Lena Niemann* and Meinald T. Thielsch

Evaluation of Basic Trainings for Rescue Forces

https://doi.org/10.1515/jhsem-2019-0062
Received October 31, 2019; accepted July 7, 2020

Abstract: Since members of rescue forces such as firefighters have to deal with sometimes extreme and dangerous situations, high-quality basic trainings are indispensable for their professional success. There is therefore an obvious need for standardized tools assessing the training quality. This paper aims to develop and validate such an evaluation instrument. In Study 1, a qualitative analysis (*N* = 21) was used to identify core characteristics of good firefighter basic trainings and served as theoretical basis for the generation of corresponding items. In Study 2 (*N* = 257), the item set was piloted and reduced, its structure was assessed in exploratory factor analyses, and first validations were conducted. Study 3 (*N* = 451) tested the proposed factor structure via confirmatory analyses and validated the questionnaire comprehensively. Factor analyses showed a six-factor structure. The scales of the newly created Feedback Instrument for Rescue forces Education – Basic education (FIRE-B) are to be judged as reliable. Moreover, there are several clear indications of validity. Thus, the present research contributes to the understanding of critical factors and processes of basic trainings. Furthermore, the FIRE-B has a high practical relevance, both in the assessment of training quality and in the identification of opportunities for improvement.

Keywords: evaluation, vocational training evaluation, firefighter, fire service, rescue forces, questionnaire

1 Introduction

Worldwide, numerous people work in rescue services and fire brigades. Their activities are frequently characterized by danger, physical and mental exertion as

***Corresponding author: Lena Niemann**, Department of Psychology, University of Münster, Fliednerstr. 21, 48149 Münster,Germany, E-mail: lena.niemann@uni-muenster.de. https://orcid.org/0000-0002-6793-4188
Meinald T. Thielsch, Department of Psychology, University of Münster, Münster, Germany, E-mail: thielsch@uni-muenster.de. https://orcid.org/0000-0001-8493-9071

well as high unpredictability (Taber, Plumb, and Jolemore 2008, 280). Therefore, firefighters, along with other rescue workers, need a variety of skills to be optimally prepared for their specific work requirements. These skills include cognitive, physical and social skills (Burke 1997; Henderson 2010). The main way firefighters develop the necessary knowledge and skills is through a good firefighter basic training. When such a training is subpar, the likelihood of both mistakes and work-related injuries increases (Moore-Merrell et al. 2008), which, in turn, is significantly associated with symptoms of burnout and PTSD (Katsavouni et al. 2016). Consequently, high-quality trainings for firefighters not only serve the interest of society, which expects competent personnel to respond to critical incidents, but they also serve firefighters' own interests, as they help to minimize their risk of personal injury while also protecting their mental health (e.g., Katsavouni et al. 2016; Moore-Merrell et al. 2008).

Thus, it is important to evaluate firefighter training programs in order to assess the quality of training and to identify potential areas for improvement. Such evaluations can be understood as the systematic judgment of a program's worth or value (Steele 1970). Yet, according to the current state of knowledge, there is no scientifically developed instrument for evaluating firefighter basic trainings that would enable trainees to validly assess the training quality. Thus, the present research answers this need for a valid and scientifically based quality assessment instrument for firefighter basic trainings.

1.1 Firefighter Basic Trainings

Internationally, the structure of fire brigades is quite diverse, and different nations have different distributions of volunteer and full-time firefighters (Brushlinsky et al. 2019). Nonetheless, Bukowski and Tanaka (1991) proposed four central points as international performance code for firefighters: "Prevent the fire or retard its growth and spread, protect building occupants from the fire effects, minimize the impact of fire [and] support fire-service operations" (175–176). Thus, firefighters' goals are to save lives, property, and the environment. For these purposes every firefighter must receive appropriate training.

In Germany, fire protection is provided by professional fire brigades as well as volunteer fire brigades (cf. Brushlinsky et al. 2019). In both cases, future firefighters must first complete appropriate basic training (cf. FwDV 2 2012), which is provided by municipalities, districts and cities. At the voluntary level, basic training includes troop training and technical training. At the professional level, basic training involves several years of full-time training. Troop training is dedicated to the smallest tactical fire brigade unit, the troop. This usually consists of two firemen: A troop leader and a troop man. The troop leader bears the responsibility for his troop man and usually takes orders from a group leader, who, in turn, is subordinate to a platoon leader. Accordingly, the troop training can be subdivided into troop man training and troop leader training. The curriculum for troop training at the voluntary fire brigade level consists of various modules on topics such as legal bases, vehicle knowledge, equipment knowledge and life-saving emergency measures. It comprises a total of 185 h. Complementing and building on this, the second part of basic training at the voluntary fire brigade level is technical training (FwDV 2 2012). According to FwDV 2 (2012), the technical training includes, for example, additional training courses for radio operators, respirator wearers and machine operators as well as special training courses on how to behave when working with hazardous substances. The combination of troop and technical training at volunteer fire brigades is roughly comparable with the training required to become a professional firefighter at the professional fire brigade level. A central difference is that the basic training at the professional fire brigade level additionally includes training as a paramedic as well as internships at fire stations.

Regulations (FwDV 2 2012; VAP1.2-Feu 2015) specify the design of basic trainings, including components such as adequate content, learning objectives, time targets and methods. However, these training descriptions only represent minimum requirements. They are regarded as recommendations that can be supplemented. Accordingly, slight differences exist in the implementation of basic trainings depending on the respective federal state, district, municipality and trainer (Buchenau 2020). For example, even though the specified learning objectives must be achieved in the specified time frame (FwDV 2 2012), trainers can use methods and didactic approaches that are not described in the regulations. Furthermore, selection of trainers for the professional fire brigade focuses more on the trainers' technical rather than pedagogical aptitude (VAP1.2-Feu 2015; Meyer and Stiegel 2012).

This training situation underlines the need for an evaluation to assess the current quality of basic training and to identify potential areas for further improvement. The desired result, namely high-quality firefighter basic trainings, will not only support good practices in emergency situations but will also support the physical and psychological health of firefighters. Specific evaluations can be helpful for recording and improving the quality of firefighters' initial trainings in order to reduce the risks associated with the profession. For this reason, the concept of program evaluations will be explained in more detail in the following section.

1.2 Program Evaluation

Widely used in the context of evaluating training programs is the four-level model by Kirkpatrick (1979) (Blau et al. 2012). According to the model, the evaluation of trainings should take place at the four consecutive levels of reaction, learning, behavior and results (see Figure 1). According to Kirkpatrick (1979), the first level *reaction* covers the subjective, emotional evaluation of a training. This can be operationalized with questions about training relevance, materials and exercises, and reactions to the trainer or premises (Blanchard and Thacker 2010; Kirkpatrick 2007). Positive reactions not only indicate that participants are highly motivated but also that they are paying close attention – processes that are presupposed for participants' learning success (Blanchard and Thacker 2010). Secondly, according to Kirkpatrick (1979), evaluations should investigate the level of *learning*. This refers to the extent to which participants expand their knowledge, develop their skills or change attitudes through a training. The trainees' learning can be assessed with items such as "After the training, I know substantially more about the training contents than before" (Grohmann and Kauffeld 2013, 142). This evaluation criterion is of central importance to determine how well trainers promote the learning of participants and to uncover potential improvements (Kirkpatrick 2007). Thirdly, at the level *behavior*, an evaluation should examine whether a change in behavior, i.e., a transfer of learned training content to everyday situations, has taken place as a result of a training (Kirkpatrick 1979). Finally, at the level of *results*, an evaluation should record the broad organizational effects of a training. For example, a training might lead to reduced costs or improved service quality (Kirkpatrick 1979). Kirkpatrick and Kirkpatrick (2006) stress that the four levels of the model are to be understood as successive stages. Each individual level has an influence on the next level. For this reason, Kirkpatrick (1979) suggests evaluating a program on higher levels only after it has been shown to be successful on lower levels. However, implementing subsequent evaluation levels steadily becomes more difficult and resource intensive (Kennedy et al. 2014).

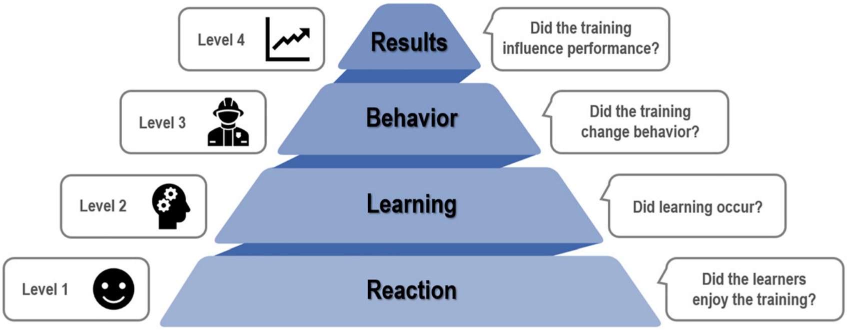
Level 4
Results
Did the training influence performance?
Level 3
Behavior
Did the training change behavior?
Level 2
Learning
Did learning occur?
Level 1
Reaction
Did the learners enjoy the training?

Figure 1: Kirkpatrick's four levels of training evaluation.

Therefore, we aim to develop an evaluation questionnaire that covers the levels of reaction and learning. At the reaction level, we will record firefighters' affective reactions to the trainings and their perceived usefulness (cf. Blanchard and Thacker 2010). At the learning level, self-assessments of whether the firefighters acquired specific competences should provide information on subjective learning success (cf. Kirkpatrick and Kirkpatrick 2006). The present work aims at covering both of these levels in a standardized questionnaire (cf. Blanchard and Thacker 2010; Kirkpatrick and Kirkpatrick 2006). Before describing the questionnaire, the next section examines the extent to which evaluations are common in firefighter basic trainings.

1.3 Evaluation in the Context of Firefighter Basic Trainings

Regarding firefighter trainings, evaluations are critically important to achieve optimum training outcomes. In basic training courses, it is already common to evaluate participants' performances after practical exercises. Such evaluations often take place in the form of "lessons learned" (Berlin and Carlström 2014, 199), conversations that happen after the training exercises with the aim of discussing in the training group what worked and what could use improvement. Also, Childs (2005) states the importance of such critical reflection in firefighter education programs. Similarly, Sommer and Njå (2011) propose that sharing experiences is a good learning method in firefighter basic trainings.

