Int J Disaster Risk Sci (2020) 11:300–315
https://doi.org/10.1007/s13753-020-00279-6

www.ijdrs.com
www.springer.com/13753

ARTICLE

Evaluation of Fire Service Command Unit Trainings

Meinald T. Thielsch[1] · **Dzenita Hadzihalilovic**[1]

Published online: 25 May 2020

Abstract The lack of routine and training of command units and emergency managers is among the main causes of suboptimal decisions and could lead to serious consequences. To ensure optimal standards of emergency management training, specific and valid evaluation tools are needed—but are lacking. Thus, the present study's purpose is to validate instruments for the evaluation of tactical and strategic leader trainings, in particular command unit trainings, based on survey data of $n = 288$ German Command Unit members. Resulting questionnaires were named "FIRE-CU" (Feedback Instrument for Rescue forces Education – Command Unit) and "FIRE-CPX" (Feedback Instrument for Rescue forces Education – Command Post eXercise scale). Results of confirmatory factor analyses show a good fit for the postulated four-dimensional structure of process scales in the FIRE-CU (*trainer's behavior*, *structure*, *overextension*, *group*), for the two-dimensional structure of outcome scales in the FIRE-CU (*self-rated competence*, *transfer*), and for the one-dimensional structure of the FIRE-CPX. Further, strong evidence is found for reliability as well as for convergent, divergent, and concurrent validity of both the FIRE-CU and FIRE-CPX. Implications for research and practical application are also discussed to enable broad applicability in various educational programs for public security and crisis management.

Keywords Command post exercise · Command staff · Crisis management · Emergency management · Rescue forces · Vocational training evaluation

1 Introduction

Crises like the nuclear disaster in Fukushima in 2011, the terror attack at the Stade de France in Paris 2015, or the wildland fires in California in 2018 and Australia in 2019/2020 pose high risks for human lives and the environment. In these cases, emergency responders bear the heavy responsibility of managing the crises. But individual fire departments or rescue services cannot overcome these major challenges on their own, especially if there are several deployment sites at once (Heath 1998; Wybo and Kowalski 1998; Lamers 2016). Therefore, crisis management command units are required during major incidents to deal with the situation and coordinate the various emergency and rescue forces. Command units have to predict and control situational change to ensure rescue and survival of those affected.

Because crisis managers and rescue services bear the weighty responsibility of protecting human lives and the environment, personnel at this management level need to be excellently trained. Crucially, inadequate preparation and training of emergency managers are among the main causes of suboptimal decisions with sometimes serious consequences (Useem et al. 2005). Thus, in order to meet the expected standards of emergency management, high-quality trainings and valid training evaluation tools are needed. Although the existing literature provides valuable insights into the optimal design of trainings for rescue services and emergency management (Sommer and Njå 2011; Berlin and Carlström 2014; Grunnan and Fridheim 2017), to our knowledge, there is currently no validated

✉ Meinald T. Thielsch
thielsch@uni-muenster.de

[1] Department of Psychology, University of Münster, 48149 Münster, Germany

evaluation tool specifically designed for trainings of emergency management, particularly fire service command units. Thus, the aim of the present research is to adapt and validate existing scales from the evaluation of firefighter leadership education for command unit trainings.

2 Theoretical Background

Crisis management command units are confronted with major incidents or disasters, have to deal with the respective situation and coordinate the individual forces of the emergency and rescue services. While it is fortunate that major incidents requiring command units are rare, this results in command unit members having little experience with real incidents. Therefore, appropriate training and its evaluation to ensure high quality is essential. The following sections provide brief background information on command units, their training, and approaches for the corresponding training evaluation.

2.1 Command Units

The command unit or the command staff, here used as synonyms, has its origin in the military and represents a team that supports the leader or incident commander[1] (Heath 1998; Lamers 2016). Nowadays, command units are present in some civilian areas, such as the fire services. The command unit is defined as a consulting and supporting council that assists the decision-making incident commander through specific roles, structures, and information flows (Hofinger and Heimann 2016). Command units can be regarded as emergency management teams dealing with incidents that are too large for local forces such as single fire brigades or rescue services. As such, these units have to process information quickly and come to decisions in a short time (Heath 1998; Wybo and Kowalski 1998). Thus, command units consist of different groups of experts and senior emergency managers. Their work can be described as a "flexible yet robust decision environment that uses both centralised and delegatory decision processes" (Heath 1998, p. 141). In Germany, the command unit consists of higher and senior service firefighters and experienced leaders of the volunteer fire brigades who have several years of service and have completed a specific training course (Feuerwehr-Dienstvorschrift (FwDV) 2 2012; Hofinger and Heimann 2016). Unlike the lower management levels, the command unit works as rear leading support, which means it is usually not at the site of the incident (Lamers 2016). Rather, it is located in specially prepared command unit rooms or a command center assisting and coordinating the work of the formation leaders.[2] Though command units have little experience in real incidents—according to Lamers (2016), in Germany such incidents statistically occur about every 25 years—command units are preemptively deployed during scheduled major public events such as the Football World Cup. Therefore, these units have to practice in command post exercises.

2.2 Command Unit Structure and Training

The main phases of disaster and crisis management are generally known as prevention/mitigation, preparedness, response, and recovery (Heath 1998; Grunnan and Fridheim 2017). In this context, for a command unit to work efficiently it must create a specific organizational structure that allows it to deal with the emergency situation, to anticipate upcoming situational changes, and to coordinate the emergency teams on site (Heath 1998; Wybo and Kowalski 1998; Lamers 2016). In Germany, command unit members are divided into subject areas, whereby the subject areas (S) 1–4 and the incident commander are mandatory and subject areas 5 and 6 and the command unit leader are optional. The subject areas have different responsibilities, such as coordination of all personnel activities (S1), analysis of the situation (S2), action planning (S3), supply and sustenance (S4), press and media relations (S5), and information and communications systems (S6). Additionally, certain members coordinate all information sent to the command unit, and other members communicate with other organizations such as the police or technical rescue services (Lamers 2016). The German system is similar to the Incident Command System (ICS), as both have the same origin in NATO staffs (Lamers 2016). Therefore, similar principles and structures are applied: the "ICS Operations Section" is comparable to subject area S3, "Planning" is mainly found in S2 (and parts of S3), and "Logistics" is mostly covered in S1 and S4. There are differences, however, in that in Germany there is no command staff subordinate to the incident commander (their tasks are largely assigned to individual subject areas) and that the tasks of the "ICS Finance/Administration Section" are either handled by the entire staff or—in case of large-scale disasters—by a separate crisis committee dealing with all administrative measures related to the incident (Lamers 2016).

[1] The incident commander coordinates the command unit and is its representative (Karsten 2012). For definitions and German translations of fire service specific terms see Appendix A in the online supplement at https://doi.org/10.5281/zenodo.3816781.

[2] A formation leader is responsible for up to five platoons and leads and coordinates the fire service teams on site (Feuerwehr-Dienstvorschrift (FwDV) 2 (2012).

To successfully accomplish their tasks command unit members need to be very well trained. Their training consists of classical teaching elements and exercises; an excellent overview with a focus on exercises can be found at Grunnan and Fridheim (2017). In Germany, prerequisite for attending the command unit training is a successfully completed training for formation leaders (Lamers 2016).[3] During the command unit training, several teaching methods are used, such as knowledge transfer during lectures, table top exercises, teamwork tasks, and command post exercises (Hofinger and Heimann 2016).

According to the literature on command unit work, critical success factors for teams include processing, coordinating, and integrating complex information as well as developing a shared comprehension within the entire team (Wybo and Kowalski 1998; Hagemann et al. 2012; Thieme and Hofinger 2012). Taking a closer look, these attributes infer that command units must have "shared mental models" (SMM). Shared mental models are a commonly discussed model in team decision making and is defined as "knowledge structures held by members of a team that enable them to form accurate explanations and expectations for the task, and […] to coordinate their actions and adapt their behavior to demands of the task and other team members" (Cannon-Bowers et al. 2013, p. 228). To generate these shared understandings of complex situations during major incidents, the command unit members need to communicate efficiently (Hagemann, Kluge, and Ritzmann 2012), which implies that they must share information with their colleagues and process information received. Due to the significance of SMMs, generating them is a competence worth developing during command unit training.

