory factor analyses: Fit indices

Factors and rater group	χ^2	df	p	NNFI	CFI	GFI	SRMR	RMEA
Self-reports				N = 224				
Team meeting outcome and Results of the emergency response								
Model A: 4 factors	518.682	183	.00	.88	.90	.81	.06	.10
Model B: 5 factors	387.939	179	.00	.93	.94	.85	.06	.08
Model C: 6 factors	433.006	194	.00	.92	.93	.84	.05	.08
Shadow team members				N=82				
Team meeting outcome								
Model A: 1 factor	106.092	44	.00	.86	.89	.78	.06	.15
Model B: 2 factors	23.119	19	.23	.98	.99	.92	.04	.06
Model C: 3 factors	29.102	24	.21	.98	.99	.91	.04	.06
Response trainers				N=254				
Results of the emergency response								
Model A: 2 factor	262.532	64	.00	.88	.90	.80	.05	.15
Model B: 3 factors	150.865	62	.00	.95	.96	.87	.05	.10
Model C: 4 factors	100.766	59	.00	.97	.98	.91	.04	.07

Table 6. Reliabilities and scale statistics

		Cronbach's Alpha	N	Mean	SD
Team outcome	Image building	.90	214	5.62	.967
	Wrapping up the meeting	.82	202	5.86	.882
Results of the response	Quality of actions	.91	212	5.96	.796
	Goal achievement	.88	214	5.75	.845
	Error rate	.84	213	5.53	1.124

Convergent validity. We examined the relationship between the performance output scores and a general team effectiveness measure. We expected respondents with high scores on the one scale logically to have high scores on the other. Since the OSCT members differ in organisation, educational background, tasks and experiences we chose the multi-national team effectiveness scale developed by Gibson, Zehlmer-Bruhn, and Schwab (2003), which acknowledges differences among team members. This scale originally contained 30 items referring to the team goal, customers, timelines, quality and productivity.

We asked the 32 participants who filled in the questionnaire in the first phase of this study to mark the items they judged relevant for the OSCT (DeVellis, 2012). Analyses showed that 7 of the 30 items of the Gibson et al. (2003) scale were marked as relevant for the OSCT by 90% or more of the respondents ('This team accomplishes its objectives' [93%]; 'This team achieves its goals' [90%]; 'This team serves the purpose it is intended to serve' [100%]; 'This team wastes time' [100%]; 'This team makes mistakes' [92,6%]; 'This team needs to improve the quality of its work' [93%]; 'This team is efficient' [93%]).

The Cronbach's Alpha of this scale was .70 (N = 203, *M* = 5.29, *SD* = .783), which increased to .79 when the item 'This team needs to improve the quality of its work' was deleted. Because of this improvement we excluded this item from our analyses. Correlations between the ratings on the general team effectiveness scores and the subscales of the performance effectiveness scale were all significant: $r = .65$, $p < .01$ for 'Image building', $r = .56$, $p < .01$ for 'wrapping up the meeting', $r = .54$, $p < .01$ for 'quality of actions', $r = .53$, $p < .01$ for 'goal achievement', and $r = .42$, $p < .01$ for 'error rate'. Therefore, convergent validity is showed.

Discriminant validity. The variable 'team identification' served as a comparison construct for our scale, since it expresses a certain satisfaction with the team in terms of affective commitment which could prevent participants from objectively rating performance effectiveness. We measured team identification by using the four highest-loading items from Allen and Meyer's (1990) affective commitment scale (as did Van der Vegt and Bunderson, 2005): 'I feel emotionally attached to this team', 'I feel a strong sense of belonging to this team', 'I feel as if the team's problems are my own', and 'I feel like part of the family in this team' (N = 215, *M* = 5.71, *SD* = .912; $\alpha = .86$). A CFA of a model including the performance effectiveness measure and team identification subscales yielded acceptable goodness-of-fit indices ($\chi^2 = 599.951$, df = 265, p = .00, NNFI = .90, CFI = .92, IFI = .92, GFI = .82, SRMR = .18, RMSEA = .08), with all indices falling within acceptable ranges (Tabachnick & Fidell, 2007). This indicates discriminant validity.

