Evaluation of Firefighter Leadership Trainings

forthcoming in the International Journal of Emergency Services
and available via https://doi.org/10.1108/IJES-03-2018-0020.

Please use the following reference for this article:

Schulte, N. & Thielsch, M. T. (2018). Evaluation of firefighter leadership trainings. *International Journal of Emergency Services*. doi: 10.1108/IJES-03-2018-0020

Evaluation of Firefighter Leadership Trainings

Abstract

Purpose

The training of highly skilled officers in rescue forces is essential for success and performance of fire brigades in their daily work. The purpose of this paper is to develop a validated instrument assessing the quality of leadership trainings in firefighter education.

Design/methodology/approach

In Study 1, relevant factors of teaching quality in this specific context are established using semi-structured interviews (n=5 trainer, n=59 trainees), and a pool of corresponding survey items is tested in a pilot sample (n=7 trainer, n=26 trainees). In Study 2 (n=263 trainees), we select best-fitting items and explore the structure of latent variables via exploratory factor analyses. Study 3 (n=45 trainer, n=380 trainees) tests this structure by means of confirmatory analyses and validates the questionnaire using scales from other evaluation instruments for higher education.

Findings

Analyses resulted in a six-dimensional questionnaire reflecting relevant training processes and outcomes. Results suggest that the newly created Feedback Instrument for Rescue forces Education (FIRE) meets all relevant psychometric quality criteria.

Originality/value

By examining critical factors of training quality, we enhance the understanding of critical processes in programs for rescue forces education. The developed questionnaire provides trainers and educational institutions with a validated tool to measure these relevant processes and the desired training outcomes. Therefore, the FIRE scales can contribute to an ongoing improvement of rescue forces trainings.

Key words: Evaluation, vocational training evaluation, firefighter, fire service, rescue forces, questionnaire

1 Introduction

Leadership positions in fire departments require a wide range of skills to handle challenging operating conditions safely. Before they take on these high responsibility jobs, prospective commanding officers are provided with special training. The quality of these trainings is not only important for future subordinates, who depend on their leaders in high risk situations, but it is also particularly important to the public. To consistently ensure that commanding officers are being offered high-quality courses, these trainings should be evaluated periodically, where *evaluation* is defined as a systematic investigation of a program's worth or merit that is derived from comprehensible, empirical qualitative and/or quantitative data (Beywl, 2003). However, to our knowledge, there is currently no such instrument for reliably and validly collecting data on trainings for firefighters. To bridge this gap, the Feedback Instrument for Rescue forces Education (FIRE; in German "Feedback-Instrument zur Rettungskräfte-Entwicklung") was developed in a series of three studies: In Study 1, we conducted interviews with trainers and participants to establish factors of excellent teaching in the context of the aforementioned trainings. Based on this, we created a set of evaluation items. In Study 2, after testing for comprehensibility, relevance, and completeness, we performed an exploratory factor analysis (EFA) to reveal the underlying structure of factors that determine the quality of firefighter leadership trainings. Accordingly, we selected items which measure these factors best. Building on these results, Study 3 was supposed to validate this factor structure by means of confirmatory factor analyses (CFA) and to assess the construct validity of the extracted scales (e.g. whether the questionnaire captured the constructs it was hypothesized to cover). Agreement of trainers and trainees was assessed with a trainer's version of the questionnaire developed for this investigation.

1.1 Structure of Trainings for Leadership Positions within the German Fire Service System

Our studies were conducted in cooperation with the Institut der Feuerwehr Nordrhein-Westfalen (State Fire Service Institute North Rhine-Westphalia, IdF NRW), which is the largest of the 16 German state-run academies for fire service forces. At the IdF NRW approximately 16,000 firefighters and members of crisis committees receive trainings for all types of leadership positions in fire service every year. Within the current set of studies, we focused on trainings for group and platoon leaders, the two most prevalent training types for incident commanders. In Germany, a group leader has command over up to eight firefighters, i.e. the entire crew of a single vehicle like a pumper (Feuerwehrdienstvorschrift (FwDV) 3 [German Fire Service regulation 3], 2008). A platoon typically consists of a command car, the engines, and a ladder truck. A platoon leader supervises up to 21 firefighters (FwDV 3, 2008).