Less common is the evaluation of firefighter basic trainings themselves. However, such an evaluation is equally important, as various training conditions can impede learning. For instance, learning effects are absent or very low if a training is badly structured, if exercises are unrealistic, if a training mainly focuses on existing knowledge or if trainers impart new knowledge using bad didactics (Berlin and Carlström 2014). To the best of our knowledge, the Feedback Instrument for Rescue forces Education (FIRE, Schulte and Thielsch 2019) questionnaire is currently the only systematically developed and validated published general evaluation tool in the context of firefighter education programs. The instrument is directed at future group and platoon leaders who can use it to rate the quality of firefighter leadership trainings in which they participate. The FIRE questionnaire consists of the six scales *trainers' behavior*, *structure*, *overextension*, *group*, *competence* and *transfer*.

To better understand what constitutes a validated questionnaire, the concept of validity and the aim of the present research is explained in more detail in the following section.

1.4 Validation of Questionnaires and Aim of the Present Study

In general, validity is understood as the measure of accuracy with which a questionnaire measures the concept that it intends to measure (Goldstein and Simpson 2002). Goldstein and Simpson (2002) propose to assess validity by examining three different facets of validity: Content validity, construct validity and criterion validity. First, content validity examines the extent to which a questionnaire represents the characteristic to be measured (Haynes, Richard, and Kubany 1995). It can be achieved by involving experts from the respective field in the development of the construct's definition and items for the questionnaire (American Educational Research Association et al. 2014). Next, construct validity can be confirmed if relationships between the behavior in the test situation and underlying psychological characteristics can be demonstrated (Goldstein and Simpson 2002). As such, construct validity can include the investigation of a questionnaire's factorial structure: Factorial validity exists if there is a good fit between the theoretical model on which a questionnaire is based on and the empirical data obtained with it (Guilford 1946; Thompson and Daniel 1996). Yet, as described by Campbell and Fiske (1959), construct validity can be divided into two subtypes: convergent validity and divergent validity. A questionnaire is regarded as being convergently valid if high correlations with construct-related questionnaires can be proven. The idea is to test whether constructs that are expected to be related are, in fact, related. In turn, a questionnaire can be described as being divergently valid if only low correlations with other independent constructs are found. Thus, it is the idea to check whether constructs measured with other questionnaires that are not expected to be related do not, in fact, have any relationship. The third aspect of validity, criterion validity, aims at demonstrating that questionnaire scores are related to or, more precisely, predict concrete real-life outcomes. Again, two subtypes can be specified: A questionnaire is said to have concurrent validity if it can predict criteria measured at the same time (e.g., expert or global ratings), whereas it has predictive validity if it forecasts criteria measured some point after the questionnaire scores were obtained (e.g., learning or performance outcomes) (Cronbach and Meehl 1955).

The present study aims to respond to the need for a valid tool to evaluate firefighter basic trainings[1]. Therefore, our first goal is to systematically develop a profound theoretical basis for appropriate items (Study 1). Further, we aim at a piloting these items as well as comprehensively validating the developed

[1] Hypotheses for validation were preregistered prior to data collection (https://aspredicted.org/tv95j.pdf). The questionnaire referred to as FIRE-G in the preregistration was later renamed to FIRE-B.

questionnaire using different samples. To examine the questionnaire's factorial validity, we conduct exploratory and confirmatory factor analyses. Beyond that, to investigate convergent and divergent construct validity as well as concurrent criterion validity, we make use of correlative validation methods (Study 2 & 3).

2 Study 1

In Study 1, we determined the factors related to the success and quality of a good firefighter basic training program at municipal and district level as theoretical basis for the item construction.

2.1 Method

In Study 1, in a qualitative research approach we asked N = 21 experts (n = 13 trainees, n = 4 trainers, n = 4 persons mainly having a managing function in a fire brigade school) what they personally consider to be important aspects of good training at the municipal and district level. All participants were German and male. Ages ranged from 18 to 31 years (M = 22.46; SD = 3.60) for the trainees, from 25 to 49 years (M = 34.00; SD = 10.52) for the trainers, and from 33 to 46 years (M = 37.50; SD = 5.80) for the persons with a managing function. The survey was answered with regard to trainings for professional fire brigades by 48% of the participants, whereas 52% answered it with regard to trainings for voluntary fire brigades. Factors related to the success and quality of basic trainings were recorded in an online survey by means of the Critical Incident Technique (CIT, Flanagan 1954). The CIT is a qualitative research method based on expert surveys. It is frequently used as an effective exploratory tool to better understand specific human activities or to get an information base for further research (Butterfield et al. 2005). The idea is to let experts clearly and comprehensibly describe critical situations (i.e., situations including particularly effective or ineffective behaviors) from which critical categories or items can be derived. Additionally, participants were able to directly name important aspects of a good training in four open questions: The first concerned aspects of good training in the fire brigade. The second asked for characteristics of a good trainer and his or her teaching style. The third asked about relevant framework conditions, and the fourth asked for ways in which trainees themselves can contribute to good training. The study was available online from the end of August to October 2017. Participation was voluntary and anonymous. As compensation, the respondents received a result report after Study 1 was completed.

2.2 Results and Discussion

Based on a qualitative content analysis (Mayring 2015), the experts' statements were clustered into categories of good firefighter basic trainings by two independent evaluators. The analyses led to eight categories of a good basic training at a fire brigade: *Didactics*, *motivation & engagement*, *personality*, *content & methods*, *structure & organization*, *materials & facilities*, *group* and *achievement of learning objectives* (see Table 1). A standard procedure in questionnaire construction is the creation of a large item pool for a first draft of a questionnaire (e.g., Kline 2000; Nunnally 1975; Rossi, Wright, and Anderson 2013). This makes it possible to remove unsuitable items after subsequent item analyses while still retaining a sufficiently large number of items. Thus, in accordance with the identified

Table 1: Category system: characteristics of a good basic training.

Category	Examples	CIT	E	C	P
Didactics	Good instructions, technically good instructors, answering questions	15 (17%)	17 (19%)	7 (10%)	38 (37%)
Motivation & engagement	Fun in training, good preparation, self-motivation of the trainers	10 (11%)	8 (9%)	4 (6%)	12 (12%)
Personality	At eye level / as partners, calm, serious appearance in the right moments	8 (9%)	5 (5%)		52 (51%)
Content & methods	Realistic exercises, proper debriefing, practical work	15 (17%)	19 (21%)	6 (9%)	
Structure & organization	Sufficient time, well-developed curriculum, structured process	12 (13%)	9 (10%)	20 (30%)	
Material & facilities	Equipment of the training facilities, latest technology, good learning materials	9 (10%)	16 (18%)	16 (24%)	
Group	Respectful interaction, team spirit, willingness to learn	7 (8%)	13 (14%)	12 (18%)	
Achievement of learning objectives	Direct applicability in practice, high learning effect	10 (11%)	1 (1%)		
Other		4 (4%)	3 (3%)	2 (3%)	

Note. Shown is the number of entries that could be assigned to the respective category for the four different questions. The percentages are given in brackets. If there was no information, this category was not mentioned in the question. CIT = Critical Incident Technique (one entry per category was counted for each situation), E = Education (question about characteristics of a good education), C = Conditions (question on general conditions of good training), P = Person (question about characteristics of a good trainer).

categories, a pool of 51 items, which comprehensively and fully depicted the mentioned success-critical aspects, was created for a preliminary version of the questionnaire (see Table A1 in the online supplement at https://doi.org/10.5281/zenodo.3948173).

The results of Study 1 indicate that the categories experts described as being important for good firefighter basic trainings are similar to the success factors for firefighter leadership trainings (FIRE questionnaire). However, the results also revealed aspects that are not covered by the existing FIRE questionnaire (Schulte and Thielsch 2019), such as specific teaching methods and outcomes, motivational aspects, personality aspects of the trainers and required materials and facilities. Beyond that, the results of Study 1 revealed parallels to the characteristics of good teaching at universities, as many items were similar to items that have been described in established evaluation instruments used at universities (TRIL, Gläßer et al. 2002; MFE-ZGr, Grötemeier and Thielsch 2010a; MFE-ZHa, Grötemeier and Thielsch 2010b; MFE-V, Hirschfeld and Thielsch 2009; HILVE, Rindermann and Amelang 1994; FEVOR, FEPRA, Staufenbiel 2000; MFE-Sr, Thielsch and Hirschfeld 2012).

Based on these findings, we developed an adapted questionnaire for the evaluation of firefighter basic trainings. Thus, the initial item pool consisted of 51 questions newly created based on interview results of Study 1 as well as items originating from existing instruments that were adapted to the technical context of the fire brigade as well as to the training context at municipal and district level (see Table A1 in the online supplement at https://doi.org/10.5281/zenodo.3948173). The resulting questionnaire was named Feedback Instrument for Rescue forces Education – Basic education (FIRE-B). The aims of the following studies were thus to shorten this draft version by removing items proven to be unsuitable in item analyses, to facilitate practical application and to check the psychometric quality of the resulting final instrument.

3 Study 2

In Study 2, the preliminary questionnaire version developed in Study 1 was piloted with members of various fire brigades in Germany. The aims of this study were to shorten the FIRE-B draft by selecting items on the basis of the descriptive item parameters, to uncover the factor structure via an exploratory factor analysis, and to carry out initial validations.

3.1 Method

3.1.1 Sample

The sample for the item and exploratory factor analysis consisted of $N = 257$ persons from Germany (229 men, 26 women, 2 not specified) with an age range of 16 to 51 years ($M = 25.75$; $SD = 6.48$). An overview of the initial sample and the exclusion criteria applied is given in Figure A1 in the appendix. Of the final sample, 37% was made up of (former) apprentices in training to become professional firefighters in the fire brigade, and 63% was made up of (former) apprentices in troop man or troop leader training in the voluntary fire brigade. In professional fire brigades, 39% ($n = 94$) of the trainees had completed their training. With regard to trainees in volunteer fire brigades ($n = 163$), 12% were in troop man training, 47% were between troop man training and troop leader training, 5% were in troop leader training and 36% had already completed troop leader training.

3.1.2 Measures and Procedure

Study 2 was conducted as an online survey using the software EFS Survey (provided by the Questback GmbH 2018). The main component of the survey was the set of 51 items developed in Study 1. Participants indicated their agreement with the items on a seven-point Likert scale (from 1 = *strongly disagree* to 7 = *strongly agree*) with a *denial option* to indicate that the respective item cannot be answered meaningfully. Another component of the survey included items for the initial validations. Firstly, two items of global judgment (subjective learning success (Gediga et al. 2000), global grading on a school grading scale (cf. FEVOR/FESEM, Staufenbiel 2000)) served as indicators for criterion validity. Secondly, mood served as a criterion for divergent validity. To measure the participants' mood, the five-level smiley scale by Jäger (2004) was used. A series of two studies by Jäger (2004) provided evidence for this scale's unidimensionality and equidistance and showed high correlations with the German version of the PANAS scale ($0.75 \leq r \leq 0.89$). Finally, one additional scale was measured that is not pertinent to the present study. The median response time for completing the entire survey was 12 min and 22 s.