Participant-external rater invariance. We assessed structural invariance of the OSCT performance effectiveness scale between team members and external raters (n = 336). The question was whether there was the same number of factors in the external rater scores as in the self-reports, and if the same items were related to the same fac-

tors. The five factors were scored by two different groups: 82 shadow team members scored the factors 'image building' and 'wrapping up the meeting' (team meeting outcome), and 254 trainers scored the factors 'quality of actions', 'goal achievement', and 'error rate' (results of the emergency response). The shadow team members observed the team by following one team member and so had a good view of the team meeting outcome. The trainers facilitated the training session by observing the actions initiated by team members on the virtual scene and repeatedly providing information about the development of the incident. Thus, the trainers had a good view of the team outcome.

Mean age of the shadow team members was 46 (*SD* = 7.4), and 10 % were women. Of the shadow team members 64% held a Bachelor's degree or higher. The average practical experience of the shadow team members was 19 incidents (*SD* = 18.9; range 0 – 150), and 15 incidents in exercises (*SD* = 13.1; range 0 – 60). The mean age of the response trainers was 43 (*SD* = 12.3), and 10 % were women. Of the response trainers 41% held a Bachelor's degree or higher.

For the ratings by the shadow team members for the two factors 'image building' and 'wrapping up the meeting', we conducted a CFA for three models with one, two, and three factors, respectively (see Table 6). Comparison of their goodness-of-fit indices confirmed that the two-factor model had the best fit (χ^2 = 23.119, p = .23, NNFI = .98, CFI = .99, GFI = .92, SRMR = .04, RMSEA = .06), with all indices falling within acceptable ranges (Tabachnick & Fidell, 2007). The model included the five-item factor 'image building' (α = .90, N = 69, M = 5.14, SD = 1.032), and the three-item factor 'wrapping up the meeting' (α = .78, N = 69, M = 5.50, SD = .862).

For the dimension 'result of the emergency response' (including three factors) as rated by response trainers we conducted a CFA for three models with two, three, and four factors, respectively (see Table 5). The best solution was the three-factor model (χ^2 = 150.865, p = .00, NNFI = .95, CFI = .96, GFI = .87, SRMR = .05, RMSEA = .10), with all indices falling within acceptable ranges and all factors having three or more items (Costello and Osborne, 2005). The model included the six-item factor 'quality of actions' (α = .94, N = 150, *M* = 5.77, *SD* = 1.001), the four-item factor 'goal achievement' (α = .96, N = 149, *M* = 5.16, *SD* = 1.177), and the three-item factor 'error rate' (α = .73, N = 148, *M* = 5.05, *SD* = 1.068).

These results show that the scale has the same number of factors across team members and external raters, which indicates dimensional variance (Gregorich, 2006) between these groups. Furthermore, each factor contained the same items in both groups, so configural invariance is supported as well (Gregorich, 2006).

Aggregation analysis. We assessed intragroup agreement (Rwg) on the scale with the multi-item formula presented by James, Demaree, and Wolf (in Lebreton & Senter, 2008) to confirm that the scores were similar enough to be aggregated into a team score. The Rwg of the subscales varied between .87 and .96 for the self-reports, and between .88 and .95 for the external raters. Thus, both team member responses and external ratings on the subscales were quite homogenous. We calculated the intraclass

correlation coefficient ICC (1) using Cohen and Doveh's (2005) formula, which is suitable for unequal group sizes. The proportion of variance in the study variables that can be explained by team membership ranged from 61% to 74% for the self-reports, and from 46% to 53% for the external ratings. These results indicate that aggregating members' scores as well as external ratings to the team level of analysis was statistically justified.

7. General discussion

Our study has provided initial evidence that the 21-item OSCT performance effectiveness scale is a statistically valid measure that can be used for OSCTs and similar teams with a command-and-control task. The scale consists of five factors. Two factors refer to the team meeting outcome. The factor 'image building' measures the image of the emergency situation developed by the team members (example: 'The team managed to create a shared image of the situation in a short time'). The factor 'wrapping up the meeting' measures the translation of actions into assignments for people or disciplines (example: 'The decisions are translated into an assignment for a specific professional or a team'). Three factors reflect the results of the response at the scene of the incident: The factor 'quality of actions' (example: 'The coordination of each unit is adjusted to the coordination of the other units'); the factor 'goal achievement' (example: 'The source is reduced efficiently and effectively'); and the factor 'error rate' (example: 'There are no unnecessary victims').