In professional fire departments[1], firefighters are provided with training for leadership positions after working several years as crew members or directly after graduating from university. Voluntary fire departments dispatch participants based on their experience and the needs of the department. Depending on the intended position (group vs. platoon leader) and background (professional vs. volunteer fire department), the duration of the tactical trainings ranges from two to eight weeks, supplemented with a theoretical module (e.g. law) with a

[1] The German fire service system is organized on a voluntary basis with firefighters having other regular jobs and being alarmed on demand. Only larger cities or large companies employ full time firefighters.

length of up to two years. The curriculum consists of different modules on various topics, such as general leadership techniques, leadership in chemical, biological, radiological or nuclear emergencies, and legal regulations, and there is a train-the-trainers module for instructing firefighters that rank lower[2] in the organizational hierarchy (AG-BIII, 2007; AG-FIII, 2005; AG-F IV, 2007). The content and complexity of these modules vary according to the aspired position.

Over the course of the trainings, a wide variety of teaching methods are used. While some content is taught in lectures, other parts comprise map exercises, group work with presentations, and independent (home) work. The skills acquired during these theory-driven lessons are applied in several mission simulations. On special training grounds, many facilities such as apartment buildings, a hospital, and laboratories set the scene for realistic mission exercises and leadership tasks (IdF NRW, 2012; IdF NRW, 2013).

1.2 Common Concepts for Program Evaluation

To assess the quality of the applied training concepts accurately, and to identify potential for further improvements, a proper evaluation system is needed. When measuring training outcomes, four steps are commonly distinguished: reaction, learning, behavior, and results (Kirkpatrick, 1979). *Reaction* is defined as how well the participants liked the evaluated training. The favorable reaction of trainees is an important precondition for learning processes (Blanchard and Thacker 2010; Kirkpatrick, 1998), as it promotes attention and motivation which are crucial cognitive processes for effective social learning (Bandura, 1977). On the second level, *learning* describes the degree to which principles, facts and techniques are understood by the trainees (Kirkpatrick, 1979). In a third step, changes in the participants' *behavior* on the job are determined. Finally, one can aim for measuring consequences for *organizational results*, e.g., increases in service quality or a reduction of costs (Kirkpatrick, 1979). Despite intensive literature research, we were not able to find measurement instruments for any of the above described evaluation stages for trainings of firefighters. Kirkpatrick (1979) suggests evaluating a program on higher levels only after it has been shown to be successful on lower levels. We therefore aim to develop a questionnaire that covers Kirkpatrick's (1979) first two steps of evaluation (reaction and learning).

Trainees' reactions are commonly assessed by means of standardized questionnaires (Blanchard and Thacker, 2010). Two types can be distinguished here: *affective* and *utility questionnaires*. "An affective questionnaire measures general feelings about training ('I found this training enjoyable'), whereas the utility questionnaire reflects beliefs about the value of training ('This training was of practical value')" (Blanchard and Thacker, 2010, pp. 333 f.). Here, we use the latter type as it is more conducive to identify indications for change.

On level two, the use of tests with a pre-post comparison in a control group design is recommended (Kirkpatrick, 1979). Other methods to measure the (subjective) learning success are self-ratings of participants, trainers or external raters (Holling, 1999). This approach is applied here to supplement existing exams trainees have to take at the beginning, in the middle, and/or at the end of their training.

A second classification of evaluation methods is the separation between process and outcome evaluation (Blanchard and Thacker, 2010). As processes are only covered on level

[2] In Germany, regular crew members and squad leaders (a squad consists of two or three men or women) are trained on municipal or city level. Only trainings for higher positions are offered by state-run academies.

one of the Kirkpatrick model, level one is essential for identifying parts that might have gone wrong. In identifying such areas, the trainers derive benefits from the evaluation and the quality of the training programs can be improved. Outcome measures provide important information on whether the training goals are achieved or not.