The study was available online from January to March 2018. Participation in the survey was voluntary, anonymous and possible via an access link. It could be carried out on computers or other internet-enabled devices and consisted of three different sections (see Figure A2 in the appendix). As compensation, the respondents received a result report after Study 2 was completed.

3.1.3 Statistical Analyses

All data analyses of Study 2 were performed with the program IBM SPSS Statistics - Version 24. Before starting the analyses, the inverted items were reversed so that a high value for all items is equivalent to a good evaluation of the training. Missing values of the training evaluation items (those for which participants had selected the denial option) were imputed using the expectation maximization algorithm. Missing values occurred in 18% of the respondents. Overall, only 0.5% of the data were missing.

The 51 items of the preliminary questionnaire version were evaluated primarily with regard to their distribution, response rates and item intercorrelations as well as on the basis of their correlation with the mean value of eight items on the self-assessed acquisition of competence (these included all seven items on the above-mentioned scale acquisition of competence as well as one item that was not included in the final instrument due to content redundancy). The latter correlation was regarded as an indication of how relevant the items were regarding content and practicability in the feedback process. On the basis of these analyses, an initial item selection was made. The reduced set of items (see Section 3.2.1) was included in an explorative factor analysis (EFA, main axis analysis with oblique promax rotation) in order to uncover the factor structure underlying the data and to further reduce the pool of items. Finally, bivariate correlations between the scales of the draft questionnaire and the mentioned validation measures were calculated to exploratorily assess construct and criterion validity.

3.2 Results and Discussion

3.2.1 Item Selection

In the first selection phase, the item set was reduced to 44 items: One item was excluded due to an unfavorable answer distribution in the histogram and a low correlation with the self-assessed competence acquisition. Another item was excluded due to a high item intercorrelation. In addition, three items concerning overextension in the training (originating from the FIRE scale for leadership training evaluation, Schulte and Thielsch 2019) seemed to be somewhat irrelevant for basic training evaluation: They had comparatively high mean values and low standard deviations, did not correlate with the self-assessed acquisition of competence and were assessed as less relevant by an expert from a fire brigade school. Thus, those three items were excluded. Lastly, two items were excluded as they had comparatively high mean values and low standard deviations.

Additionally, both items could possibly be problematic for the feedback process, as they referred to stable personality traits of the trainers. For a detailed description of the reasons for exclusion, see Table A1 in the online supplement at https://doi.org/10.5281/zenodo.3948173. See Table A2 in the online supplement at https://doi.org/10.5281/zenodo.3948173 for the final FIRE-B-items with an indication of the original items that served as basis.

3.2.2 Exploratory Factor Analyses

The factor number was determined based on the Kaiser criterion (eigenvalues > 1; Guttman 1954; Kaiser and Dickmann 1959), the scree plot (Cattell 1966) and the minimum average partial test (MAP test, Velicer 1976). The Kaiser criterion argued for a solution with seven factors, while a visual inspection of the scree plot as well as the original version of the MAP test (Velicer 1976) suggested six factors, and the revised version of the MAP test (Velicer, Eaton, and Fava 2000) indicated five factors. Because the Kaiser criterion generally tends to overestimate the number of factors (Moosbrugger and Schermelleh-Engel 2008) and because only solutions with five or six factors seemed conceptually meaningful, subsequent content-related deliberations finally led to a solution with six factors.

Based on the results of the EFA, a second item selection was carried out, which further reduced the item pool from 44 to 30 items. A total of eight items were excluded due to their loading pattern (and in some cases due to additional criteria): Five items were removed due to low loadings < 0.5 or double loadings > 0.3, two items were excluded due to comparatively low loadings ≤ 0.54 and low correlations with the mean value of self-assessed competence acquisition ($r \leq 0.27$), and one item was excluded due to a double loading (0.32) and a high item intercorrelation ($r = 0.68$). In contrast, five items with low loadings < 0.4 and/or double loadings ≥ 0.3 were considered relevant in terms of content due to the high correlation of $r \geq 0.5$ with the mean self-assessed competence acquisition. Accordingly, these items were retained. The content relevance of the items thus represented the more important decision criterion. In addition, two items were excluded due to low to moderate correlations with the criterion ($r \leq 0.34$), and four items were excluded due to high item-total correlations ($r_{\text{it}} \geq 0.67$) and high item intercorrelations ($r \geq 0.65$), which could indicate the content redundancy of the items. In this case, high item-total correlations were chosen as a reason for exclusion in order to obtain factors that cover as many different facets of the construct as possible. For a detailed description of the reasons for exclusion, see Table A1 in the online supplement at https://doi.org/10.5281/zenodo.3948173.

3.2.3 Extracted Scales and Their Interpretation

Subsequently, following the recommendations of Beavers and colleagues (2013), a new EFA (main axis analysis with oblique promax rotation) with the final 30 items was calculated to obtain the factor structure of the optimized solution. This EFA led to a solution with six factors. In terms of content, factor 1 (*competence*) represents the acquisition of competences in training. Factor 2 (*structure & didactics*) concerns the structure of training and the didactic abilities of the trainers. Factor 3 (*materials & facilities*) describes the quality and availability of materials and facilities. Factor 4 (*support & encouragement*) represents the support and promotion of the trainees by the trainers. Factor 5 (*group*) refers to the group of trainees and, finally, factor 6 (*practice*) concerns the practical orientation of training. Thus, the instrument consists of the outcome scale for competence acquisition (factor 1), which focuses on the consequences or effects of training, and of five process scales (factors 2–6), which make it possible to assess the execution and implementation of a training (cf. Blanchard and Thacker 2010).

The scales and items of the final questionnaire as well as the corresponding item statistics are presented in Table 2.

3.2.4 Initial Validation

Concerning a first validation, in Study 2 participants' mood was not strongly related to the assessment of the training ($0.23 \leq r \leq 0.34$, $p < 0.001$). Comparable correlations were found in the validation of the related FIRE questionnaire (Schulte and Thielsch 2019). Consequently, the evaluation results can be distinguished from the participants' mood, initially indicating divergent construct validity. Regarding a first criterion validation, the evaluation results on the five process scales of the FIRE-B in Study 2 showed average to high correlations with the subjective learning success ($0.33 \leq r \leq 0.53$, $p < 0.001$) as well as with the global grading of the training ($0.40 \leq r \leq 0.72$, $p < 0.001$). See Table 4 for single values. As these first validation results seem promising, further in-depth analyses were necessary and performed in the following study, Study 3.

4 Study 3

In Study 3, confirmatory factor analyses (CFA) were carried out using a different sample to verify the questionnaire's factor structure proposed in Study 2. In addition, bivariate correlations between the scales and selected validation measures served to broadly assess construct and criterion validity.

Table 2: Mean values (*M*), standard deviations (*SD*), item-total correlations and factor loadings of the 30 items in Study 2.

No.	Dimension	Item	*M*	*SD*	Item-total correlation	Factor loading
1a.	Structure & didactics	I think the training was clearly structured.	5.09	1.54	0.72	0.97
1b.		I could always follow the training process during the training.	5.33	1.45	0.65	0.84
2.		I think the training gave a good overview of the subject area.	5.58	1.16	0.68	0.57
3.		The trainers explained complicated things in an understandable way.	5.60	1.10	0.68	0.50
4.		I find that the trainers were able to adapt the lesson/exercises flexibly to my knowledge and skills.	5.04	1.56	0.68	0.43
5.		The trainers prepared the topics in an interesting way.	5.35	1.22	0.69	0.42
6.		The trainers knew the contents of the training very well.	5.86	1.15	0.67	0.42
7.	Support & encouragement	The trainers had an open ear for the problems of the trainees.	5.77	1.25	0.73	0.96
8.		I think the trainers were interested in my learning success.	5.74	1.32	0.70	0.73
9.		The trainers motivated me to get involved.	5.40	1.44	0.71	0.59
10.		The trainers provided peace and quiet in difficult situations.	5.63	1.27	0.69	0.57
11.		The trainers were able to supplement the training well with their own experience of working in the field.	5.80	1.40	0.57	0.31
12.	Group	The other trainees played an active role.	5.52	1.16	0.68	0.76
13.		The other trainees were motivated to participate.	5.58	1.07	0.67	0.75
14.		I think there was good cohesion among the trainees	5.95	1.17	0.64	0.72

Table 2: (continued)

No.	Dimension	Item	*M*	*SD*	Item-total correlation	Factor loading
15.	Practice	I think the practical exercises were realistic.	5.09	1.42	0.60	0.55
16.		I think the exercises were discussed sufficiently.	5.48	1.43	0.64	0.55
17.		There were plenty of opportunities to practice what I had learned.	5.16	1.49	0.63	0.53
18.	Materials & facilities	Rooms and training facilities were sufficient and available in good quality.	5.60	1.44	0.63	0.77
19.		The quality of the media used in theoretical teaching (presentations, films, sketches, etc.) was good.	5.40	1.42	0.65	0.71
20.		Teaching materials (books, exercise books, worksheets, etc.) for the theoretical lessons were sufficient and available in good quality.	5.33	1.71	0.59	0.68
21.		Vehicles, devices and equipment were sufficient and available in good quality.	5.63	1.52	0.54	0.67
22.		Exercise and visual materials (dummies, scrap cars, etc.) were sufficiently available.	5.38	1.64	0.58	0.59
23.	Competence	Training makes it easier for me to react to dangers.	5.74	1.14	0.70	0.82
24.		The training enables me to handle fire-fighting equipment and vehicles much better.	5.80	1.41	0.70	0.79
25.		The training makes me feel very well prepared for my next assignment.	5.45	1.34	0.70	0.77
26.		Through the training, I know my personal limits better.	5.29	1.59	0.63	0.73
27.		Through the practical exercises during training, I have gained the necessary confidence to take on my tasks in the fire brigade in various missions.	5.65	1.24	0.66	0.72
28.		The training improved my ability to work in a team.	5.47	1.36	0.60	0.67
29.		The training enabled me to significantly expand my theoretical knowledge.	5.67	1.46	0.62	0.57

Note. Item 1b was excluded from the final FIRE-B in the CFA of Study 3. Scale from 1 = *strongly disagree* to 7 = *strongly agree*, 248 ≤ *N* ≤ 257 (mean values and standard deviations) or *N* = 257 (item-total correlations and factor loadings). The dataset with imputed values was used to calculate the item-total correlations and factor loadings.