The scale is incident independent, and allows comparisons between different teams and incidents. Convergent and discriminant validity tests suggest that the scale behaves as expected because it is related to a similar construct (general team effectiveness) and distinct from the construct 'team identification', which it is not intended to measure. We also showed participant-external rater invariance and found that aggregation of individual data to the team level is justified.

7.1 Scientific contributions and implications

In addition to the review studies by Cohen and Bailey (1997) and Delgado Piña and colleagues (2007), our study has shown that for the emergency management command-and-control team performance effectiveness is the most relevant dimension of team effectiveness. Behavioural outcomes such as absenteeism and turnover did not seem to be meaningful for the OSCT, nor did attitudinal outcomes such as team member satisfaction, commitment or trust in the team leader.

Using the scale we developed for team research purposes, it is possible to make comparisons between the performances of different teams under different circumstances and discover what input factors (e.g. team member experience) and process

factors (e.g. decision-making competencies) do and do not contribute to the output following an input-process-output model (Tannenbaum, et al., 1992). For this purpose the scale needs to be filled in after task accomplishment not only by team members, in order to prevent common method bias. External observers, such as trainers, and members of teams that cooperate directly and interdependently in response to environmental contingencies should also fill in the scale.

The scale can help to diagnose best and poorly performing teams. Consequently, the value of different critical success factors extracted from case studies and reviews of emergency management research (e.g. Zhou, et al., 2011), as well as basic organisational requirements can be determined by relating them to performance effectiveness. The results of such research can help to improve protocols and training programs for emergency management teams.

7.2 Practical implications

Besides research purposes, the scale is also useful for teams when applied in actual situations. The scale can add to the evaluation of the performance effectiveness of teams operating both during emergency management exercises (simulations) and actual emergency situations. Filling in the questionnaire after participating in an emergency or emergency training as a team is an individual reflective activity for team members which can make them realize to what extent they are satisfied with the results achieved. Collectively discussing these results in the light of the chosen approach and the experienced team processes enhances team members' insights into successful cooperation and communication. If interpreted in the context of what occurred in the specific situation (Segers & Van der Haar, 2011), the scale can thus feed the dialogue between team members. Such evaluations will yield new knowledge about what works and what does not, and will therefore support learning from experiences (Segers & Van der Haar, 2011). This dialogue can be expanded to include the cooperation and communication with other teams operating in the multi-team system, if they also fill in the questionnaire and participate in the dialogue.

We suggest that in these evaluative dialogues teams are facilitated by external evaluators who have observed the team processes and scored team effectiveness. Filling in the questionnaire and collectively interpreting and discussing the results can be done both after exercises and after actual emergency situations. Since during exercises a time-out is possible, trainers can also consider creating a reflective dialogue during cooperation, in order to learn from what happened so far and create the opportunity for improvement during the next phase of the exercise. Evaluation results can be used for training purposes and to improve protocols.

7.3 Limitations and future research directions

This study has several limitations that should be addressed in future research and validity testing. First, we tested the validity of the scale in 50 teams participating in an emergency management exercise. A bigger sample size would have provided stronger evidence, and a third study, with a new sample, could have shown if our findings can be replicated.

Second, since we only included teams participating in exercises and not teams that managed actual incidents we need to investigate further the validity of the instrument for evaluation purposes in real-life settings. Using the scale as an evaluation tool during or after an operation would yield ample data as a basis for researching the instrument in an actual setting. In this way we could reveal not only its statistical but also its practical validity, which means that the users experience working with the scale as valuable. Furthermore, in future research we could investigate whether the scale can function as an intervention that improves emergency management processes and outcomes.

Third, we tested the invariance of the scale with two separate external rater groups rating different factors. Therefore it is questionable if the evidence we found for the invariance is strong enough. In future research the invariance should be tested with one external rater group scoring all factors.

Fourth, there were many statements by the respondents about the team process (271 in all), as answers to the question how they had determined that a team had good results. Therefore, it would be useful to investigate the influence of process variables and their relation with performance effectiveness. We can also use the process variable as additional support for the discriminant validity of the scale, which would meet the basic requirement that for discriminant validity ideally more than one other variable should be used (Campbell & Fiske, 1959).

8. Conclusion

This study was a first attempt to develop a methodology for measuring the team performance effectiveness of command-and-control teams in emergency management. The 21- item scale we developed seems highly relevant for future research in this domain. It can also be used to support the evaluation and optimization of the processes used by these emergency management command-and-control teams.