Therefore, the aim of the presented studies is to develop an evaluation questionnaire which allows trainers and training facilities to assess the quality of their trainings reliably and validly and to identify potential areas of further improvements.

2 Study 1

As a first step towards the development of a questionnaire, we first had to determine the critical factors of successful teaching in firefighter trainings.

2.1 Method

We conducted semi-structured face-to-face interviews with n = 5 trainers and n = 3 trainees at IdF NRW. Additionally, we administered paper-based interviews with the same questions to n = 56 trainees. All participants were male, ranging from 23 to 56 years old (M = 34.97, SD = 7.9). They were instructed to remember prior leadership trainings and to state in detail the main characteristics of good teaching in the context of firefighter and rescue forces education. Additionally, they were asked to describe good trainers, good trainees, necessary context conditions, and potential bias variables that, in their opinion, may influence how trainees judge the quality of trainings (but do not actually influence the training quality itself; Marsh and Roche, 1997; Spiel, 2001). Participation took about 15 to 20 minutes and was voluntary, anonymous, and without any compensation.

2.2 Results and Discussion

Participants made 330 statements on characteristics of good teaching. Those were clustered in sub-categories via content analysis (Mayring, 2000) by two independent and trained observers. In the same manner, the aspects of a good trainer (209 statements), a good trainee (132 statements), good context conditions (158 statements), and potential biases (46 statements) were categorized (see Table A1 in the OSF material at osf.io/m39ug). This served as a basis for the subsequent generation of items, which were directly derived from the categorization scheme or, if applicable in the firefighting context, from other existing teaching evaluation instruments.

In a second step, we piloted the resulting list of 116 items to assess their comprehensibility and relevance (all items are available from the corresponding author upon request). Additionally, participants were asked whether the presented items covered all relevant aspects that affect the quality of leadership trainings in the context of firefighter and rescue forces education. The pilot sample consisted of n = 7 trainers (male: 6; age: 27 to 45 years M = 38.9, SD = 6.0) and n = 26 trainees (male: 25; age: 21 to 43 years M = 30.5, SD = 6.9) at IdF NRW. The items were rated as understandable, appropriate and exhaustive. Thus, it appears legitimate to conclude that from the perspective of trainers and trainees, the item list covered all important aspects of trainings for leadership positions at fire departments. This indicates high content and face validity of the constructed item set. Furthermore, items could be separated in two groups: First, 65 questions regarded general aspects of firefighter education that are applicable in nearly every possible training course (Table A2 of the OSF material at osf.io/m39ug). Second, 51 questions included very specific teaching methods used

at IdF NRW (such as different forms of mission simulations, homework or group tasks) that are not applied in every step of the leadership trainings as well as items regarding potential bias factors. In the following two studies, we will focus our analyses on the items that were generally applicable.

3 Study 2

The aim of the second study was to explore factors that are most relevant for a general training's quality and to reduce the item set. For this purpose, we employed a quantitative empirical approach using exploratory factor analysis (EFAs), an adequate method to narrow down data from a large item pool to a smaller set of underlying latent factors (see Costello and Osborne, 2005). Additionally, EFAs allow for the inclusion and exclusion of items based on their ability to capture these factors.

3.1 Method

3.1.1 Sample

A total of 263 trainees from eleven group- and platoon-leader courses completed the evaluation, resulting in a response rate of 97%. Data from 20 participants were excluded from further analyses as they either did not agree with the use of their data (n = 10), did not respond to more than 10% of the items (n = 7), or had a monotonous or unrealistic answering style (n = 3). The final sample consisted of 243 participants (96% male, which represents a typical proportion in these trainings) ranging from 21 to 55 years old (M = 31.8, SD = 6.5). The mean job experience was 13.8 (SD = 8.2) years with on average 9 (SD = 15.5) emergency incidents per month. In the sample, 14% worked at professional fire departments, 70% at volunteer departments and 16% stated to be engaged in both. Participants took part voluntarily and on an anonymous basis without any compensation.