4.1 Method

4.1.1 Sample

The sample in Study 3 consisted of N = 451 (414 men, 37 women) German firefighters aged between 18 and 63 years (M = 34.02; SD = 9.92). An overview of the initial sample and the exclusion criteria applied is given in Figure A3 in the appendix. The final sample size meets the minimum requirement of N = 400 based on the recommendation for CFAs with three indicator variables per factor and loadings of 0.6 (Gagne and Hancock 2006). It is also suitable for the calculation of correlation coefficients, which, according to Schönbrodt and Perugini (2013), are sufficiently robustly estimated from a sample size of about 250 persons, assuming medium effect sizes. The participants were asked to assess the training they currently completed or had last completed at the time of the survey. Of those questioned, 24% assessed their training for becoming professional firefighters. Furthermore, 19% assessed the troop man training and 33% assessed the troop leader training within the voluntary fire brigade. Other trainings provided at municipal or district level were assessed by 24% of the sample.

4.1.2 Measures

In addition to the items of the FIRE-B (see Table 2), scales from other well-established evaluation instruments as well as from FIRE validation studies were used for the investigation of convergent construct validity. The participants' current mood, their level of education and their experience served as divergent criteria. Third, participants' overall satisfaction with the training and their learning success were assessed for criterion validation. Unless specified differently, participants indicated their agreement with the items on a seven-point Likert scale (from 1 = *strongly disagree* to 7 = *strongly agree*). An additional denial option (*unanswerable*) could be ticked if participants perceived an item as not applicable, for instance, because they had never been involved in the activity described in an item (cf. Chyung et al. 2017). Table A3 in the online supplement at https://doi.org/10.5281/zenodo.3948173 gives an overview of all validation items of Study 3 as well as their source and response format.

4.1.2.1 Items For Construct Validation

Scales from well-established German evaluation instruments for higher education (HILVE II (Rindermann 2009); TRIL (Gläßer et al. 2002); FEPRA (Staufenbiel 2000)) were used for convergent construct validation of the four scales *structure & didactics*, *support & encouragement*, *materials & facilities* as well as *competence*. In past studies, the HILVE could be assessed as a very stable measure over time that correlates with performance criteria and external-rater judgments (Rindermann 1994). Likewise, the results from the TRIL in a program evaluation also correlated with the judgments of external observers (Gollwitzer and Schlotz 2003). For the validation of the scale *group*, three items were used which were also used for the validation of the group scale from the FIRE questionnaire (Schulte and Thielsch 2019). The validation of the scale *practice* was carried out with four items from the validation of the transfer scale from the FIRE questionnaire (Schulte and Thielsch 2019).

For divergent construct validation, the current mood was again (as in Study 2) measured with a five-point smiley scale (Jäger 2004), and the level of education was assessed by the highest level of school-leaving certificate achieved. Additionally, the variable experience of participants was assessed by an inquiry about previous experience, for example regarding previous membership in a youth fire brigade as well as the monthly mission experience.

4.1.2.2 Items for Criterion Validation

To investigate the first criterion, participants' overall satisfaction, three single-item measures were used: The item "All in all, the attendance of this training was worthwhile for me" was taken from the TRIL (Gläßer et al. 2002), the item "I would recommend the training to a good friend" was used according to the MFE-Sr (Thielsch and Hirschfeld 2012), and the last item asked the participants to rate the training on a school grade scale (1 = *very good*; 6 = *insufficient*) (FEVOR/FESEM, Staufenbiel 2000). Learning success served as second criterion and was first measured by the item "I learned a lot during my training" taken from the KIEL (Gediga et al. 2000). Second, participants were asked whether they had passed the evaluated basic training.

4.1.3 Procedure

For data collection, an online survey was created using the survey software EFS Survey (provided by the Questback GmbH 2018). Participation in the survey was voluntary, anonymous and possible via an access link. It could be carried out on computers or other internet-enabled devices and consisted of three different sections (see Figure A4 in the appendix). The median response time for completing the entire survey was 12 min and 59 s. The study was available online from July to September 2018. Again, as did Study 2, it aimed at interviewing German firefighters. As compensation, the respondents received a result report after Study 3 was completed. Moreover, they were able to take part in a raffle for an annual subscription to a firefighter-specific magazine.

4.1.4 Statistical Analysis

The statistical data analyses were carried out with RStudio (RStudio Team 2016, Version 1.1.456). In particular, the packages lavaan (Rosseel 2012, version 0.6.3), plyr (Wickham 2011, version 1.8.4), psych (Revelle 2018, version 1.8.10) and semPlot (Epskamp 2017, version 1.1) were used. A robust maximum-likelihood estimator with Huber–White standard errors and a scaled test statistic asymptotically comparable to the Yuan–Bentler test statistic (MLR) was used to calculate the confirmatory factor analysis (cf. Steinmetz 2015). In addition, bivariate correlations between the scales of the questionnaire and selected validation criteria were calculated to assess construct and criterion validity.

4.2 Results and Discussion

4.2.1 Confirmatory Factor Analysis

A confirmatory factor analysis (CFA) was conducted to review the factor structure proposed in Study 2. Modification indices indicated a high correlation between item 1a (“I think the training was clearly structured”) and item 1b (“I could always follow the training process during the training”) of the scale *structure & didactics*. After considerations of content, both items were judged to be redundant. Since item 1a seemed more global and understandable, item 1b was removed from the model. According to Schermelleh-Engel, Moosbrugger, and Müller (2003), the model fit for the final FIRE-B with 29 items on six scales can be classified as good (RMSEA = 0.05; SRMR = 0.04) to acceptable (CFI = 0.95; TLI = 0.95). The χ^2-test was significant (χ^2 (362) = 585.12, p < 0.001), which is common for large samples (Tanguma 2001). However, related to the degrees of freedom, the χ^2-value is good (χ^2/df = 1.62). Results thus provide support for a six-factorial structure. Figure 2 illustrates the specified model including all path coefficients. The intercorrelations of the scales are medium to high and can be found in Table A1 in the appendix.

In sum, the results of the CFA are in line with the findings of the EFA in Study 2, confirming that the items of the FIRE-B load on six distinct factors (*structure & didactics, support & encouragement, group, practice, materials & facilities, competence*). Overall, the six-dimensional questionnaire structure demonstrates that various quality factors contribute to good firefighter basic trainings and that firefighters should have a wide range of skills for successful action (cf. Kleinmann et al. 2010).

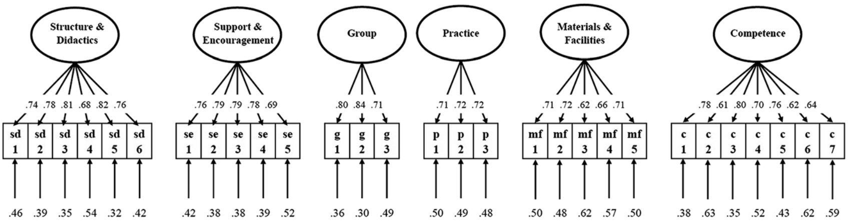

Figure 2: Results of confirmatory factor analysis. Standardized coefficients are reported.

4.2.2 Descriptive Statistics and Reliability

Table A4 in the online supplement at https://doi.org/10.5281/zenodo.3948173 contains means, standard deviations and correlations for all measures used in Study 3. Table 3 reports an overview of the reliability estimates and the associated measurement model tests for all FIRE-B scales based on the data of Study 2 and 3. Cronbach's α should at best assume values between 0.70 and 0.90 (Tavakol and Dennick 2011). Likewise, ω_H can be classified (Schweizer 2011). Accordingly, all scales reach a good to acceptable level of reliability. To this extent, the coefficients are comparable with those of other established evaluation instruments (e.g., HILVE II, Rindermann 2009).

4.2.3 Correlation Analysis

4.2.3.1 Convergent Construct Validity

All FIRE-B scales showed consistently high positive correlations with their corresponding validation scales (see Table 4): $r = 0.81$, $p < 0.001$ between the scale *structure & didactics* (FIRE-B) and the scale *structure & didactics* (TRIL); $r = 0.88$, $p < 0.001$ between the scale *support & encouragement* (FIRE-B) and the scale

Table 3: Reliability coefficients and measurement model tests for all FIRE-B scales.

Scale	Study 2	Study 3				
	Cronbach's α	Cronbach's α	ω_H	$\Delta\chi^2$	*df*	*p*
Structure & didactics	0.88	0.89	0.89	19.05	7	0.008
Support & encouragement	0.86	0.88	0.88	18.90	5	0.002
Group	0.81	0.83	0.83	15.16	3	0.002
Practice	0.78	0.76	0.76	1.97	3	0.830
Materials & facilities	0.81	0.80	0.80	5.22	5	0.390
Competence	0.87	0.87	0.88	24.98	7	<0.001

Note. $N_{\text{Study 1}} = 257$, $N_{\text{Study 2}} = 451$. The χ^2-difference test compares essentially tau-equivalent measurement models with congeneric measurement models.

lecturer management (HILVE II); $r = 0.52$, $p < 0.001$ between the scale *group* (FIRE-B) and the validation item for the FIRE scale group; $r = 0.78$, $p < 0.001$ between the scale *practice* (FIRE-B) and the validation items for the FIRE scale mission exercises; $r = 0.79$, $p < 0.001$ between the scale *material & facilities* (FIRE-B) and the question of quantity and quality of equipment and materials (FEPRA); $r = 0.77$, $p < 0.001$ between the scale *competence* (FIRE-B) and the scale *learning-quantitative* (HILVE II); $r = 0.75$, $p < 0.001$ between the scale *competence acquisition* (FIRE-B) and the scale *learning-qualitative* (HILVE II). In addition, all FIRE-B scales correlated positively with scales for the validation of other FIRE-B scales. However, these were (in some cases significantly) lower (e.g., $r = 0.33$, $p < 0.001$ between the scale of *competence acquisition* [FIRE-B] and the question of the quantity and quality of the equipment and materials [FEPRA]). Overall, the results support the convergent validity of the FIRE-B.

Table 4: Correlations between FIRE-B scales and convergent (structure & didactics to learning – qualitative), divergent (mood to mission experience) and criterion-related (school grade to learning success) validation indicators in Study 2 and 3.