Thieme and Hofinger (2012) emphasize that information exchange might be seen as an antecedent of SMM, which can be also found in the team learning beliefs and behaviors (TLBB) model developed by van den Bossche et al. (2006). They describe the information exchange within a team as the "team learning behavior" that promotes "mutually shared cognition." However, mutually shared cognition is defined as the mutual understanding and shared perception of a problem or task (van den Bossche et al. 2006). To generate a mutually shared cognition, team members need to "construct" (share) and "co-construct" (validate) information across the team and engage in "constructive conflicts" (discuss and amend information) about disagreements (van den Bossche et al. 2006), assuring a shared understanding of the incident and the upcoming actions. Following these definitions, SMM and mutually shared cognition describe closely related constructs. Further, the TLBB model was also confirmed for police and fire service teams by Boon et al. (2013), which suggests that the processes of construction, co-construction and constructive conflict could be applicable in fire service teams for developing mutually shared cognitions. This assumption is encouraged by Thieme and Hofinger (2012), who describe a procedure for developing and sustaining a shared mental model in command units. In this procedure, the command unit leader takes in all relevant information from each command unit member (construction) and recapitulates the gathered facts (co-construction). While the command unit leader is summarizing the information, each member is encouraged to confirm, adapt, or add to the condensed information (constructive conflict) (Thieme and Hofinger 2012).

2.3 Evaluation of Training

Evaluation is generally defined as the systematic assessment of an intervention's merit, worth, or significance (Scriven 1999). It can be checked whether the intervention is worth the resources invested, achieves its goals, or causes unintended consequences (Scriven 1999). Furthermore, evaluations that apply quantitative and qualitative research methods offer the opportunity to gather information about the appropriateness of used methods and how to improve the intervention (Kirkpatrick and Kirkpatrick 2006). Thus, the choice of evaluation criteria is crucial for evaluating the effectiveness of a training (Arthur et al. 2003). Several theories and models exist for training evaluation, but the four-level model developed by Kirkpatrick (1979) is still the most popular and commonly used model for training evaluation criteria (Salas and Cannon-Bowers 2001; Arthur et al. 2003; Grohmann and Kauffeld 2013).

2.3.1 The Four-Level Model of Training Evaluation

The four-level model of Kirkpatrick (1979) represents a hierarchical system that indicates training effectiveness through four categories of evaluation: reaction, learning, behavior, and results (Kirkpatrick 1979). The first level (level I) of evaluation represents the reaction by the participants (Kirkpatrick 1979), which can be described as the trainees' attitudinal and affective response to the training (Arthur et al. 2003; Kirkpatrick and Kirkpatrick 2006). A favorable reaction is important, because otherwise the trainees will lack learning motivation, which is the prerequisite for level II, learning (Kirkpatrick and Kirkpatrick 2006). The assessment of trainees' reaction is operationalized with self-assessment reaction sheets using standardized questions and written comments immediately

[3] For definitions and German translations of fire service specific terms see Appendix A in the online supplement at https://doi.org/10.5281/zenodo.3816781.

after the training (Kirkpatrick and Kirkpatrick 2006). Moreover, reaction questionnaires can be distinguished into affective and utility questionnaires (Blanchard and Thacker 2013), whereby affective questionnaires reflect the enjoyment or general feelings about the training, and utility questionnaires expresses the training's value (Blanchard and Thacker 2013; Ritzmann et al. 2014). Following Schulte and Thielsch (2019), the present study focuses on the utility measures, as they provide more suitable leverage points for improvements (Blanchard and Thacker 2013). Furthermore, meta-analytical evidence suggests that utility reactions have higher correlation with learning and behavior outcomes than affective reaction measures (Alliger et al. 1997).

The learning level of the model (level II) needs to be assessed to ensure the training was more than just a pleasant experience. Trainers or evaluation managers should obtain data about acquired knowledge, developed or improved skills, and changed attitudes, since these are prerequisites for change in behavior, which represents level III (Kirkpatrick and Kirkpatrick 2006). The present study uses self-assessment questionnaires for learning outcomes. Kirkpatrick and Kirkpatrick (2006) recommend an immediate examination of level II to assure a high response rate and allow conclusions about the entire amount of learned content, as a delayed assessment excludes once learned but already forgotten content (Blanchard and Thacker 2013).

The evaluation criteria for behavior, level III, ask whether a change in on-the-job behavior occurred as a result of the training's attendance (Kirkpatrick and Kirkpatrick 2006). Because applying the learned content requires some time for adjustment, the assessment of level III cannot be done promptly after the training (Kirkpatrick 1979; Blanchard and Thacker 2013). The same holds true for the level IV, namely the results. For evaluating the results, evaluation mangers want to figure out what impact occurred on an organizational level due to the training (Arthur et al. 2003; Kirkpatrick and Kirkpatrick 2006). Apparently, much more time is needed to detect effects on this macro criteria (Arthur et al. 2003). Further, level III and level IV need more sophisticated assessment methods like 360-degree performance appraisals or utility analysis estimates (Arthur et al. 2003; Kirkpatrick and Kirkpatrick 2006).

2.3.2 Process and Outcome Evaluation

A second classification of training evaluation data can be done by separating the process evaluations and the outcome evaluations (Blanchard and Thacker 2013). Learners' processes are covered on level I of the Kirkpatrick model, thus this level is essential for identifying parts of a training that might have gone wrong. One can derive benefits from comparing whether the intended training program matches up well with the implemented training program. Thereby, process data require an evaluation focus on trainer, training techniques, and learning objectives (Blanchard and Thacker 2013). Outcome measures provide important information on whether the training achieved its intended goals. When only outcome data are assessed, it is possible to judge whether the training accomplished its objectives, but it may be difficult to identify the underlying causes (Blanchard and Thacker 2013). Thus, a combination of both data types is desirable in a training evaluation.

2.4 Evaluation in the Context of Emergency Services Education

Evaluation is of prime importance in ensuring that trainings for educating emergency and rescue services personnel are of optimal quality. There are two different foci of evaluation: First, an evaluation can involve trainees' reflections on how they performed during trainings. Second, an evaluation can assess the training itself, including its structure, exercises, and instructor behavior. A typical approach in trainings is to schedule time for reports and discussion on what went well in the exercises and where participants can optimize their behavior (Berlin and Carlström 2014; Grunnan and Fridheim 2017); without such time to reflect, participants' learning can be hindered, particularly if evaluations were given long after a training was conducted (Berlin and Carlström 2014). Thus, with respect to crisis management exercises, Grunnan and Fridheim (2017, p. 80) stress that evaluative reflections are important and "should always be part of an exercise."

Aside from lack of reflection time, several other factors can impede learning in fire service trainings, such as lack of structure, inappropriate instructor behavior, unrealistic training scenarios, or difficulties in knowledge transfer (Berlin and Carlström 2014). To evaluate training quality, the four-level model of Kirkpatrick (1979) could be applied. Several general tools are available for an evaluation of team training (such as the Q4TE; Grohmann and Kauffeld 2013 or the TEI, Ritzmann et al. 2014). However, such instruments only allow for a global screening and cannot provide the trainers of crisis teams with detailed feedback on the perception of context-specific aspects and possibilities for improvement. At the moment, to the best of our knowledge, only one validated instrument in the context of firefighter education is published: The *Feedback Instrument for Rescue forces Education* (FIRE, in German: *Feedback Instrument zur Rettungskräfte Entwicklung*) (Schulte and Thielsch 2019), which is an evaluation tool for the training of group and platoon leaders.

The FIRE questionnaire was created in Germany in cooperation with the State Fire Service Institute North-

Rhine-Westphalia (in German: *Institut der Feuerwehr Nordrhein-Westfalen* (IdF NRW)), and it is based on a series of three consecutive studies: As there was an absence of published previous work on fire service training evaluation, the first study was of qualitative design, whereby the authors conducted interviews with the trainers for firefighter leaders and the trainees, ascertaining crucial attributes for excellent firefighter training (Schulte and Thielsch 2019). By consulting with professionals and topic experts, the authors were able to gather data on this topic despite missing theoretical work (Wroblewski and Leitner 2009). Furthermore, including the perspectives of trainers and trainees from the beginning increases the accuracy of self-ratings on performance (Blanchard and Thacker 2013) as well as the acceptance of the developed evaluation tool (Wroblewski and Leitner 2009). Based on this first study's 64 semistructured interviews, the authors deduced a number of factors related to excellent teaching of firefighter leaders. These factors were then used to build an initial set of evaluation questions (Schulte and Thielsch 2019). As the authors recognized several similarities between the teaching methods in fire service training and university courses (for example, lectures, group discussions, and group work with presentations) the initial questionnaire was supplemented with items from existing and validated instruments for evaluation in higher education. In the subsequent second study, the resulting initial 116 items were tested for comprehensibility, completeness, and relevance by seven trainers and 26 trainees. Afterwards (with n = 263 trainees), an exploratory factor analysis (EFA) was conducted with the remaining 65 items to reveal the underlying structure of factors (Schulte and Thielsch 2019). In a final third study (with n = 45 trainer and n = 380 trainees), the found structure was validated using confirmatory factor analysis (CFA) and associations with related scales. The resulting core version of the FIRE questionnaire for group and platoon leader trainings consists of 21 items assessing six main factors: *trainer's behavior*, *structure*, *overextension*, *group*, self-rated *competence*, and *transfer* (Schulte and Thielsch 2019).