3.1.2 Measures and Procedure

Participants filled out the paper and pencil questionnaires at the end of the trainings but before a final exam. We collected data from 44 items measuring processes and 21 items measuring outcomes of the trainings (Table A2 of the OSF material at osf.io/m39ug). The items were rated on a seven-point Likert scale (1 = *strongly disagree*, 7 = *strongly agree*). Besides these, other items were part of the questionnaire but not analyzed in this context as they measured bias variables, provided first validation estimates or asked for special teaching methods at IdF which are not part of the core evaluation questionnaire presented here.

3.1.3 Statistical Analyses

All data analyses were performed with R (R Core Team, 2016) using the packages e1071 (Meyer at al., 2015), psych (Revelle, 2016), and lavaan (Rosseel, 2012). We calculated descriptive statistics (means, standard deviations, response rate, skewness, and excess) as well as Pearson item inter-correlations for the purpose of item selection, i.e., to avoid items not capable of describing differences in course quality adequately and independently from the information generated by other items. With the selected items, we conducted exploratory factor analyses (principal component analyses). Items reflecting the behavior of trainers were not excluded for skewness, excess, means or standard deviations to make sure there were enough items for this scale in this step. To maximize the acquired information per factor

while accounting for the expectable dependencies between process and outcome factors, we conducted one EFA for process and one for outcome variables, each with varimax rotation.

3.2 Results and Discussion

3.2.1 Item Selection

Items had to be removed if they (a) were answered by less than 95% of the participants (none of the items relevant for this study met this criterion), (b) had an absolute skewness value of >2 (none of the relevant items) or an absolute excess value of >7 (none of the relevant items), (c) showed item inter-correlations >.75 (five items), or (d) had a mean ≥6 and a standard deviation of ≤1 (three items). Additionally, one item was excluded due to its limited relevance in combination with an unfavorable answer distribution. The final item set included in the exploratory factor analyses comprised 39 process and 17 outcome items.

3.2.2 Exploratory Factor Analyses for Process Items

The scree test (Cattell, 1966) suggested that two or four factors should be extracted, the parallel analysis (Horn, 1965) suggested this should be four factors. Therefore and because of content-related deliberations, four factors were extracted. Based on the loading patterns, we excluded 23 items as they showed the lowest factor loading on the respective factors. For the remaining items, we conducted discriminatory power analyses which did not lead to additional exclusions. To further reduce the number of items, we calculated reliabilities (raw alphas) if an item was dropped. In this step, two items were excluded due to dropped α value in combination with content-related reasons and one item was excluded based on its dropped α value exclusively. For detailed information on all items and the corresponding reasons for exclusion, see Table A2 in the OSF material at osf.io/m39ug; for the factor matrix, see Table A3 in the OSF material.

3.2.3 Exploratory Factor Analyses for Outcome Items

EFAs were performed the same way for outcome items as they were for process items. The scree test (Cattell, 1966) suggested one factor, the parallel analysis (Horn, 1965) two factors. We decided to extract two factors. The discriminatory power was sufficient for all tested items, and estimates for α if the item was dropped did not differentiate between the items.

Five items were excluded as they exhibited high cross-loadings on both factors. An additional four items were excluded with regard to the content (e.g. very specific focus) and the loading patterns. Item-specific reasons for exclusion are presented in Table A2, and relevant item parameters are shown in Table A4 of the OSF material at osf.io/m39ug.