Variable	S & D	S & E	G	P	M & F	C
Structure & didactics (TRIL)	0.81***	0.72***	0.44***	0.62***	0.60***	0.57***
Lecturer management (HILVE II)	0.76***	0.88***	0.52***	0.54***	0.46***	0.55***
Group (FIRE validation)	0.39***	0.44***	0.52***	0.32***	0.19***	0.43***
Mission exercises (FIRE validation)	0.59***	0.55***	0.41***	0.78***	0.48***	0.62***
Equipment & material (FEPRA)	0.50***	0.38***	0.27***	0.46***	0.79***	0.31***
Learning – quantitative (HILVE II)	0.57***	0.54***	0.40***	0.42***	0.35***	0.78***
Learning – qualitative (HILVE II)	0.56***	0.52***	0.40***	0.45***	0.34***	0.75***
Mood	0.21*** (0.31***)	0.23*** (0.34***)	0.16*** (0.26***)	0.17*** (0.32***)	0.14** (0.23***)	0.21*** (0.33***)
Level of education[a]	−0.11*	−0.12**	−0.06	−0.04	−0.11*	−0.17***
Previous experience	−0.09	−0.12*	−0.04	−0.06	−0.08	0.01
Mission experience	0.00	−0.01	−0.04	0.00	−0.06	0.09*
School grade (*r*)	0.75*** (0.72***)	0.64*** (0.60***)	0.42*** (0.40***)	0.61*** (0.58***)	0.52*** (0.54***)	0.58*** (0.62***)
Overall satisfaction	0.67***	0.64***	0.47***	0.54***	0.42***	0.69***
Learning success	0.63*** (0.53***)	0.58*** (0.51***)	0.41*** (0.44***)	0.49*** (0.51***)	0.40*** (0.33***)	0.75*** (0.75***)

Note. $^*p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$ (two-sided). [a]Spearman rank correlation. $N = 451$ (Study 3), $N = 257$ (Study 2). S & D = Structure & didactics, S & E = Support & encouragement, G = Group, P = Practice, M & F = Materials & facilities, C = Competence, *r* = recoded. In brackets are the values of the initial validation in Study 2.

4.2.3.2 Divergent Construct Validity

According to the assumption, in Study 3 the participants' current mood correlated significantly but only to a small extent with the FIRE-B scales ($0.12 \leq r \leq 0.23$, $p < 0.01$). For the educational level of the participants, small and only partially significant correlations with the FIRE-B scales ($-0.17 \leq r \leq -0.05$, p between <0.001 and 0.31) were found. With regard to the length of previous experience in the work of fire brigades, there were consistently small, sometimes insignificant correlations ($-0.10 \leq r \leq -0.02$; $0.09 \leq p \leq 0.74$). Similarly, the monthly experience in professional and voluntary fire brigades was only slightly related to the assessment of the FIRE-B scales ($-0.05 \leq r \leq 0.09$, $0.05 \leq p \leq 0.95$). All results point to divergent validity of the FIRE-B. See Table 4 for detailed results.

4.2.3.3 Criterion Validity

Study 3 showed moderate to large correlations of the FIRE-B scales with the school grade awarded ($0.42 \leq r \leq 0.75$, $p < 0.001$). In addition, there were moderate to large highly significant relationships with the other items measuring overall satisfaction ($0.42 \leq r \leq 0.69$, $p < 0.001$). With regard to learning success, moderate to high highly significant correlations between the FIRE-B scales and the subjective learning success were noted ($0.41 \leq r \leq 0.75$, $p < 0.001$). The correlation between the evaluation results and the passing of the examination could not be meaningfully examined due to the lack of variance in the data. Thus, 99% of the respondents passed the examination directly, and only 1% passed after a subsequent examination. All other results support the criterion validity of the FIRE-B. See Table 4 for detailed results.

5 General Discussion

High-quality basic training is critical to the development of firefighters' knowledge and skills. Only with such a training firefighters can successfully perform their demanding tasks and, at the same time, be prepared for possible negative physical (e.g., work-related injuries, see Moore-Merrell et al. 2008) or psychological (e.g., burnout and PTSD, see Katsavouni et al. 2016) consequences. In this regard, the newly created evaluation questionnaire for firefighter basic trainings (FIRE-B) addressed the lack of a valid and scientifically based tool to assess these trainings. The questionnaire was developed and validated in a series of three studies. Results clearly show that the FIRE-B meets all central psychometric standards and, therefore, can and should be used.

Through these studies, we ensured high content validity, meaning that the constructed item set represents all relevant facets of firefighter basic trainings:

First, an expert survey about the characteristics of a good firefighter education served as basis for the development of the item pool (Study 1). Second, in Study 2 only few participants made additions to the questionnaire despite explicit requests. Moreover, we found further clear indications for validity of the FIRE-B: The factor structure proposed in Study 2 was confirmed with an independent sample in Study 3, indicating factorial validity. Beyond that, bivariate correlations in Study 2 and Study 3 served to investigate convergent and divergent construct validity as well as criterion validity. The patterns of correlations between the FIRE-B and the validation scales clearly support the assumption that the FIRE-B measures the intended content. Thus, the results consistently confirmed the validity of the FIRE-B. In addition, the internal consistencies of the scales can overall be regarded as sufficient to good (see Table 3), which is especially promising because most of the scales are brief. Therefore, applying the FIRE-B will lead to reliable results. Reliability will be further ensured because training evaluations will only be performed for trainings with a fairly large group of participants in order to avoid answer bias based on individual opinions (see the scoring instructions in online supplement at https://doi.org/10.5281/zenodo.3948173).[2]

Other benefits of the FIRE-B are its time efficiency, usefulness and relevance. Regarding efficiency, even though the six different scales allow for a comprehensive assessment of various aspects concerning basic training, the items can be processed in about 4 to 5 min. Regarding usefulness, no other scientifically developed, published evaluation tool for rescue service basic education is available in the literature, making the FIRE-B a highly useful tool. In addition, the evaluation questionnaire is of practical relevance, as its results can describe the current quality of fire brigade trainings and serve as a starting point to derive concrete measures for improving trainings.

In theoretical terms, the current study contributes to the question of which quality factors are relevant to evaluate firefighter basic trainings. Altogether, the identified six-factor questionnaire structure confirms that various quality factors contribute to a good training and that firefighters should have a wide range of skills for successful action (cf. Kleinmann et al. 2010). Similarly, Schulte and Thielsch (2019) concluded that evaluations of firefighter leadership trainings should be multidimensional. The final dimensions of the FIRE-B largely correspond to the scales of the FIRE questionnaire (Schulte and Thielsch 2019). For example, both the

[2] Targeted tests of objectivity did not take place. However, objectivity is a prerequisite for reliability – and reliability, in turn, is a prerequisite for validity. Conversely, this also means that the reliability values found here provide support for an objective applicability of the FIRE-B. In other words, if the FIRE-B is applied according to the given instructions, then its results will not be biased by the person who is calculating the test results.

FIRE-B and the FIRE questionnaire include the scale *group*. The scale *structure & didactics* of the FIRE-B is comparable to the scales *trainers' behavior* and *structure* in the FIRE questionnaire. In addition, the scale *competence* includes aspects of the two original FIRE scales *competence* and *transfer*. However, differences also exist, showing that basic and leadership training differ from each other and should be evaluated with different instruments. Firstly, the FIRE-B contains the scale *support & encouragement*. In contrast, the FIRE only contains a few motivation-related items on its scale *trainers' behavior*. Furthermore, the scales *practice* and *materials & facilities* are part of the FIRE-B but do not exist in the FIRE questionnaire. These aspects seem to be more important for basic trainings than for experienced participants of leadership trainings. Conversely, the scale *overextension* is part of the FIRE but not of the FIRE-B questionnaire. The low failure rate in examinations of firefighter basic trainings confirms the low relevance of this scale for the FIRE-B.

5.1 Practical Application

On a practical level, the FIRE-B for the first time offers the possibility to assess the quality of firefighter basic trainings at municipal and district level from the trainee's point of view. The differentiation of process and outcome scales according to Kirkpatrick (1979) helps to distinguish information on learning outcomes from information on possible ways to adapt the training process. In this way, the process-related items of the FIRE-B capture judgments about the trainer, the organization of the training, the group as well as about exercises, materials and facilities. The result-related items assess the extent to which the training has contributed to the acquisition of knowledge and skills. Thus, the FIRE-B provides information on the current quality of basic firefighting trainings and helps to identify possible areas for improvement within the implementation of the training and the achievement of the learning objectives. If, for example, trainees indicate not having achieved certain learning objectives, the trainer can check whether and at what point in the process there was a problem. To facilitate the questionnaire's practical application, we provide additional information in the online supplement at https://doi.org/10.5281/zenodo.3948173, including questionnaire templates (in English and German) and scoring instructions.

Generally, an evaluation with the FIRE-B should take place directly after a firefighter training course. The 29 items can be answered quickly (in our experience in about 4–5 min), and the six different scales provide a comprehensive picture of the training quality. The items are consistently to be rated on a seven-point response scale with a denial option, allowing a simple data analysis and interpretation. Depending on the evaluation context, we recommend the additional use of four optional items (see Table A5 in the online supplement at https://doi.org/10.5281/zenodo.3948173). Also, as it is important to ensure anonymous evaluation, one should not collect variables (e.g., demographic) that allow conclusions about individual persons. If scales are not applicable or if the use of all scales is considered too time consuming, single FIRE-B scales can be omitted, as they have been validated separately. However, one should not remove individual items. As the scales are already very brief, this may impair psychometric quality. Similarly, one should not change the wording of the individual questions. Exceptions are minor adjustments to ensure the questionnaire's comprehensibility and its optimal adaptation to the evaluation context.

In the analysis, mean values are calculated for the individual scales across participants and courses (see scoring instructions in the online supplement at https://doi.org/10.5281/zenodo.3948173). A high number of missing values (see scoring instructions in the online supplement at https://doi.org/10.5281/zenodo.3948173) from many participants may indicate a lack of fit of the questionnaire in the respective evaluation context. Additionally, an evaluation analysis should only take place if a minimum number of completed evaluation questionnaires are available (see scoring instructions in the online supplement at https://doi.org/10.5281/zenodo.3948173). Further, to keep effort and calculation errors to a minimum, we recommend that evaluations are analyzed using simple data evaluation programs.

Beyond that, organizers and trainers should consider the subjective nature of this type of evaluation: feedback gathered with the FIRE-B questionnaire is an opportunity for organizers and trainers to obtain important information about their own teaching activities from their trainees' points of view. Thus, organizers and trainers should meet with the trainees to discuss the results of the evaluation. In addition, the responsible organization should support the evaluation both technically and in terms of its content. Particularly, organizers should offer trainers help if evaluations repeatedly reveal areas for improvement or offer them praise for good teaching quality.