With respect to Kirkpatrick's model, Schulte and Thielsch (2019) focused on the first two levels—reaction and learning. As each level of Kirkpatrick's model builds on the previous one (Kirkpatrick and Kirkpatrick 2006), the results for the first evaluation levels must assure (for level I) that participants had favorable reactions to the training and (for level II) that they actually learned during the training before evaluations can examine levels III or IV, the behavior or result outcomes. Furthermore, for learned content to manifest in subsequent behavior or organizational results, not only on the learning outcomes but also on the on-the-job environment is important (Arthur et al. 2003). Considering the multifactorial influences on level III and level IV as well as the accompanying challenges in assessing these levels, the current evaluation study on the fire service command unit will as well focus on level I and II. Further, by integrating Kirkpatrick's four-level model and the process and outcome model of Blanchard and Thacker (2013), the FIRE questionnaire captures process data on the reaction level and outcome data on the learning level (Schulte and Thielsch 2019), which will be continued in the present study. Specifically, the constructs *trainer's behavior*, *structure*, *overextension*, and *group* represent process data, and self-rated *competence* and *transfer* constitute outcome data.

The FIRE questionnaire has been successfully adapted and tested in the context of firefighter basic trainings at municipal and district levels (Thielsch et al. 2019). Furthermore, additional questionnaires and scales have been created for more specific evaluation purposes: For example, a questionnaire was developed to evaluate the quality of examinations of firefighters from the viewpoint of the candidates (Thielsch et al. 2018), and a four-item short questionnaire was developed to evaluate mission exercises in trainings of group and platoon leaders (Röseler et al. 2020). The latter questionnaire is based on items regarding mission exercises created in the first study of Schulte and Thielsch (2019) that were not included in the main FIRE questionnaire but were instead validated as a separate scale (Röseler et al. 2020).

2.5 Adaption of the FIRE Questionnaire for Command Unit Trainings

Considering the literature on command unit training, the applied teaching methods for training command units seem generally identical to the ones described in Schulte and Thielsch (2019) for group and platoon leader trainings. This impression was confirmed in a discussion with the deputy head of department "Crisis Management and Research" of IdF NRW in Germany. He explained that the teaching methods used in the trainings are comparable, while the contents differ due to the different audiences. Proceeding from these similarities, we decided to adapt the FIRE process scales for *trainer's behavior*, *structure*, *overextension*, and *group* by only changing the names from group/platoon leader to command unit member. This applies to the command post exercises as well, which use similar teaching methods as those used for the mission exercises of group and platoon leaders. As mentioned earlier, the command post exercises are very important for the command units. Therefore, the present study aims to validate one additional scale of the FIRE, the mission exercise scale (Röseler et al. 2020). For the questions on this scale, the term "mission exercise" was changed to "command post exercise."

With respect to outcome measures, the FIRE *transfer* scale was slightly amended in wording to better fit with the command unit training content.[4] However, contrary to group and platoon leaders, the command unit leads form the rear. Thus, major changes were necessary in the self-rated *competence* scale of the FIRE, as main key competences in the command unit are adequate processing of information, coordination, and communication (Wybo and Kowalski 1998; Hofinger and Heimann 2016). Particularly, the competence scale was amended based on the evidence presented above on team learning and shared mental models.

2.5.1 Amendment of the FIRE Competence Scale

Having considered the relevance of SMM (and its antecedent, information exchange) in command unit duties, we assume that such mental models develop during command unit trainings and conclude that they should also be explicitly taught; this, in turn means that evaluations should assess whether SMM were successfully developed during trainings. However, measuring the degree of SMM is challenging, and, so far, no consistently used methodology exists for doing so (Mohammed et al. 2010). In addition, although measuring SMM cannot be realized by a self-assessment evaluation form, it is possible to examine the acquired competence in construction, co-construction, and constructive conflict during training (van den Bossche et al. 2006; Boon et al. 2013). Therefore, based on the work of van den Bossche et al. (2006) and Boon et al. (2013), the competence dimension of the FIRE scale is extended by three items measuring construction, co-construction, and constructive conflict. In doing so, the nine items of van den Bossche et al. (2006) were condensed to three items, namely "Through my participation in the course, I learned to better communicate the information relevant to my colleagues," "My participation in the course has made it easier for me to process information received from my colleagues," and "Through my participation in the course, I am able to critically check the information provided by my colleagues for my tasks."

2.5.2 Amendment of Scale Names

Since the scale for self-rated *competence* differs from the original FIRE scale and a number of changes in item wording were conducted, the name for the tool was also amended to avoid any possibility of confusion. So, the main evaluation questionnaire for command units is titled FIRE-CU (**F**eedback **I**nstrument for **R**escue forces **E**ducation – **C**ommand **U**nit). The additional amended mission exercise scale is named FIRE-CPX (**F**eedback **I**nstrument for **R**escue forces **E**ducation – **C**ommand **P**ost e**X**ercise scale). See Table 1 for the final FIRE-CU and FIRE-CPX items to be validated in the present study.

2.6 Validation of Evaluation Instruments and Application in the Present Study

The aim of the present research is to validate the FIRE-CU and FIRE-CPX in the evaluation of trainings for the highest management level of the fire services, the command unit. Validity is usually described as the most important characteristic of psychometric tests and instruments (Clark and Watson 1995; Irwing and Hughes 2018), whereby assessing the validity of a test or instrument implies determining its accuracy and appropriateness (American Educational Research Association et al. 2014; Irwing and Hughes 2018). Validity is referred to as a unitary concept, which means that instead of there being distinct types of validity (American Educational Research Association et al. 2014), there are different types of evidence for validity that require a series of investigations (Clark and Watson 1995). For example, assessing the evidence for content validity involves determining whether the test's content relates well with the construct it is intended to measure (American Educational Research Association et al. 2014). Content validity can be ensured by developing the construct definition and the items in cooperation with experts in the field (American Educational Research Association et al. 2014), as was done in the development phase of the FIRE as described above.

Furthermore, testing the factorial structure by using confirmatory factor analysis (CFA) provides evidence of whether there is a sufficient model fit between the empirical data and the theoretically assumed model (Thompson 2004). Following the results of Schulte and Thielsch (2019) for the original FIRE questionnaire, a six-dimensional structure is also assumed for the FIRE-CU: The original FIRE process items show a four-dimensional structure reflecting *trainer's behavior*, *overextension*, *structure*, and *group*; the original FIRE outcome items show a two-dimensional structure reflecting self-rated *competence* and *transfer*. The additional module for mission exercises was assessed with a one-factor model (Röseler et al. 2020), which is also expected for the FIRE-CPX scale.

The construct validity of the FIRE-CU and FIRE-CPX questionnaires in this study are investigated with convergent, divergent, and concurrent scales (Nunnally 1978). Convergent measures were chosen based on the relatedness of fire service evaluations to teaching evaluations in higher education, as there are no other comparable specific evaluation instruments for fire service training. Therefore,

[4] See Appendix B in the online supplement at https://doi.org/10.5281/zenodo.3816781.

Table 1 Final FIRE-CU and FIRE-CPX items

Scale	Items
FIRE-CU	
Trainer's behavior	The trainers condensed difficult topics concisely
	I think the trainers gave useful feedback
	The trainers motivated me to participate actively in the course
	I think the trainers were interested in the participants' learning success
Overextension	I was overexerted by the amount of subject matter
	The speed of impartation was too high
	The course content was too difficult for me
Structure	I think the course was well structured
	I was always able to follow the structure of the course
	I think the course gave a good overview of the subject area
Group	The other trainees participated actively
	The participants supported each other
	I think there was a strong solidarity within the course
Competence	Through my participation in the course, I learned to better communicate the information relevant to my colleagues
	After the training, it is easier to make decisions in critical situations
	After this training, I know my personal limitations better than before
	After this training, I think I am more capable of staying calm in stressful situations
	My participation in the course has made it easier for me to process information received from my colleagues
	Through my participation in the course, I am able to critically check the information provided by my colleagues for my tasks
Transfer	I feel very well prepared for the next mission I will perform as a command unit member
	By participating in the exercises during the course, I gained the necessary self-assurance to perform missions as a command unit member
	I can use the acquired knowledge for my future assignment as a command unit member
FIRE-CPX	I learned a lot during the command post exercises
	The trainers provided useful feedback during the command post exercise
	During the command post exercises, I was able to apply my newly acquired knowledge
	The command post exercises' level of difficulty was appropriate

For German items see Appendix F in the online supplement at https://doi.org/10.5281/zenodo.3816781

positive relationships are expected for convergent scales. Evidence for divergent validity is assessed by a mood scale and the correlation between participants' age and their evaluations. With regard to mood, we note that mood may not only influence quality assessments in the form of a bias variable, but mood can also be the result of training quality, as learning success can also have a positive effect on the trainee's mood. In this context, small or medium correlations therefore do not argue against validity. Larger correlations, however, would call the construct validity of the scales into question. In addition, there should be little or no correlation between the age of the trainees and their assessment of training. Further evidence for construct validity can be provided with test-criterion relationships, whereby the instrument is assumed to predict a certain criterion (American Educational Research Association et al. 2014). An instrument is said to have predictive validity if it can forecast criteria measured at a later point in time, whereas it has concurrent validity by predicting criteria obtained at the same time (American Educational Research Association et al. 2014). The present study intends to assess all scales immediately after the trainings, reducing the effort for the participants and, thereby, concentrating on concurrent validity measures for FIRE-CU and FIRE-CPX. These are realized by an overall assessment of the entire training and a grade on a six-point scale. In sum, the present study focuses on testing the applicability, factorial structure, reliability, and the validity of the FIRE-CU and FIRE-CPX in command unit trainings.