3.2.2 Extracted Scales and Their Interpretation

The final questionnaire measures the four process factors *trainers' behavior*, *structure*, *overextension*, and *group* as well as the two outcomes factors *self-rated competence* and *transfer*. The first scale of the version tested here (trainers' behavior) is supposed to reflect the degree to which the trainers motivate participants, deliver their lessons concisely and give useful feedback. Moreover, it asks whether trainers are interested in trainees' learning success. The trainer scale consists of four items. Clarity of structure during the whole training is assessed with the three-item structure scale. The overextension scale (three items) covers task difficulty, speed of impartation, and the number of topics addressed. As the last process variable, the group

scale deals with the active participation and mutual social support of trainees (three items). Competence acquisition was designed to cover five different learning areas associated with action regulation in emergency situations. Finally, the transfer scale asks for self-rated ability to use the acquired knowledge appropriately on the job (three items). Items of the final questionnaire are presented in Table 1; values for reliability in terms of internal consistency in Table 2.

Table 1
Final FIRE Items

Item no.	Dimension	Item
1.	trainer	The trainers condensed difficult topics concisely.
2.	trainer	I think the trainers gave useful feedback.
3.	trainer	The trainers motivated me to participate actively in the course.
4.	trainer	I think the trainers were interested in the learning success of the participants.
5.	overextension	I was overexerted by the amount of subject matter. (reversed)
6.	overextension	The speed of impartation was too high. (reversed)
7.	overextension	The course content was too difficult to me. (reversed)
8.	structure	I think the course was well-structured.
9.	structure	I was always able to follow the structure of the course.
10.	structure	I think the course gave a good overview of the subject area.
11.	group	The other trainees participated actively.
12.	group	The participants supported each other.
13.	group	I think there was a strong cohesion within the course.
14.	competence	After this training, I can identify dangerous situations earlier.
15.	competence	After the training, it is easier to make decisions in critical situations.
16.	competence	After this training, I know my personal limitations better than before.
17.	competence	After this training, I dare to keep calm in stressful situations better.
18.	competence	The training enabled me to give more specific and clearer assignments.
19.	transfer	I feel prepared very well for my next mission as a leader.
20.	transfer	By participating in the field trainings, I gained the necessary self-assurance for leading a mission.
21.	transfer	I can use the acquired knowledge on the job.

Note. Trainer = Trainers' behavior, competence = competence acquisition. Note that high scores one overextension items indicate high levels of overextension and therefore low trainer performance. Reversal of the item scores for these items is therefore recommended. See scoring instruction in the online supplement for details.

Taken together, Study 2 resulted in a short questionnaire measuring six scales with 21 items. The scales correspond in part directly to the variables of successful teaching in firefighter trainings identified in Study 1. The scales *trainer* and *group* were identified in both investigations. We assume the factor *trainees* was merged in the group scale as both constructs have a definitional overlap. In contrast, the broad *trainer* category derived from Study 1 turned

out to be empirically divisible into general trainer behavior and the training's structure. Likewise, *teaching success* (as identified in study 1, see Table A1) can be divided into competence acquisition and transfer. For printable FIRE versions (English and German) as well as for scoring instructions, see the OSF material at osf.io/m39ug.

Table 2

Reliability coefficients and measurement model tests for all FIRE scales

	Study 2		Study 3				
Scale	Cronbach's α	ω_H	Cronbach's α	ω_H	$\Delta\chi^2$	*df*	*p*
Trainers' behavior	.83	.83	.73	.78	156.50	4	<.001
Structure	.83	.83	.80	.81	89.80	3	<.001
Overextension	.86	.87	.86	.86	22.44	3	<.001
Group	.74	.76	.79	.81	138.04	3	<.001
Competence	.85	.85	.82	.82	127.24	5	<.001
Transfer	.78	.80	.75	.76	73.83	3	<.001

Note. $N_{Study\ 1} = 243$, $N_{Study\ 2} = 382$. χ^2-difference tests compare essentially tau-equivalent with congeneric measurement models.

4 Study 3

The aim of the third study was to cross-validate the questionnaire's internal structure proposed in Study 2 by means of confirmatory factor analyses (CFA's). Additionally, we collected information on the construct validity of the instrument and investigated the agreement of trainers and trainees.