Finally, while the FIRE-B was developed specifically for the context of fire brigades, in this context it pursues a rather global approach, meaning that the questionnaire can be used to evaluate a wide range of firefighter basic trainings. If, beyond that, people wish to investigate more specific aspects as part of an evaluation, we recommend using more specific evaluation scales, such as scales for the evaluation of firefighter examinations, mission exercises or command unit trainings (Röseler et al. 2020; Schulte and Thielsch 2019; Thielsch, Busjan, and Frerichs 2018).

5.2 Limitations and Future Research

This study has some limitations but also some possibilities for future research. Both will be discussed below.

Regarding the application of the FIRE-B, one must consider that German firefighters served as the basis for constructing the questionnaire, such that the proportion of voluntary versus professional firefighters as well as the gender distributions were representative of German fire brigades (cf. Deutscher Feuerwehrverband (DFV) 2015). Fire brigades in other countries may have different distributions or different work and trainings structures. Thus, organizers in different countries should first check whether the FIRE-B validly covers the relevant areas of the respective training courses or whether it needs to be adapted. Similarly, the use of the FIRE-B questionnaire in other training contexts is conceivable, as its items do not contain a fire brigade-specific vocabulary. The content should be generally relevant within rescue training courses, such as paramedic trainings. Before applying a translation or adapted version, the validity should first be checked. Therefore, we recommend as a minimum requirement performing an expert assessment of the content validity for the intended context, and we highly welcome specific validation studies.

Another aspect to consider when using the FIRE-B is its subjectivity: The FIRE-B asks for judgments from the trainees' points of view. As indicated above, such information is primarily helpful for trainers to reflect on their own teaching activities and, if necessary, to think about possibilities for improvement. At the same time, the evaluation's subjectivity increases the risk of misunderstandings. We, therefore, recommend that organizers meet with the trainers and trainees after the evaluation to discuss the results as a group, to collect ideas for improvement, and to uncover and clarify possible misunderstandings. In this regard, an evaluation from the trainers' viewpoints might also be of interest, especially as Schulte and Thielsch (2019) showed that the judgments of trainees are, to some extent, different from those of the trainers. For example, trainers rated trainees' overextension to be higher than the trainees actually experienced. As such, considering the views of both parties could contribute to a more global assessment of the quality of firefighter basic trainings and can also benefit trainees' learning (Berlin and Carlström 2014; Childs 2005; Sommer and Njå 2011).

Additionally, the results of our study also carry some limitations. First, there was no proof of a relationship between the FIRE-B scales and passing the final exam. In view of the marginal failure rate, the question of passing the final examination for basic trainings seems to be an inadequate validation criterion. In the case of a uniform grading system, the grade achieved in a final examination might be a more suitable criterion. Furthermore, an objective proof of examination (e.g., training diplomas) could prevent possible effects of social desirability. Second, to avoid high exclusion rates, the date of the evaluated training was used as a filter criterion only in Study 2 but not in Study 3 (see Figures A1/A3 in the appendix). In Study 3, there were few significant correlations between the number of years since the start of training and FIRE-B scales. However, this finding does not affect the validity of this work, since controlling partial correlations were used for statistical data analysis in Study 3.

Regarding further research, the four-level model according to Kirkpatrick (1979) calls for methodically expanded follow-up studies investigating how to evaluate firefighter basic trainings at the levels of behavior and results. For the assessment of transfer effects at the behavioral level, future studies should collect data during subsequent work as a firefighter. Conceivable sources are again subjective self-judgments (e.g., "I successfully manage to apply the training contents in my everyday work" (Grohmann and Kauffeld 2013, 142), but also judgments from the perspectives of trainers, colleagues or superiors. In addition, objective observations of behavior based on standardized evaluation criteria can be a useful supplement (Blanchard and Thacker 2010). For the evaluation at the level of results, follow-up studies could examine whether a high quality of training leads to better organizational outcomes and whether, for example, fire brigade operations can be carried out more quickly or more successfully. Equally, future studies could check whether high-quality trainings imply a lower accident rate among firefighters themselves. A particularly challenging aspect of such an evaluation at the result level is the difficulty in attributing positive effects only to the object of evaluation and not to other unrecorded influences (Kennedy et al. 2014; Kirkpatrick 1979; Praslova 2010).

5.3 Conclusion

The present paper provides, for the first time, a systematically developed and validated evaluation questionnaire for firefighter basic trainings. The FIRE-B is a useful, efficient, reliable and valid feedback instrument. Thus, it can and should be used in rescue service education. The regular and long-term use of the evaluation questionnaire may not only contribute to the recording of current quality standards but may also reveal areas for improvement or possible changes in the basic training of firefighters. By providing a tool that may help improve the quality of firefighter basic trainings, we hope to ultimately contribute to society as a whole, as an optimized education hopefully leads to the most competent firefighters capable of responding optimally to various forms of emergencies.

Appendix

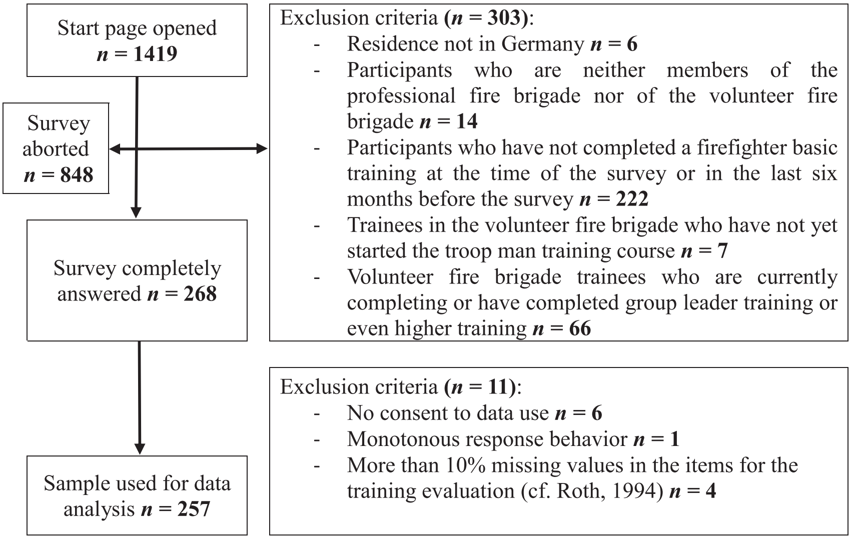
Start page opened
n = 1419
Survey aborted
n = 848
Exclusion criteria (*n* = 303):
- Residence not in Germany *n* = 6
- Participants who are neither members of the professional fire brigade nor of the volunteer fire brigade *n* = 14
- Participants who have not completed a firefighter basic training at the time of the survey or in the last six months before the survey *n* = 222
- Trainees in the volunteer fire brigade who have not yet started the troop man training course *n* = 7
- Volunteer fire brigade trainees who are currently completing or have completed group leader training or even higher training *n* = 66
Survey completely answered *n* = 268
Exclusion criteria (*n* = 11):
- No consent to data use *n* = 6
- Monotonous response behavior *n* = 1
- More than 10% missing values in the items for the training evaluation (cf. Roth, 1994) *n* = 4
Sample used for data analysis *n* = 257

Figure A1: Flowchart of the sample of Study 2 and the exclusion criteria applied. Because some participants met multiple exclusion criteria, the participants who meet each exclusion criterion do not add up to the number of participants excluded in the first step.

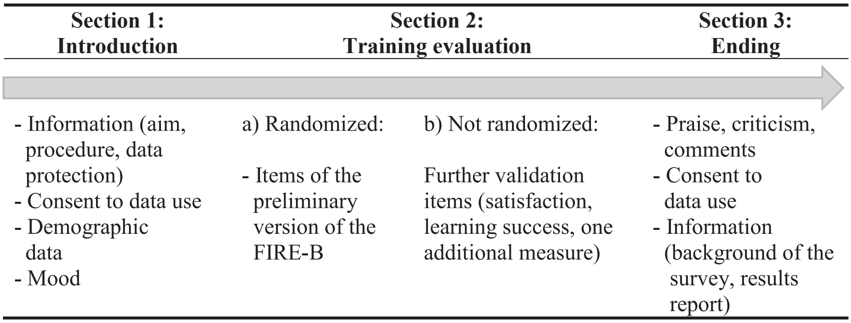
Section 1: Introduction
Section 2: Training evaluation
Section 3: Ending
- Information (aim, procedure, data protection)
- Consent to data use
- Demographic data
- Mood
a) Randomized:
- Items of the preliminary version of the FIRE-B
b) Not randomized:
Further validation items (satisfaction, learning success, one additional measure)
- Praise, criticism, comments
- Consent to data use
- Information (background of the survey, results report)

Figure A2: Procedure and contents of the online survey of Study 2.

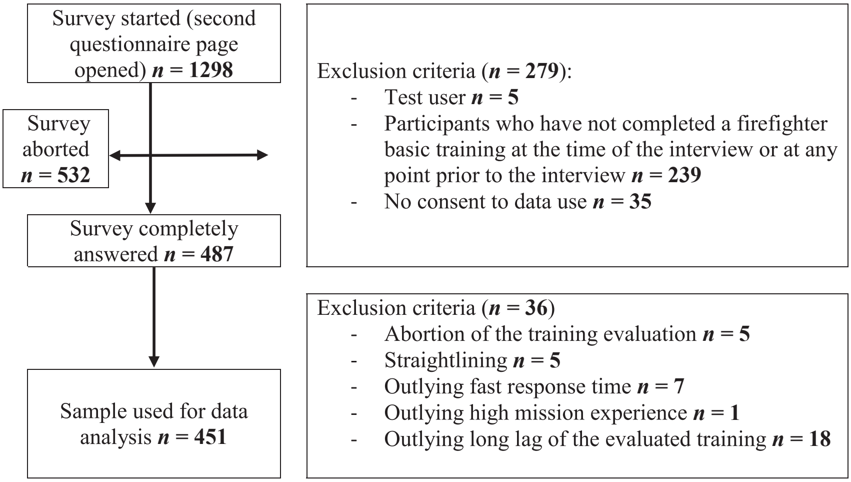
Survey started (second questionnaire page opened) ***n* = 1298**
Survey aborted ***n* = 532**
Survey completely answered ***n* = 487**
Sample used for data analysis ***n* = 451**
Exclusion criteria (***n* = 279**):
- Test user ***n* = 5**
- Participants who have not completed a firefighter basic training at the time of the interview or at any point prior to the interview ***n* = 239**
- No consent to data use ***n* = 35**
Exclusion criteria (***n* = 36**)
- Abortion of the training evaluation ***n* = 5**
- Straightlining ***n* = 5**
- Outlying fast response time ***n* = 7**
- Outlying high mission experience ***n* = 1**
- Outlying long lag of the evaluated training ***n* = 18**

Figure A3: Flowchart of the sample of Study 3 and the exclusion criteria applied.