3 Method

The following sections give a brief overview of sampling, measurements, and the procedure for data collection and analysis.

3.1 Sample

A total sample of $n = 294$ participants was surveyed.[5] Only participants who affirmed the informed consent were included in the analysis; 6 participants were excluded because of missing data on relevant items, resulting in a final sample of $n = 288$. In total, 277 were male (96.18%), 9 were female (3.13%), and 2 were not specified (0.69%), which resembles the usual gender ratio for the fire services in Germany. The ages ranged from 23 to 64 (Mean= 44.00, Standard deviation = 8.69). Of the participants, 30.21% were professional firefighters, 46.52% were volunteer firefighters, and 16.67% indicated that they were both professional and volunteer firefighters.

3.2 Measures

Both the adapted FIRE scales for assessing training quality and corresponding validation questionnaires were used.

3.2.1 Quality of Training

The quality of the training courses was assessed with the FIRE-CU and the FIRE-CPX scale (see Table 1). All items were scored on a 7-point scale from 1 (= strongly disagree) to 7 (= strongly agree) with an "unanswerable" option, if not specified differently.

3.2.2 Validation Scales

Well-established scales for teaching evaluation were used to underpin the convergent validity of the FIRE-CU dimensions.[6] The trainer's behavior was validated with the dimension "teaching competence" (three items, $\alpha = 0.78$), and for the overextension dimension, an eponymic scale was used (four items, $\alpha = 0.57$); both scales were obtained from the frequently used and validated inventory for teaching evaluation HILVE I (*Heidelberger Inventar zur Lehrveranstaltungs-Evaluation*, English: Heidelberg Inventory for Teaching Evaluation, Rindermann 2001). The dimensions "structure and organization" and "fellow students" (four items each) of the first HILVE edition (Rindermann and Amelang 1994) served as convergent measures for the dimensions structure and group, respectively. Convergent evidence for the competence dimension was gathered with the sub-scale "professional competence" (four items, $\alpha = 0.81$) of GEKo (*Grazer Evaluationsmodell des Kompetenzerwerbs*, English: Graz Evaluation-Model of Competence Acquisition, Paechter et al. 2011). As there was no adequate scale for convergent validity of the transfer dimension, two self-developed items were used instead: "The training prepared me well for my upcoming duties in the command unit" and "By participating in the training, the action processes during an incident became clear." The same lack of adequate scales applied to the FIRE-CPX scale, which was validated with the items "I perceived the command post exercise as very realistic" and "The command post exercise might happen in the same manner in real deployments."

Evidence for concurrent validity was gathered with five individual items, which were adapted from different validated inventories.[7] Mood was measured with a five-point smiley scale (Jäger 2004). For this scale, Jäger (2004) provided evidence for its unidimensionality and equidistance as well as high correlations with the German version of the PANAS scale ($0.75 \leq r \leq 0.89$).

3.3 Procedure

This study used a questionnaire design, and participants were asked to complete the evaluation immediately after each training, whereby the participants were assured that the evaluation was voluntary and anonymous. A pre-test ($n = 8$) preceded the main evaluation period, in which the amended FIRE items were tested for applicability during one command unit seminar. The pre-test was conducted as an online survey, where participants were asked to rate how applicable the items are to their current training. The items were scored on a 7-point scale from 1 (= strongly disagree) to 7 (= strongly agree), and an "unanswerable" option was given to indicate items that are not applicable for command units. Besides testing the items, the pre-test was also used to check the response rate in online surveys; it showed that online surveys have a very low response rate in this specific fire service context. As a result, the main evaluation was conducted as a paper-and-pencil survey. Furthermore, results of the pre-test were discussed with the head of the department and the deputy head of the department "Crisis Management and Research" of IdF NRW. Based on these pre-tests, we replaced item no. 14 "After this training, I can identify dangerous situations earlier," and some other items were also amended.[8] The main evaluation period started on 2 July 2018 and lasted for 16 weeks, during

[5] Some recommendations indicate that for satisfactory effects in CFAs with three factors and loadings of 0.6, a sample size of $n = 100$ might be sufficient (Gagne and Hancock 2006), which seemed quite small for a CFA. Therefore, the intended sample size for the present study was $n = 250$, assuring stable correlations and following the recommendations of Schönbrodt and Perugini (2013).

[6] For an overview see Appendix C in the online supplement at https://doi.org/10.5281/zenodo.3816781.

[7] For detailed description see Appendix C in the online supplement at https://doi.org/10.5281/zenodo.3816781.

[8] See Appendix B in the online supplement at https://doi.org/10.5281/zenodo.3816781.

which the trainers distributed the evaluation questionnaire among participants after each training, and the participants returned the completed forms to them in envelopes. The survey first asked for demographic data followed by the mood assessment. Subsequently, the FIRE-CU and FIRE-CPX items and the validation scales were assessed. All questionnaires were completed in the seminar room immediately after the training, assuring a high response rate. The data were collected in seminars, exercises, courses, and command post exercises, as each training was part of the crisis management apprenticeship.

3.4 Statistical Analysis

EvaSys (version 7.0) was used to create and scan the paper-and-pencil based questionnaires. All statistical analyses were computed with R (R Core Team 2018) using the packages psych (Revelle 2018), plyr (Wickham 2011), lavaan (Rosseel 2012), ggformula (Kaplan and Pruim 2018), Hmisc (Harrell 2018), and mice (van Buuren and Groothuis-Oudshoorn 2011). Assessing the reliability of the scales, an ANOVA was used to test whether a congeneric or an essentially τ-equivalent measurement model better fits the data. For testing the postulated four-factor model for the process level of the FIRE-CU, the two-factor model for the outcome level of the FIRE-CU, and the one-factor model for FIRE-CPX scale, we used a confirmatory factor analysis.

4 Results

Descriptive data for the FIRE-CU and the FIRE-CPX in terms of means, standard deviations, and answer distribution parameters can be found in Table D1 in the online supplement. Additional descriptive data and correlations of all included measures are presented in Table E1 in the online supplement (at https://doi.org/10.5281/zenodo.3816781). The main analyses presented in the following section focus on factorial structure, reliability, and validity of FIRE-CU and FIRE-CPX.

4.1 Factorial Structure

To test the assumptions made about the factorial structure, confirmatory factor analyses were conducted. As outlined earlier, a four-factor model was expected for the FIRE-CU process scales, which show a good model fit according to Hu and Bentler (1999) (RMSEA = 0.06 [0.04–0.07], SRMR = 0.05, CFI = 0.96, TLI = 0.95). Only the χ^2 indicates a poor model fit ($\chi^2(59) = 111.55$, $p < 0.001$), which is typically found for large sample sizes and therefore always requires further indices to judge the fit. Modification indices of this model suggest an inter-item correlation between the second and third items of the group scale, which results in a slightly improved model fit ($\chi^2(58) = 96.451$, $p < 0.01$, RMSEA = 0.05 [0.03–0.06], SRMR = 0.04, CFI = 0.97, TLI = 0.96). Just as in the initial model, the χ^2 is significant and indicates a poor model fit, but according to Schermelleh-Engel et al. (2003), in relation to the degrees of freedom the χ^2 is acceptable (χ^2/df = 1.66). According to the similar content of these items, a correlation between them seemed appropriate. The whole model is shown in Fig. 1.

The two-factor model for the FIRE-CU outcome scales shows a poor fit in all fit indices for the initial model ($\chi^2(26) = 140.96$, $p < 0.001$, RMSEA = 0.12 [0.11–0.14], SRMR = 0.07, CFI = 0.87, TLI = 0.82). Checking the modification indices, a few inter-item correlations might improve the model fit, namely intercorrelations between the first and second item of the transfer scale and intercorrelations between the third item of the competence scale and the second, fifth, and sixth items of the same scale, and between the fifth and sixth competence items. Theoretically, the competences of command unit members are complex and interdependent (Lamers 2016); therefore these model modifications seem appropriate. Adding these intercorrelations improves the model fit, which then displays a good fit $\chi^2(21) = 45.90$, $p < 0.01$, RMSEA = 0.06 [0.04–0.09], SRMR = 0.04, CFI = 0.97, TLI = 0.95]. Figure 2 shows the complete model.