4.1 Method

4.1.1 Sample

All firefighters who received training as group or platoon leaders at IdF NRW in the first quarter of 2017 were asked to participate. The intended sample size was $N = 400$ based on the recommendation for CFAs with three indicator variables per factor and loadings of .6 (Gagne and Hancock, 2006). A power analysis for correlative validity measures showed that this would be sufficient even for small effects ($|\rho| = 0.12$, power = .8). The sampling procedures yielded a total sample size of 382 trainees (from 18 different courses). The response rate was 88%. Two participants were excluded from analyses due to missing values on all major variables of the study and another one because of straight-lining. The ages of the participants were between 20 and 55 years ($M = 33.3$, $SD = 6.9$), and 95% of them were male (a representative number in these kinds of trainings). The mean job experience of voluntary firefighters was 14.42 ($SD = 6.6$) years with on average 6.35 ($SD = 12.0$) emergency incidents per month. The mean job experience of professional firefighters was 9.35 ($SD = 7.3$) years with 24.62 ($SD = 23.9$) emergency incidents per month. In the sample, 54% worked at professional fire departments, 25% at volunteer departments and 21% stated to be engaged in

both. Additionally, we collected 45 evaluations from the trainers' perspective.[3] All participants took part voluntarily and anonymously without any compensation.

4.1.2 Measures

In addition to the FIRE items (German version; see Table A5 in the OSF material at osf.io/m39ug), scales from other well-established evaluation instruments for college lectures were used for the validation of individual FIRE scales. Items measuring the overall satisfaction served as a criterion (for details, see OSF material at osf.io/m39ug). Bias variables formed a third group of measures. Together with these measurements, data for the construction and validation of additional, more specific scales not pertinent to the present paper were collected. Please refer to the corresponding author for a detailed list of items used here. Unless specified differently, participants indicated their agreement with the statements on a seven-point scale (from 1 = *strongly disagree* to 7 = *strongly agree*) with an *unanswerable* option.

4.1.2.1 Scales Corresponding to Individual FIRE Scales

We used scales from other validated evaluation questionnaires as convergent criteria for specific FIRE scales, primarily from two German evaluation instruments for higher education (HILVE, Rindermann, 2001 and TRIL, Gläßer et al., 2002). None of the established evaluation scales cover the group of trainees with an according scale. Therefore, we asked about group-related behaviors with three self-developed items (e.g. "On how many evenings of the course did you spend at least two hours sitting together?"). For a detailed description, including reliabilities and sample items of each scale employed here, we refer the reader to the supplemental material.

4.1.2.2 Bias Variables

We used the following single item measures for bias variables: "I felt very well prepared for this training." (preparation prior to the training), "The amount of time I spent with the training was appropriate for me." (time expenditure), "The group size was adequate." (group size), "I was able to fully concentrate on the training." (concentration), "I feel very well prepared for the exam." (preparedness for exams), "I am proud to be trained for this kind of leadership position (group leader/platoon leader)." (proud of the participation in a leadership training). Mood was measured with a five-point equidistant smiley scale (Jäger, 2004).

4.1.2.3 Valuation by the Trainers

The trainers' perspective was captured by a specially developed version of the FIRE questionnaire. Out of 21 items, 15 were adapted slightly (e.g., "After this training, the participants can identify dangerous situations earlier." instead of "After this training, I can identify dangerous situations earlier.").

[3] Trainers were between 28 and 67 years old (M = 39, SD = 7.58). As it is representative for trainers at IdF, 91% were male. Their mean job experience at voluntary fire departments was M = 17.09 years (SD = 9.4) and at professional fire departments M = 8.87 years (SD = 7.44). Trainers reported to have M = 2.54 emergency incidents per month (SD = 2.2) at voluntary fire departments and M = 18.34 (SD = 27.21; Mdn = 5.5) incidents at professional fire departments.

4.1.3 Procedure

All participants completed the evaluation questionnaire at the end of the training but before the exam, if there was one. The questionnaires were handed out by the trainers and were returned to them in a sealed envelope. Participants responded voluntarily and anonymously. Each trainer was requested to complete the trainer version of the questionnaire. Trainers were not assured of anonymity, as the study design requested that their questionnaires matched with the questionnaires of participants who took their course.