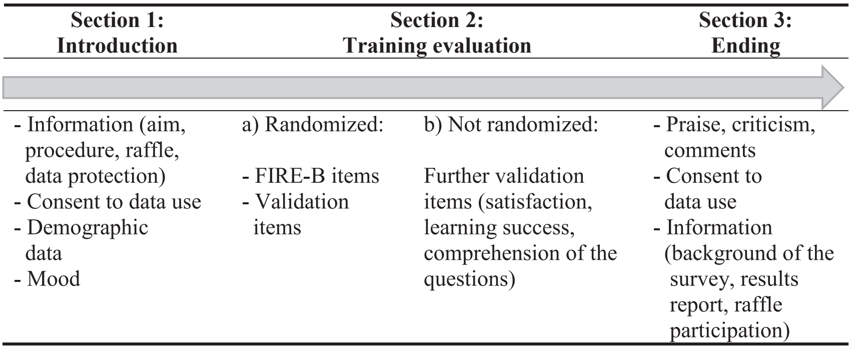
Section 1: Introduction
Section 2: Training evaluation
Section 3: Ending
- Information (aim, procedure, raffle, data protection)
- Consent to data use
- Demographic data
- Mood
a) Randomized:
- FIRE-B items
- Validation items
b) Not randomized:
Further validation items (satisfaction, learning success, comprehension of the questions)
- Praise, criticism, comments
- Consent to data use
- Information (background of the survey, results report, raffle participation)

Figure A4: Procedure and contents of the online survey of Study 3.

Table A1: Correlations between the scales of the FIRE-B.

Variable	S & D	S & E	G	P	M & F	C
Structure & didactics		0.75***	0.47***	0.62***	0.55***	0.64***
Support & encouragement	0.77***		0.51***	0.57***	0.51***	0.60***
Group	0.48***	0.52***		0.53***	0.39***	0.47***
Practice	0.61***	0.53***	0.41***		0.48***	0.60***
Materials & facilities	0.58***	0.48***	0.27***	0.56***		0.41***
Competence	0.56***	0.53***	0.40***	0.53***	0.38***	

Note. ***p < 0.001 (two-sided). Values above the diagonal are based on data from Study 2 (N = 257), values below the diagonal are based on data from Study 3 (N = 451). S & D = Structure & didactics, S & E = Support & encouragement, G = Group, P = Practice, M & F = Materials & facilities, C = Competence.

Acknowledgments: The authors would like to thank the Institut der Feuerwehr Nordrhein-Westfalen (State Fire Service Institute North Rhine-Westphalia, IdF NRW), in particular Yannick Ngatchou, Thomas Löchteken and Berthold Penkert, for their comprehensive support in creating this instrument. We would like to thank Carsten Hahn, Heinz Kamphausen and Ferdinand Hintz from the Düsseldorf Fire Brigade School for their insight into the basic training and their support in the recruitment of Studies 1 and 2. Furthermore, we thank Lisa Streppel for her great work in conducting Studies 1 and 2 as well as Guido Hertel for his valuable feedback on this project. Finally, we thank Sarah Weber and especially Celeste Brennecka for feedback on earlier versions of this manuscript.

References

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

Beavers, A. S., J. W. Lounsbury, J. K. Richards, S. W. Huck, G. J. Skolits, and S. L. Esquivel. 2013. “Practical Considerations for Using Exploratory Factor Analysis in Educational Research.” *Practical Assessment, Research and Evaluation* 18 (6): 1–13.

Berlin, Johan M., and Eric D. Carlström. 2014. “Collaboration Exercises—The Lack of Collaborative Benefits.” *International Journal of Disaster Risk Science* 5 (3): 192–205.

Blanchard, N. P., and J. W. Thacker. 2010. *Effective Training: Systems, Strategies, and Practices*. Prentice Hall: Pearson Education.

Blau, G., G. Gibson, M. Bentley, and S. Chapman. 2012. “Testing the Impact of Job-Related Variables on a Utility Judgment Training Criterion beyond Background and Affective Reaction Variables.” *International Journal of Training and Development* 16 (1): 54–66.

Brushlinsky, N. N., M. Ahrens, S. V. Sokolov, and P. Wagner. 2019. “World Fire Statistics No. 24.” https://www.ctif.org/sites/default/files/news_files/2019-04/CTIF_Report24_ERG.pdf (accessed July 23, 2020).

Buchenau, S. 2020. “So verschieden ist die Truppmannausbildung der Freiwilligen Feuerwehr [So different is the troop man training of the volunteer fire brigade].” https://www.feuerwehrmagazin.de/wissen/soverschieden-ist-die-truppmannausbildung-der-freiwilligen-feuerwehr-in-deutschland-68442 (accessed July 23, 2020).

Bukowski, R. W., and T. Tanaka. 1991. “Toward the Goal of a Performance Fire Code.” *Fire and Materials* 15 (4): 175–80.

Burke, Eugene. 1997. “Competence in Command: Recent R&D in the London Fire Brigade.” *Journal of Managerial Psychology* 12 (4): 261–79.

Butterfield, L. D., W. A. Borgen, N. E. Amundson, and A.-S. T. Maglio. 2005. “Fifty Years of the Critical Incident Technique. 1954–2004 and beyond.” *Qualitative Research* 5 (4): 475–97.

Campbell, D. T., and D. W. Fiske. 1959. “Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix.” *Psychological Bulletin* 56 (2): 81–105.

Cattell, R. B. 1966. “The Screen Test for the Number of Factors.” *Multivariate Behavioral Research* 1 (2): 245–76.

Childs, M. 2005. “Beyond Training: New Firefighters and Critical Reflection.” *Disaster Prevention and Management* 14 (4): 558–66.

Chyung, S. Y, K. Roberts, I. Swanson, and A. Hankinson. 2017. “Evidence-Based Survey Design: The Use of a Midpoint on the Likert Scale.” *Performance Improvement* 56 (10): 15–23.

Cronbach, L. J, and P. E Meehl. 1955. “Construct Validity in Psychological Tests.” *Psychological Bulletin* 52 (4): 281.

Deutscher Feuerwehrverband, E. V. 2015. “Feuerwehr-Statistik [Fire Brigade Statistics].” http://www.feuerwehrverband.de/statistik.html?&L=0Dietmar%3F1%3D1Christian (accessed July 23, 2020).

Epskamp, S. 2017. “SemPlot. Path Diagrams and Visual Analysis of Various SEM Packages’ Output. (Version 1.1) [Computer Software]”. https://CRAN.R-project.org/package=semPlot (accessed July 23, 2020).

Feuerwehr-Dienstvorschrift (FwDV) 2. 2012. Ausbildung der Freiwilligen Feuerwehren [Training of Volunteer Fire Brigades]. http://www.idf.nrw.de/service/downloads/pdf/fwdv_2_stand_01_2012.pdf (accessed July 23, 2020).

Flanagan, J. C. 1954. “The Critical Incident Technique.” *Psychological Bulletin* 51 (4): 327–58.

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.

Gediga, G., K. von Kannen, S. Frank, S. Köhne, H. Luck, and B. Schneider. 2000. *KIEL. Ein Kommunikationsinstrument für die Evaluation von Lehrveranstaltungen [A communication tool for the evaluation of courses]*. Bissendorf: Methodos.

Gläßer, E., M. Gollwitzer, D. Kranz, C. Meiniger, S. Wolff, T. Schnell, and V. Andreas. 2002. “Trierer Inventar zur Lehrevaluation [Trier inventory for teaching evaluation].” https://www.uni-wuerzburg.de/fileadmin/ext00267/Dokumente/Evaluation/Beispielfragebogen_Lehrveranstaltungsevaluation_TR.pdf (accessed July 23, 2020).

Goldstein, J. M., and J. C. Simpson. 2002. “Validity: Definitions and Applications to Psychiatric Research.” In *Textbook in Psychiatric Epidemiology*, edited by M.T., Tsuang and H., Tohen, 149–63. New York: Wiley-Liss.

Gollwitzer, M., and S. Wolff. 2003. “Das ‘Trierer Inventar zur Lehrveranstaltungsevaluation’ (TRIL): Entwicklung und Erste Testtheoretische Erprobungen [The ‘Trier Inventory for Teaching Evaluation’ (TRIL): Development and First Test Theoretical Tests].” In *Psychologiedidaktik und Evaluation* [*Psychology Didactics and Evaluation*, edited by G. Krampen and H., Zayer, 114–28. Bonn: Deutscher Psychologen Verlag.

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): 135–55.

Grötemeier, I., and M. T. Thielsch. 2010a. “Münsteraner Fragebogen zur Evaluation – Zusatzmodul Gruppenarbeit (MFE-ZGr) [Münster Questionnaire for Evaluation – Additional Module Group Work].” *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis103. https://zis.gesis.org/skala/Gr%C3%B6temeier-Thielsch-M%C3%BCnsteraner-Fragebogen-zur-Evaluation-Zusatzmodul-Gruppenarbeit-(MFE-ZGr) (accessed August 9, 2020).

Grötemeier, I., and M. T. Thielsch. 2010b. “Münsteraner Fragebogen zur Evaluation – Zusatzmodul Hausaufgaben (MFE-ZHa) [Münster Questionnaire for Evaluation – Additional Module Homework].” *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis100. https://zis.gesis.org/skala/Gr%C3%B6temeier-Thielsch-M%C3%BCnsteraner-Fragebogen-zur-Evaluation-Zusatzmodul-Hausaufgaben-(MFE-ZHa) (accessed August 9, 2020).

Guilford, Joy P. 1946. “New Standards for Test Evaluation.” *Educational and Psychological Measurement* 6 (4): 427–38.

Guttman, Louis. 1954. “Some Necessary Conditions for Commonfactor Analysis.” *Psychometrika* 19 (2): 149–61.

Haynes, S. N., D. C. S. Richard, and E. S. Kubany. 1995. “Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods.” *Psychological Assessment* 7 (3): 238.