In the present study, the additional scale for evaluating mission exercises was also amended and renamed FIRE-CPX. Based on prior findings, a one-factor model was tested; the model showed an overall excellent fit ($\chi^2(2) = 1.75$, $p > 0.10$, RMSEA = 0.00 [0.00– 0.11], SRMR = 0.02, CFI = 1.00, TLI = 1.01). Figure 3 shows the model with all path coefficients.

4.2 Reliability

Table 2 shows Cronbach's α and ω_H for all scales. Cronbach's α is a commonly used indicator for reliability (Trizano-Hermosilla and Alvarado 2016) and is the appropriate measure for FIRE-CU *trainer's behavior* and *group* scales, as an essentially τ-equivalent measurement model fits the data best. In contrast, the remaining scales show a better fit for the congeneric measurement model, which indicates that ω_H is the more appropriate measurement for reliability. The results of the χ^2-difference tests can be obtained in the last three columns of Table 2. The reliability in terms of internal consistency of the FIRE-CU scale *structure* and the FIRE-CPX can be judged as sufficient, and for all other FIRE-CU scales it can be judged as good, when applying the reliability standards for the assessment of program evaluation and learning success

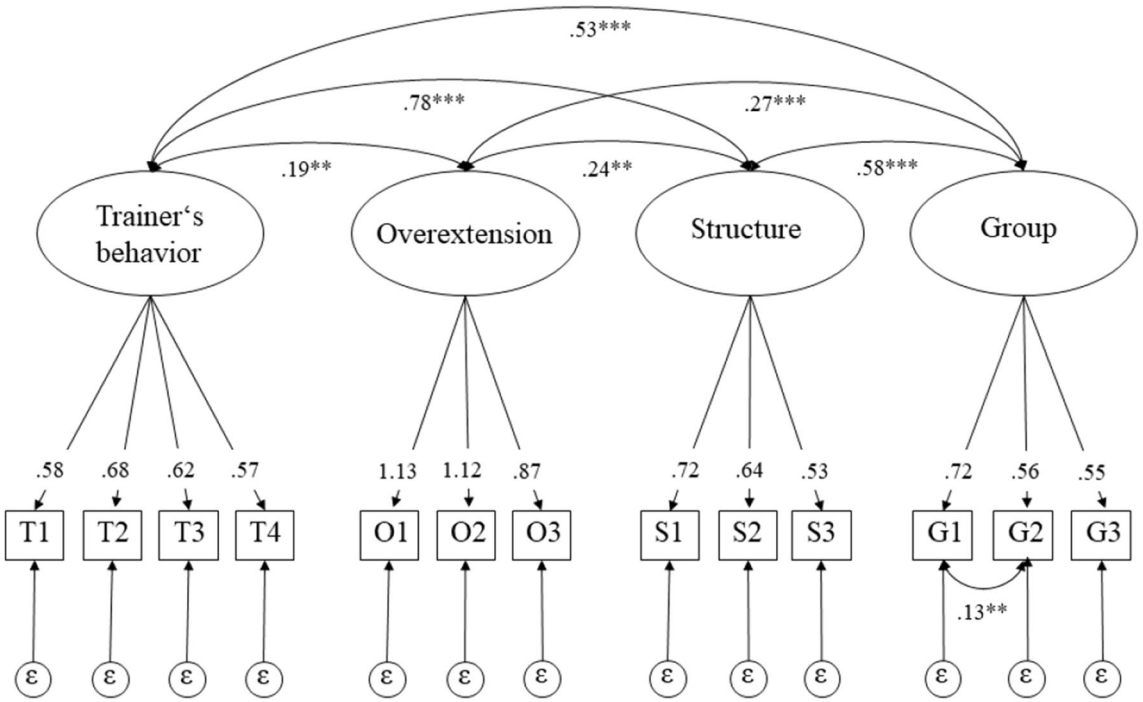

Fig. 1 Results of confirmatory factor analysis for process scales; standardized coefficients are reported. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

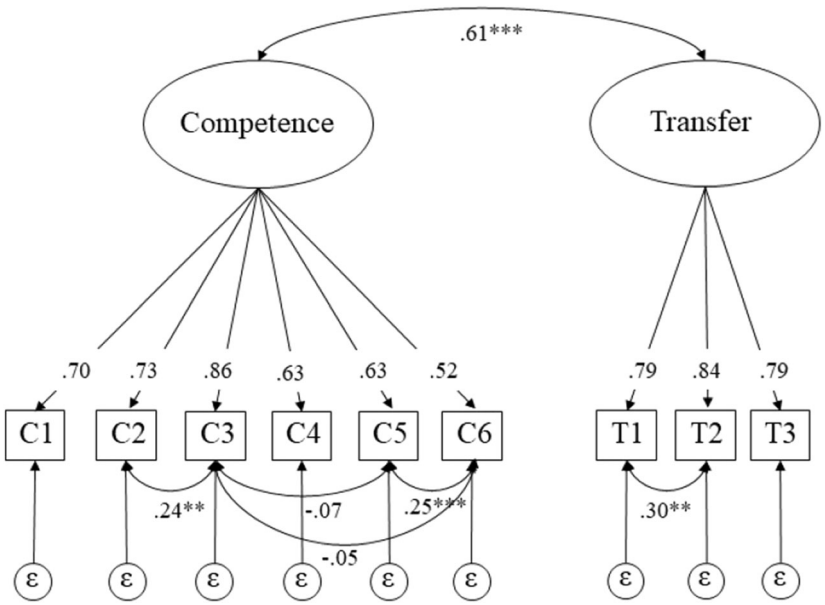

Fig. 2 Results of confirmatory factor analysis for outcome scales; standardized coefficients are reported. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

(Evers 2001). According to Nunnally (1978), the reliability for the FIRE-CU *overextension* scale can be rated as excellent.

4.3 Validity

We investigated the validity of the FIRE-CU and the FIRE-CPX by applying convergent, divergent, and concurrent validation measures. Convergent validity is the extent of agreement among theoretically highly related measures (Nunnally 1978). For all FIRE-CU scales, the associations with the corresponding scales show large and significant effects (Table 3). The process scales *trainer's behavior* and teaching competence (HILVE, Rindermann 2001) correlate with $r = 0.61$ ($p < 0.001$), while *overextension* and the eponymous HILVE scale correlate with $r = 0.57$ ($p < 0.001$). The FIRE-CU scale *structure* and the items of structure and organization (HILVE, Rindermann 2001)

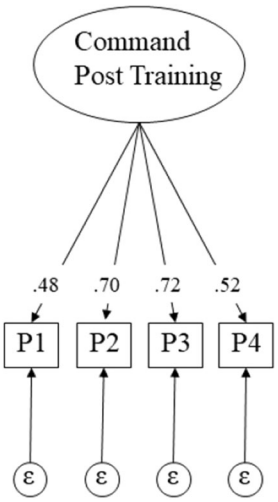

Fig. 3 Results of confirmatory factor analysis for the FIRE-CPX scale; standardized coefficients are reported. *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table 2 Reliability coefficients and measurement model tests for all FIRE-CU scales and FIRE-CPX

Scale	Cronbach's α	ω_H	*df*	$\Delta\chi^2$	*p*
Trainer's behavior	0.82	0.82	5	12.23	0.09
Overextension	0.90	0.91	2	35.91	< 0.001
Structure	0.75	0.76	2	17.99	< 0.001
Group	0.88	0.88	2	2.56	0.28
Competence	0.85	0.85	14	147.18	< 0.001
Transfer	0.86	0.87	2	42.80	< 0.001
FIRE-CPX	0.76	0.77	5	16.65	< 0.001

N = 288. χ^2-difference tests compare congeneric and essentially τ-equivalent models

show a correlation of $r = 0.50$ ($p < 0.001$), whereas *group* and fellow students (HILVE, Rindermann 2001) show $r = 0.45$ ($p < 0.001$). The FIRE-CU outcome scales show high correlation with corresponding measures as well: self-rated *competence* and professional competence (GEKo, Paechter et al. 2011) $r = 0.58$ ($p < 0.001$); *transfer* and self-developed validation items $r = 0.69$ ($p < 0.001$). The FIRE-CPX also correlates well with the corresponding self-developed validation items ($r = 0.48$, $p < 0.001$). These results support the assumption that FIRE-CU and FIRE-CPX assess the training quality by showing the expected high correlations with established and validated scales for teaching quality and additional convergent measures.