4.1.4 Statistical Analysis

The paper-pencil questionnaire was built and scanned with EvaSys (version 7.0). All data analyses were performed with R (R Core Team, 2016; version 3.3.2) using the packages psych (Revelle, 2016; 1.6.12), multilevel (Bliese, 2016; version 2.6), lavaan (Rosseel, 2012; version 0.5-22), semPlot (Epskamp, 2014; version 1.0.1), semTools (semTools Contributors, 2016; version 0.4-14), and metafor (Viechtbauer, 2010; version 1.9-9). A first model was fitted for process (trainers' behavior, structure, overextension, and group) and a second one for outcome scales (competence and transfer). For the associations of FIRE subscales and validity criterions, correlation coefficients were calculated. To determine the agreement of trainers and trainees, a meta analytic approach was applied. Based on means of trainers and trainees per course, we calculated standardized mean differences (Hedges' *g*) with pooled standard deviations for each FIRE scale. This procedure also considered the varying sample variance, which was caused by the course-specific sample size in both the trainee and the trainer group. Since the agreement depends highly on the trainer's ability to judge his or her work correctly, random effects models with restricted maximum likelihood estimation were used.

4.2 Results and Discussion

4.2.1 Descriptive Statistics and Reliability

Means, standard deviations, intra class correlations and correlations for all measures used in the current study are presented in Table A7 of the OSF material at osf.io/m39ug. Table A8 (OSF material) reports overall evaluation results from both, study 2 and 3. The reliability estimates for all FIRE scales are reported in Table 2. A very common measure of reliability is Cronbach's α. Yet, for each scale, ω_H (McDonald, 1999) is the more appropriate reliability measure, as congeneric measurement models fit better than essentially tau-equivalent models (for χ^2-difference tests results see the last three columns of Table 2). Applying the reliability standards for the assessment of learning success and program evaluation (Evers, 2001), the reliability of all FIRE scales can be judged as sufficient (trainers' behavior and transfer) or good (structure, overextension, group, and competence scale). Results based on the data of Study 3 confirm the estimates based on Study 2. Only the reliability of trainers' behavior was considerably lower in Study 3 but still acceptable. Compared with other German teaching evaluation instruments, FIRE scales reach equal (e.g., compared with HILVE teaching competence scale, TRIL structure scale) or considerably higher levels of internal consistency (e.g. compared with HILVE overextension).