Henderson, N. D. 2010. “Predicting Long-Term Firefighter Performance from Cognitive and Physical Ability Measures.” *Personnel Psychology* 63 (4): 999–1039.

Hirschfeld, G., and M. T. Thielsch. 2009. “Münsteraner Fragebogen zur Evaluation von Vorlesungen (MFE-V) [Münster Questionnaire for the Evaluation of Lectures].” *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis84. https://zis.gesis.org/skala/Hirschfeld-Thielsch-M%C3%BCnsteraner-Fragebogen-zur-Evaluation-von-Vorlesungen-(MFE-V) (accessed August 9, 2020).

Jäger, R. 2004. “Konstruktion einer Ratingskala mit Smilies als Symbolische Marken [Construction of a Rating Scale with Smilies as Symbolic Marks].” *Diagnostica* 50 (1): 31–38.

Kaiser, H. F., and K. W. Dickman. 1959. “Analytic Determination for Common Factors.” *American Psychologist* 14 (7): 425–39.

Katsavouni, F., E. Bebetsos, P. Malliou, and A. Beneka. 2016. “The Relationship between Burnout, PTSD Symptoms and Injuries in Firefighters.” *Occupational Medicine* 66 (1): 32–7.

Kennedy, P. E., Y. C. Seung, J. W. Donald, and R. O. Brinkerhoff. 2014. “Training Professionals’ Usage and Understanding of Kirkpatrick’s Level 3 and Level 4 Evaluations.” *International Journal of Training and Development* 18 (1): 1–21.

Kirkpatrick, D. L. 1979. “Techniques for Evaluating Training Programs.” *Training and Development Journal* 33 (6): 78–92.

Kirkpatrick, D. L. 2007. *The Four Levels of Evaluation*. Alexandria: American Society for Training and Development.

Kirkpatrick, D. L., and J. D. Kirkpatrick. 2006. *Evaluating Training Programs. The Four Levels*. San Francisco: Berrett-Koehler.

Kleinmann, M., M. Dietrich, S. Schumacher, A. Edwin, and Fleishman. 2010. *F-JAS. Fleishman Job Analyse System für Eigenschaftsbezogene Anforderungsanalysen [F-JAS. Fleishman Job Analysis System for Trait-Related Requirement Analyses]*. Göttingen: Hogrefe.

Kline, P. 2000. *The Handbook of Psychological Testing*. London: Routledge Press.

Mayring, P. 2015. “Qualitative Content Analysis: Theoretical Background and Procedures.” In *Approaches to Qualitative Research in Mathematics Education*, edited by A., Bikner-Ahsbahs, C., Knipping and N., Presmeg, 365–80. Berlin: Springer.

Meyer, N., and J. Stiegel. 2012. “Betriebliche Weiterbildung bei den Feuerwehren. Start einer Seminarreihe zur Pädagogischen Professionalisierung [Continuing Vocational Training at the Fire Brigades. Start of a Series of Seminars on Pedagogical Professionalisation].” http://www.erwachsenenbildung.at/magazin/12-17/meb12-17.pdf (accessed July 23, 2020).

Moore-Merrell, L., A. Zhou, S. McDonald-Valentine, R. Goldstein, and C. Slocum. 2008. *Contributing Factors to Firefighter Line-of-Duty Injury in Metropolitan Fire Departments*. Washington, DC: International Association of Firefighters.

Moosbrugger, H., and K. Schermelleh-Engel. 2008. “Exploratorische (EFA) und Konfirmatorische Faktorenanalyse (CFA) [Exploratory (EFA) and Confirmatory Factor Analysis (CFA)].” In *Testtheorie und Fragebogenkonstruktion [Test Theory and Questionnaire Design*], edited by H., Moosbrugger and A., Kelava, 307–24. Heidelberg: Springer.

Nunnally, J. C. 1975. “Psychometric Theory – 25 Years Ago and Now.” *Educational Researcher* 4 (10): 7–21.

Praslova, L. 2010. “Adaptation of Kirkpatrick’s Four Level Model of Training Criteria to Assessment of Learning Outcomes and Program Evaluation in Higher Education.” *Educational Assessment, Evaluation and Accountability* 22 (3): 215–25.

Questback GmbH. 2018. “EFS Survey (Version EFS Summer 2018) [Computer Software].” http://my.unipark.com/ (accessed July 23, 2020).

Revelle, W. 2018. “Psych. Procedures for Personality and Psychological Research. (Version 1.8.10) [Computer Software].” https://CRAN.R-project.org/package=psych (accessed July 23, 2020).

Rindermann, H. 2009. *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrveranstaltungsevaluation an Hochschulen mit einem Beitrag zur Evaluation Computerbasierten Unterrichts* [*Teaching Evaluation: Introduction and Overview to Research and Practice of Course Evaluation at Universities with a Contribution to the Evaluation of Computer-Based Teaching*]. Landau: Verlag Empirische Pädagogik.

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

Rosseel, Y. 2012. “Lavaan. An R Package for Structural Equation Modeling.” *Journal of Statistical Software* 48 (2): 1–36.

Rossi, P. H., J. D. Wright, and A. B. Anderson. 2013. *Handbook of Survey Research*. New York: Academic Press.

Röseler, S., T. Hannah, M. Hagel, T. Meinald, and Thielsch. 2020. “Feedback-Instrument zur Rettungskräfte-Entwicklung – Einsatzübungen (FIRE-E) [Feedback Instrument for Rescue Forces Education – Mission Exercises (FIRE-E)].” *Zusammenstellung sozialwissenschaftlicher Items und Skalen*. https://doi.org/10.6102/zis282. https://zis.gesis.org/skala/Roeseler-Thoelking-Hagel-Thielsch-Feedback-Instrument-zur-Rettungskr%C3%A4fte-Entwicklung-Einsatz%C3%BCbung-(FIRE-E) (accessed August 9, 2020).

RStudio Team. 2016. “RStudio. Integrated Development for R. RStudio. (Version 1.1.456) [Computer Software],” http://www.rstudio.com/ (accessed July 23, 2020).

Schönbrodt, F. D., and P. Marco. 2013. “At what Sample Size do Correlations Stabilize?.” *Journal of Research in Personality* 47 (5): 609–12.

Schulte, N., and M. T. Thielsch. 2019. “Evaluation of Firefighter Leadership Trainings.” *International Journal of Emergency Services* 8 (1): 34–49.

Schweizer, K. 2011. “On the Changing Role of Cronbach’s α in the Evaluation of the Quality of a Measure.” *European Journal of Psychological Assessment* 27 (3): 143–44.

Sommer, M., and N. Ove. 2011. “Learning Amongst Norwegian Fire-Fighters.” *Journal of Workplace Learning* 23 (7): 435–55.

Staufenbiel, T. 2000. “Fragebogen zur Evaluation von Universitären Lehrveranstaltungen durch Studierende und Lehrende [Questionnaire for the Evaluation of University Courses by Students and Teachers].” *Diagnostica* 46 (4): 169–81.

Steele, S. M. 1970. “Program Evaluation – A Broader Definition.” *Journal of Extension* 8 (2): 5–17.

Steinmetz, H. 2015. *Lineare Strukturgleichungsmodelle: Eine Einführung mit R [Linear Structural Equation Models: An Introduction with R]*. München: Rainer Hampp Verlag.

Taber, N., D. Plumb, and S. Jolemore. 2008. “‘Grey’ Areas and ‘Organized Chaos’ in Emergency Response.” *Journal of Workplace Learning* 20 (4): 272–85.

Tanguma, J. 2001. “Effects of Sample Size on the Distribution of Selected Fit Indices: A Graphical Approach.” *Educational and Psychological Measurement* 61 (5): 759–76.

Tavakol, M., and R. Dennick. 2011. "Making Sense of Cronbach's Alpha." *International Journal of Medical Education* 2: 53–55.

Thielsch, M. T., J. N. Busjan, and K. Frerichs. 2018. "Feedback-Instrument zur Rettungskräfte-Entwicklung – Prüfungen (FIRE-P) [Feedback Instrument for Rescue Forces Education – Examinatons (FIRE-P)]." *Zusammenstellung sozialwissenschaftlicher Items und Skalen* https://doi.org/10.6102/zis260. https://zis.gesis.org/skala/Thielsch-Busjan-Frerichs-Feedback-Instrument-zur-Rettungskr%C3%A4fte-Entwicklung-Pr%C3%BCfungen-(FIRE-P) (accessed August 9, 2020).

Thielsch, M. T., and G. Hirschfeld. 2012. "Münsteraner Fragebogen zur Evaluation von Seminaren – Revidiert (MFE-Sr) [Münster Questionnaire for the Evaluation of Seminars – Revised]." *Zusammenstellung sozialwissenschaftlicher Items und Skalen* https://doi.org/10.6102/zis86. https://zis.gesis.org/skala/Thielsch-Busjan-Frerichs-Feedback-Instrument-zur-Rettungskr%C3%A4fte-Entwicklung-Pr%C3%BCfungen-(FIRE-P) (accessed August 9, 2020).

Thompson, B., and L. G. Daniel. 1996. "Factor Analytic Evidence for the Construct Validity of Scores: A Historical Overview and Some Guidelines." *Educational and Psychological Measurement* 56 (2): 197–208.

Velicer, W. F. 1976. "Determining the Number of Components from the Matrix of Partial Correlations." *Psychometrika* 41 (3): 321–27.

Velicer, W. F., C. A. Eaton, and L. F. Joseph. 2000. "Construct Explication Through Factor or Component Analysis: A Review and Evaluation of Alternative Procedures for Determining the Number of Factors or Components." In *Problems and Solutions in Human Assessment*, edited by R.D., Goffin and E. Helmes, 41–71. Boston: Kluwer.

Verordnung über die Ausbildung und Prüfung für die Laufbahn des Zweiten Einstiegsamtes der Laufbahngruppe 1 des Feuerwehrtechnischen Dienstes im Land Nordrhein-Westfalen (VAP1.2-Feu). 2015. Regulation on the Training and Examination for the Career of the Second Entrance Office of the Category 1 of the Fire Brigade Technical Service in the State of North Rhine-Westphalia (VAP1.2-Feu). https://recht.nrw.de/lmi/owa/br_bes_text?anw_nr=2&gld_nr=2&ugl_nr=203014&bes_id=32744&menu=1&sg=0&aufgehoben=N&keyword=VAPmd-Feu#det0 (accessed July 23, 2020).

Wickham, H. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40: 1–29.

Supplementary Material: Supplementary material to this article can be found online at https://doi.org/10.5281/zenodo.3948173.