Divergent validity refers to the degree of disagreement between theoretically unrelated (or less related) constructs (Nunnally 1978). Associations between trainees' moods and the FIRE-CU scales are midsized, though all are significant ($0.18 \leq r \leq 0.32$, $p < 0.01$). Even though the correlations are slightly higher for some items, overall the correlations with the mood are moderate and, in general, are lower than the correlations with the corresponding scales, supporting the assumption that mood does not crucially influence trainees' ratings. Further, the age of participants showed mostly no association at all with given evaluations (Table 3), which serves as a further indicator for divergent validity.

Criterion validity refers to the ability of a measure to predict a concurrently or subsequently assessed criterion (Nunnally 1978). In the present study, the criterion validity was assessed at the same time as the main scales and, thereby, represents the concurrent validity. For both measures, FIRE-CU and FIRE-CPX, mostly midsized to high associations were found: Participants were asked to rate the training with German school grades (1 = very good to 6 = insufficient, in the results recoded for the sake of illustration). Correlations range from $r = 0.19$ ($p < 0.01$) for FIRE-CU *overextension* to $r = 0.45$ ($p < 0.001$) for *trainer's behavior*. Further, correlations to overall satisfaction range from $r = 0.19$ ($p < 0.01$) for FIRE-CU *overextension* to $r = 0.48$ ($p < 0.001$) for *transfer*. All in all, these findings suggest a valid assessment of criteria using the FIRE-CU and FIRE-CPX.

5 Discussion

The present work's aim was to provide and validate tools for the evaluation of fire service command unit trainings. The results across all analyses provided clear support for the psychometric quality of the FIRE-CU and the FIRE-CPX. The replication of the four-factor model for process data and the two-factor model for outcome data was successful. Likewise, the CFA for the one-factor model for the FIRE-CPX shows an excellent model fit. However, unlike in the original FIRE scales (Schulte and Thielsch 2019), the FIRE-CU outcome data show high inter-item correlations for the two-factor model. Taking a closer look at the functions of command unit members, it becomes obvious that they cannot concentrate on one major incident but rather split their attention between different simultaneously changing and interfering events (Lamers 2016). These situations require immediate, decisive and directive actions to predict what will ensue (Baran and Scott 2010; Dixon et al. 2017). Reacting on all incoming information requires simultaneous actions, and, therefore, the interrelations between the competences seem an appropriate finding.

Reliability of FIRE-CU and FIRE-CPX, in terms of internal consistency, was satisfactory to good, particularly in the light of the brevity of the scales. Reliability indicators are even slightly higher than those of the original FIRE questionnaire (see Schulte and Thielsch 2019, p. 39). Since evaluation data are usually processed on the group level of evaluated trainings, all reliability values, including the

Table 3 Correlations with validation measures

	Trainer's behavior	Overextension (r)	Structure	Group	Competence	Transfer	FIRE-CPX
Convergent measures (marked in italics)							
Teaching beh. (H)	*0.61****	0.08	0.51***	0.30***	0.30***	0.41***	0.44***
Overextension (H)	0.17**	*0.57****	0.26***	0.28***	0.06	0.34***	0.19***
Structure and org. (H)	0.53***	0.13*	*0.50****	0.27***	0.26***	0.34***	0.38***
Fellow students (H)	0.25***	0.27***	0.26***	*0.45****	0.17**	0.26***	0.17**
Prof. comp. (G)	0.49***	0.09	0.49***	0.36***	*0.58****	0.50***	0.51***
Transfer (own)	0.48***	0.19**	0.49***	0.40***	0.57***	*0.69****	0.57***
Com. post ex. (own)	0.29***	0.06	0.26***	0.10	0.22***	0.22***	*0.48****
Divergent measures							
Mood	0.31***	0.32***	0.25***	0.29***	0.18**	0.31***	0.28***
Age	0.09	0.06	0.09	0.03	0.01	0.24***	0.12
Concurrent measures							
Grade (r)	0.45***	0.19**	0.42***	0.33***	0.28***	0.44***	0.39**
Overall satisfaction	0.44***	0.19**	0.44***	0.39***	0.31***	0.48***	0.43***

N = 288. Teaching beh. (H) = Teaching behavior (HILVE scale); Overextension (H) = Overextension (HILVE scale); Structure and org. (H) = Structure and organization (HILVE scale); Fellow students (H) = Fellow students (HILVE scale); Prof. comp. (G) = Professional Competence (GEKo scale); Com. post ex. (own) = Command post exercises (own items); r = reverse coded; grade was rated from 1 = very good to 6 = insufficient but is recoded in this table for the sake of illustration

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

comparatively slightly lower values of the FIRE-CU scale *structure* and the FIRE-CPX, are absolutely sufficient for the intended purposes.

Lastly, the pattern of correlations with the validation scales clearly support the assumption that the FIRE-CU and the FIRE-CPX are measuring the intended content: Strong associations with concurrently surveyed criteria are found, correlations with all convergent measures are high, correlations to divergent measures are clearly lower or, with respect to the age of the interviewees, mostly even not significant. Small correlations between FIRE-CU and FIRE-CPX assessments and mood were to be expected, as they were also found for the original FIRE measure (Schulte and Thielsch 2019), and it can be argued that teaching quality and learning success should have a positive effect on the trainees' mood. Therefore, as mentioned before, mood might not be seen as a completely divergent variable.

As a whole, the present findings indicate that the FIRE-CU and the FIRE-CPX scale can be appropriately applied for the evaluation of fire service command unit trainings. The benefits of these scales are that trainers are provided with a validated and economic tool for assessing trainees' perceptions of the training quality. Further, separating the evaluation tools according to process and outcome data provides the training developer with critical information about learning achievements as well as potential ways to adjust the training process (Blanchard and Thacker 2013). In other words, if trainees indicate that they did not achieve the learning goals, the evaluation manager is able to check whether there was a problem in the process. Especially for the fire service command unit, solid training and realistic command post exercises are vital, as the units are rarely in action but when they are indeed deployed, they bear the responsibility for many lives (Lamers 2016; Grunnan and Fridheim 2017). Therefore, it is important to evaluate whether trainees manage to develop the desired competencies during training and whether they can transfer them to the command post exercises.

Even though the intention of our work was not to define effective leadership in dangerous contexts, we did, nevertheless, have to work out the competencies conveyed during leadership trainings in order to provide measurable parameters. Prior research indicated, for example, that leaders of emergency services and crisis managers should be able to handle and make decisions under stress, know their co-workers as well as their personal limitations, and communicate assignments (Sjoberg et al. 2011; Haus et al. 2016; Schulte and Thielsch 2019). For the command unit, the competencies were amended on the basis of literature and a pre-analysis of the items, leading to the conclusion that construction, co-construction, and constructive conflict are important abilities for command unit members (van den Bossche et al. 2006; Thieme and Hofinger 2012; Boon et al. 2013). The so extended FIRE-CU *competence* scale was successfully validated. Nevertheless, more research on

the competencies of leaders in dangerous contexts is desirable in order to be able to evaluate this even more precisely.

6 Practical Implications and Application

For practical application, both instruments, the FIRE-CU and the FIRE-CPX, will need to be used for evaluating trainings, as typical command unit trainings will include theoretical lessons as well as exercises. Yet, both measures were tested separately and, thus, if they are not applicable, single FIRE-CU scales or the FIRE-CPX scale can be omitted. However, individual items should not be removed from the scales, as the FIRE-CU and FIRE-CPX scales are already very brief, and further shortening may impair psychometric quality. In general, the wording of each question should not be changed, except for possible minor adjustments to ensure comprehensibility and optimal adaptation to the evaluation context. When applying the questionnaires, we recommend completely removing data from respondents who omit three or more items from the subsequent analysis. A high number of missing values in many individuals may be due to a lack of fit between the questionnaire and the evaluation context.

FIRE-CU and FIRE-CPX are rated on a 7-point Likert scale and allow a simple analysis and interpretation of the data.[9] Answering the 22 FIRE-CU and the four FIRE-CPX items will only take a few minutes (in our experience about 4–5 min). In order to protect the anonymity of the trainees, we generally recommend that an analysis should only be carried out if at least eight completed questionnaires are returned (or if at least 50% of the trainees in small training courses with 10–15 participants took part in the survey). In principle, it is important to ensure that the evaluation is anonymous. Accordingly, no (for example, demographic) variables should be collected on the basis of which conclusions can be drawn about individual persons. Following the recommendations of Kirkpatrick and Kirkpatrick (2006) it may be useful to add an open feedback field to a survey, for example with a question such as "Please give feedback to the trainers (suggestions/praise/constructive criticism)." In this case it is also important to make sure that people cannot be identified (for example, by their handwriting or comments).

Basically, the subjective character of such a training evaluation should be taken into account. This feedback enables trainers to obtain important information about their own teaching activities from the participants' point of view. It would therefore be useful, if time permits, to have a meeting with the trainees to discuss the results of the evaluation, to clarify possible misunderstandings and to collect ideas for improvements. The questionnaires are also suitable for comparison with previous similar courses or between different teaching concepts. If no comparative data are available, the descriptive statistics in Appendix D in the online supplement can serve as an anchor for the interpretation of results. In general, the responsible organization should support the evaluation both technically and in terms of its content. In particular, such support should include assistance if evaluations repeatedly reveal potential for improvement.

With respect to possible application areas, the following should be considered: Development and validation of the FIRE(-CU/-CPX) are based on data from fire service trainings, and the scales are thus particularly suitable for such training evaluations. In addition, FIRE-CU and FIRE-CPX might be suitable for the evaluation of military command unit trainings, as the fire service command unit evolved from the military command unit (Heath 1998; Lamers 2016; Schaub 2016); specifically, the units share similarities in their subject areas (Lamers 2016) and functioning (Schaub 2016). Furthermore, the fire command unit has similar duties and requirements as do high responsibility teams (HRT) in high reliability organizations (HRO), such that the FIRE-CU and FIRE-CPX might be applicable for all kinds of HRO command units. Thus, trainings for rescue services, police units, and technical rescue units might also be evaluated with the FIRE-CU and FIRE-CPX. Although, even as communication and coordination demands in these teams might be comparable, any application in other areas requires a validation for the intended use context.

7 Limitations and Further Research

There are some limitations to be considered when interpreting the findings, but at the same time they open up perspectives for possible future research. In the present study, all participants received the German version of FIRE-CU and FIRE-CPX, since all evaluations took place during trainings in Germany. Even though Appendix F in the online supplement contains English versions of both instruments, the authors recommend a validation for the English FIRE-CU and FIRE-CPX version before use. However, based on the commonalities of the command unit organization in Germany and other countries using the Incident Command System, it is generally assumed that both evaluation instruments can also be useful there. Thus, any further translations and validations into other languages are highly welcome.

[9] See Appendix F in the online supplement at https://doi.org/10.5281/zenodo.3816781 for scoring instructions.

In line with Schulte and Thielsch (2019) one might argue that the confirmation of the postulated factorial structure of the FIRE-CU suggests that trainees' perception of training quality in fire service contexts relies on similar criteria as students' evaluation of teaching in higher education. The application of comparable teaching methods (for example, seminars, group tasks, practical exercises) (Lamers 2016) suggests that the same factors might be responsible for good teaching in university contexts and good training in fire services. However, a recent meta-analysis doubts the meaningfulness of student evaluations, after no correlations were found between students' teaching evaluation and student's achievements, and the small-to-midsized effects found in previous meta-analyses could be explained by publication biases (Uttl et al. 2017). On the contrary the FIRE(-CU/-CPX) scales were developed with and are used for firefighters, who—in contrast to students—are experienced in the subject they are being taught. Following the argumentation of Grohmann and Kauffeld (2013), the fire service trainees can be rated as experts in their field who can provide valid assessment of the learned training content. Furthermore, involving experts in program evaluations allows insights into the appropriateness of the assessed content (Wroblewski and Leitner 2009). This means that the raters of the FIRE-CU and FIRE-CPX are able to assess the quality of the training content, which is in line with previous studies implying that learning achievements can be measured validly by self-reports of trainees (Kraiger et al. 1993).

Moreover, both measures were tested with a sample consisting of 96% male participants. Even though this is the usual gender ratio for the fire services, any potential gender differences in the evaluation of command unit trainings cannot be assessed, which reveals a direction for further research.

Furthermore, extending the evaluation of the fire service training on level III (behavior change) and IV (results) of Kirkpatrick's model (Kirkpatrick and Kirkpatrick 2006) would broaden the insight into the quality of the trainings. Such an extension would provide knowledge about the behavior of command unit members during real incidents, whether training content is applied the way it is taught, and what impact this has on the results. In addition, if subsequent evaluation tools use behavioral and result criteria, it might be important to consider the social context and the favorability of the post-training environment, as these factors can tremendously influence the transfer of training content (Arthur et al. 2003). Therefore, further research on the last two evaluation levels is necessary.

Lastly, for the fire service command unit or other command units, the shared mental models are of special interest (Badke-Schaub et al. 2012; Hofinger and Heimann 2016; Lamers 2016). Command units are faced with a large amount of information from which a shared understanding of the situation and the upcoming actions must be distilled (Badke-Schaub et al. 2012). A few suggestions for shared mental model assessment can be found in Mohammed et al. (2010). We tried to integrate this aspect within the FIRE-CU *competence* scale, yet with just three items this aspect is only screened. For an in-depth evaluation of shared mental models of command units during command post exercises, more detailed assessment tools have to be developed.

8 Conclusion

By building on existing scales for group and platoon leader trainings, this study provides extensively tested and validated tools for the evaluation of command unit trainings: the FIRE-CU and the FIRE-CPX measures. Although both instruments were validated in the context of educating command units, they could also be used for the evaluation of various tactical team trainings and other HRTs if the validity of the intended context is verified. Giving attention to these work and leadership environments might serve as a foundation for further research on this topic. As the present study provides critical tools for measuring the quality of command unit trainings, we hope that these tools will prove useful both for practical application and for research. Insight on the performance of command units not only affects members of the investigated organizations, but also has implications for a broader community that depends on fire services, police, and the other rescue forces in emergency situations. Hopefully, the present research can help contribute to a field in which high-quality training not only improves the work of one organization but also bolsters the standard of living for the society as a whole.

Acknowledgements This study was supported by the State Fire Service Institute NRW with non-monetary resources. The authors particularly thank Patrick Fuchs, Christoph Lamers, and Yannick Ngatchou of the State Fire Service Institute NRW for their helpful support during the evaluation project. Furthermore, we thank Guido Hertel for his valuable feedback on the study as well as Celeste Brennecka for her very helpful comments on earlier versions of this manuscript.

References

Alliger, G.M., S.I. Tannenbaum, W. Bennett, H. Traver, and A. Shotland. 1997. A meta-analysis of the relations among training criteria. *Personnel Psychology* 50(2): 341–358.

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. 2014. *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Arthur, W., W. Bennett, P.S. Edens, and S.T. Bell. 2003. Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology* 88(2): 234–245.

Badke-Schaub, P., G. Hofinger, and K. Lauche (eds.). 2012. *Human factors.* Berlin: Springer.

Baran, B.E., and C.W. Scott. 2010. Organizing ambiguity: A grounded theory of leadership and sensemaking within dangerous contexts. *Military Psychology* 22(sup1): S42–S69.

Berlin, J.M., and E.D. Carlström. 2014. Collaboration exercises – The lack of collaborative benefits. *International Journal of Disaster Risk Science* 5(3): 192–205.

Blanchard, P.N., and J.W. Thacker. 2013. *Effective training: Systems, strategies, and practices*, 5th edn. Boston: Pearson.

Boon, A., E. Raes, E. Kyndt, and F. Dochy. 2013. Team learning beliefs and behaviours in response teams. *European Journal of Training and Development* 37(4): 357–379.

Cannon-Bowers, J.A., E. Salas, and S. Converse. 2013. Shared mental models in expert team decision making. In *Individual and group decision making: Current issues*, ed. N.J. Castellan, 221–246. Hoboken: Taylor and Francis.

Clark, L.A., and D. Watson. 1995. Constructing validity: Basic issues in objective scale development. *Psychological Assessment* 7(3): 309–319.

Dixon, D.P., M. Weeks, R. Boland, and S. Perelli. 2017. Making sense when it matters most: An exploratory study of leadership in extremis. *Journal of Leadership & Organizational Studies* 24(3): 294–317.

Evers, A. 2001. The revised Dutch rating system for test quality. *International Journal of Testing* 1(2): 155–182.

Feuerwehr-Dienstvorschrift (FwDV) 2. 2012. Fire service regulations 2: Training of volunteer fire service (*Ausbildung der Freiwilligen Feuerwehr*). http://www.bbk.bund.de/SharedDocs/Downloads/BBK/DE/FIS/DownloadsRechtundVorschriften/Volltext_Fw_Dv/FwDV_2_Stand_Jan2012.pdf?__blob=publicationFile. Accessed 8 May 2020.

Gagne, P., and G.R. Hancock. 2006. Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research* 41(1): 65–83.

Grohmann, A., and S. Kauffeld. 2013. Evaluating training programs: Development and correlates of the questionnaire for professional training evaluation. *International Journal of Training and Development* 17(2): 135–155.

Grunnan, T., and H. Fridheim. 2017. Planning and conducting crisis management exercises for decision-making: The do's and don'ts. *EURO Journal on Decision Processes* 5(1–4): 79–95.

Hagemann, V., A. Kluge, and J. Greve. 2012. Measuring the effects of team resource management training for the fire service. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 56(1): 2442–2446.

Hagemann, V., A. Kluge, and S. Ritzmann. 2012. Flexibility under complexity: Work contexts, task profiles and team processes of high responsibility teams. *Employee Relations* 34(3): 322–338.

Harrell Jr, F.E. 2018. Hmisc: Harrell miscellaneous (Version 4.1–1). https://CRAN.R-project.org/package=Hmisc. Accessed 8 May 2020.

Haus, M., C. Adler, M. Hagl, M. Maragkos, and S. Duschek. 2016. Stress and stress management in European crisis managers. *International Journal of Emergency Services* 5(1): 66–81.

Heath, R. 1998. Dealing with the complete crisis – The crisis management shell structure. *Safety Science* 30(1–2): 139–150.

Hofinger, G., and R. Heimann (eds.). 2016. Handbook staff work (*Handbuch Stabsarbeit*). Berlin: Springer.

Hu, L.-T., and P.M. Bentler. 1999. Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal* 6(1): 1–55.

Irwing, P., and D.J. Hughes. 2018. Test development. In *The Wiley handbook of psychometric testing*, ed. P. Irwing, T. Booth, and D.J. Hughes, 3–48. Chichester, UK: John Wiley and Sons.

Jäger, R. 2004. Construction of a rating scale with smilies as symbolic marks (*Konstruktion einer Ratingskala mit Smilies als symbolische Marken*). *Diagnostica* 50(1): 31–38.

Kaplan, D., and R. Pruim. 2018. ggformula. https://CRAN.R-project.org/package=ggformula. Accessed 8 May 2020.

Karsten, A. 2012. Leading an operational-tactical unit: The tasks of the head of a command unit at D level in civil protection (*Leiten eines operativ-taktischen Stabes: Die Aufgaben der Leiterin/des Leiters eines operativ-taktischen Stabes der Führungsstufe D im Bevölkerungsschutz*). *Bevölkerungsschutz* 3: 16–19.

Kirkpatrick, D.L. 1979. Techniques for evaluating training programs. *Training and Development Journal* 33: 78–92.

Kirkpatrick, D.L., and J.D. Kirkpatrick. 2006. *Evaluating training programs: The four levels*, 3rd edn. San Francisco: Berrett-Koehler.

Kraiger, K., J.K. Ford, and E. Salas. 1993. Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *The Journal of Applied Psychology* 78(2): 311–328.

Lamers, C. 2016. Command units in civil protection: History, analysis and suggestions for optimisation (*Stabsarbeit im Bevölkerungsschutz: Historie, Analyse und Vorschläge zur Optimierung*). Edewecht: S + K Verlagsgesellschaft Stumpf + Kossendey mbH.

Mohammed, S., L. Ferzandi, and K. Hamilton. 2010. Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management* 36(4): 876–910.

Nunnally, J.C. 1978. *Psychometric theory*, 2nd edn. New York: McGraw-Hill.

Paechter, M., B. Maier, and D. Macher. 2011. Evaluation of university teaching by means of assessments of subjective competence acquisition (*Evaluation universitärer Lehre mittels Einschätzungen des subjektiven Kompetenzerwerbs*). *Psychologie in Erziehung und Unterricht* 58(2): Article 128.

R Core Team. 2018. The R project for statistical computing. http://www.R-project.org/. Accessed 2 May 2020.

Revelle, W. 2018. Psych: Procedures for psychological, psychometric, and personality research. https://CRAN.R-project.org/package=psych. Accessed 2 May 2020.

Rindermann, H. 2001. Student assessment of courses – Object of research and implications (*Die studentische Beurteilung von Lehrveranstaltungen – Forschungsgegenstand und Implikationen*). In Evaluation of university teaching: Between quality management and self purpose (*Evaluation universitärer Lehre: Zwischen Qualitätsmanagement und Selbstzweck*), ed. C. Spiel, 61–88. Münster: Waxmann.

Rindermann, H., and M. Amelang. 1994. The Heidelberg Inventory for Course Evaluation (HILVE). Instruction manual (*Das Heidelberger Inventar zur Lehrveranstaltungs-Evaluation (HILVE). Handanweisung*). Heidelberg: Asanger.

Ritzmann, S., V. Hagemann, and A. Kluge. 2014. The Training Evaluation Inventory (TEI) – Evaluation of training design and measurement of training outcomes for predicting training success. *Vocations and Learning* 7(1): 41–73.

Röseler, S., H. Thölking, M. Hagel, and M.T. Thielsch. 2020. Feedback Instrument for Rescue forces Education – Mission exercises (FIRE-E) (*Feedback-Instrument zur Rettungskräfte-Entwicklung – Einsatzübungen*). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis282.

Rosseel, Y. 2012. Lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48(2): 1–36.

Salas, E., and J.A. Cannon-Bowers. 2001. The science of training: A decade of progress. *Annual Review of Psychology* 52: 471–499.

Schaub, H. 2016. Command units in the German army (*Militärische Stäbe in der Bundeswehr*). In Handbook staff work (*Handbuch Stabsarbeit*), ed. G. Hofinger, and R. Heimann, 33–38. Berlin: Springer.

Schermelleh-Engel, K., H. Moosbrugger, and H. Müller. 2003. Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research* 8(2): 23–74.

Schönbrodt, F.D., and M. Perugini. 2013. At what sample size do correlations stabilize? *Journal of Research in Personality* 47(5): 609–612.

Schulte, N., and M.T. Thielsch. 2019. Evaluation of firefighter leadership trainings. *International Journal of Emergency Services* 8(1): 34–49.

Scriven, M. 1999. The nature of evaluation part I: Relation to psychology. *Practical Assessment, Research & Evaluation* 6(11). https://doi.org/10.7275/egax-6010.

Sjoberg, M., C. Wallenius, and G. Larsson. 2011. Leadership in complex, stressful rescue operations: A quantitative test of a qualitatively developed model. *Disaster Prevention and Management: An International Journal* 20(2): 199–212.

Sommer, M., and O. Njå. 2011. Learning amongst Norwegian firefighters. *Journal of Workplace Learning* 23(7): 435–455.

Thielsch, M.T., J.N. Busjan, and K. Frerichs. 2018. Feedback Instrument for Rescue forces Education – Examinations (FIRE-P) (*Feedback-Instrument zur Rettungskräfte-Entwicklung – Prüfungen*). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. http://doi.org/10.6102/zis260.

Thielsch, M.T., L. Kläpker, and L. Streppel. 2019. Feedback Instrument for Rescue forces Education – Basic Training (FIRE-B) (*Feedback-Instrument zur Rettungskräfte-Entwicklung – Basisausbildung*). *Zusammenstellung sozialwissenschaftlicher Items und Skalen (ZIS)*. https://doi.org/10.6102/zis266.

Thieme, U., and G. Hofinger, G. 2012. Command units and "Standing staffs" in the police: Security through professionalisation (*Stabsarbeit und » Ständige Stäbe « bei der Polizei: Sicherheit durch Professionalisierung*). In *Human factors*, ed. P. Badke-Schaub, G. Hofinger, and K. Lauche, 275–291. Berlin: Springer.

Thompson, B. 2004. *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* Washington, DC: American Psychological Association.

Trizano-Hermosilla, I., and J.M. Alvarado. 2016. Best alternatives to Cronbach's alpha reliability in realistic conditions: Congeneric and asymmetrical measurements. *Frontiers in Psychology* 7: Article 769.

Useem, M., J. Cook, and L. Sutton. 2005. Developing leaders for decision making under stress: Wildland firefighters in the south canyon fire and its aftermath. *Academy of Management Learning and Education* 4(4): 461–485.

Uttl, B., C.A. White, and D.W. Gonzalez. 2017. Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation* 54: 22–42.

Van Buuren, S., and K. Groothuis-Oudshoorn. 2011. Mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* 45(3): 1–67.

Van den Bossche, P., W.H. Gijselaers, M. Segers, and P.A. Kirschner. 2006. Social and cognitive factors driving teamwork in collaborative learning environments. *Small Group Research* 37(5): 490–521.

Wickham, H. 2011. Plyr: The Split-Apply-Combine Strategy for data analysis. *Journal of Statistical Software* 40(1): 1–29.

Wroblewski, A., and A. Leitner. 2009. Between scientific standards and claims to efficiency: Expert interviews in programme evaluation. In *Research methods series. Interviewing experts*, ed. A. Bogner, B. Littig, and W. Menz, 235–251. Basingstoke: Palgrave Macmillan.

Wybo, J.L., and K.M. Kowalski. 1998. Command centers and emergency management support. *Safety Science* 30(1–2): 131–138.