4.2.2 Factorial Structure

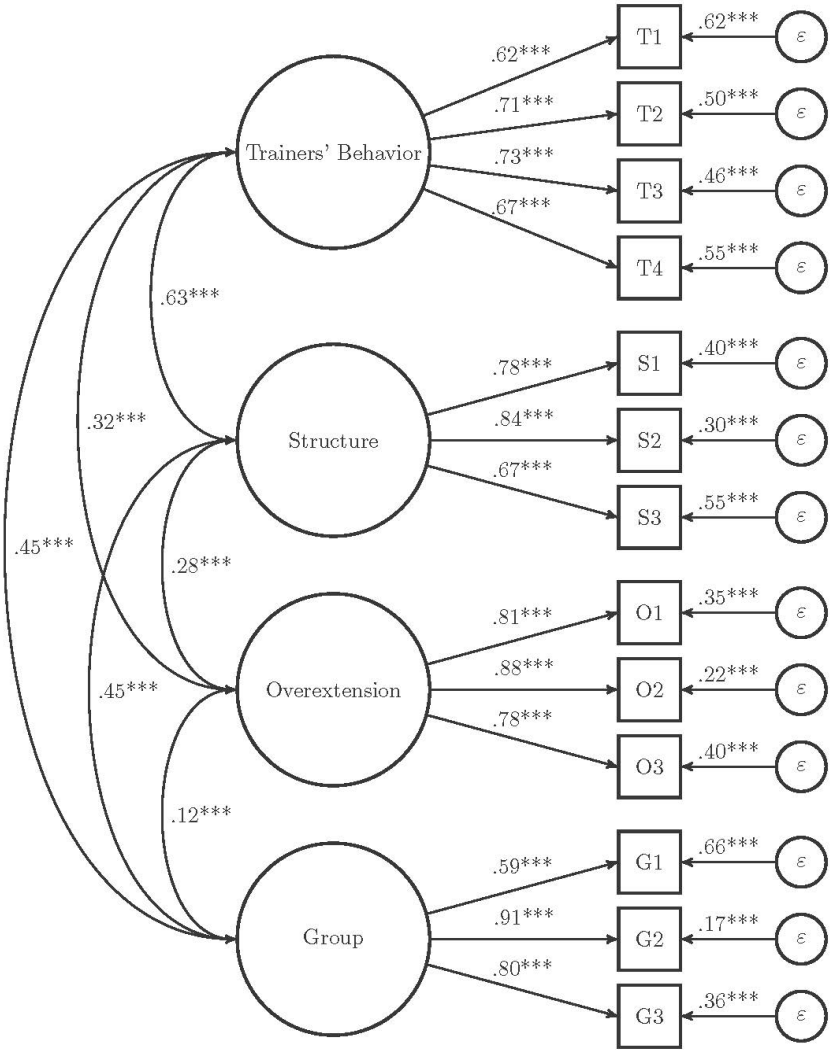

Figure 1. Results of confirmatory factor analysis for process scales, standardized coefficients. * $p < .05$. **. $p < .01$. *** $p < .001$.

To replicate the factor structures found in Study 2, confirmatory factor analyses (CFAs) were conducted testing two models. The first model consisted of the four process factors *trainers' behavior*, *structure*, *overextension* and *group*. According to the criteria laid out by Schermelleh-Engel et al. (2003), the model displayed an acceptable fit (CFI = .96, RMSEA = .06 [.05, .08], SRMR = .06. Only the TLI value of .94 did not meet the regular fit criteria. The χ^2-test was significant ($\chi^2(59) = 146.37$, $p < .001$) which is typical for large sample sizes. However, relative to the degrees of freedom, the χ^2-value was acceptable ($\chi^2/df = 2.48$). Figure 1 shows all path coefficients for this model. Overall, results support the specified model.

The second model supposed that outcome measures build two factors (Competence and Transfer). This model yielded a good fit based on the SRMR of .04 and an acceptable fit based on a CFI of .95. Two fit indices did not reach an acceptable level (RMSEA = .09 [.07, .11], TLI = .93). The χ^2-test was significant ($\chi^2(19) = 74.80$, $p < .001$) and the χ^2/df ratio was

not acceptable (χ^2/df = 3.94). Modification indices for this initial model indicated that a correlation between the first and the second item of the Competence factor should be added to the model. Item one of this scale asks about trainees' abilities to identify critical situations earlier, and item two asks about trainees' abilities to make decisions in such situations. We consider this overlapping content as a reasonable theoretical explanation for the fact that both items covary beyond the degree that is explained by the extracted factor. Thus, we allowed the two items to covay in a new model. The fit indices of this new model are acceptable (TLI = .96, RMSEA = .07 [.045, .09] to good (CFI = .97; SRMR = .03). The χ^2 test is still significant ($\chi^2(18) = 48.386$) but the χ^2/df ratio is now acceptable ($\chi^2/df = 2.69$). Fit indices and a χ^2 difference test ($\chi^2(1) = 26.41$, $p < .001$) suggest the recent model, presented in Figure 2. Results provide support for a two-factor structure.

Thus, as the EFA of Study 2 suggested, the FIRE questionnaire measures four distinct process variables (trainers' behavior, structure, overextension, and group) as well as two different outcome variables (self-rated competence and transfer). The multidimensional structure of the questionnaire demonstrates that prospective firefighter leaders differentiate among several components of effective teaching. This also implicates the absence of large halo effects, i.e. responses are not simply a generalization from some subjective feelings, external influence or an idiosyncratic response mode influencing responses to all items (Marsh, 1987).

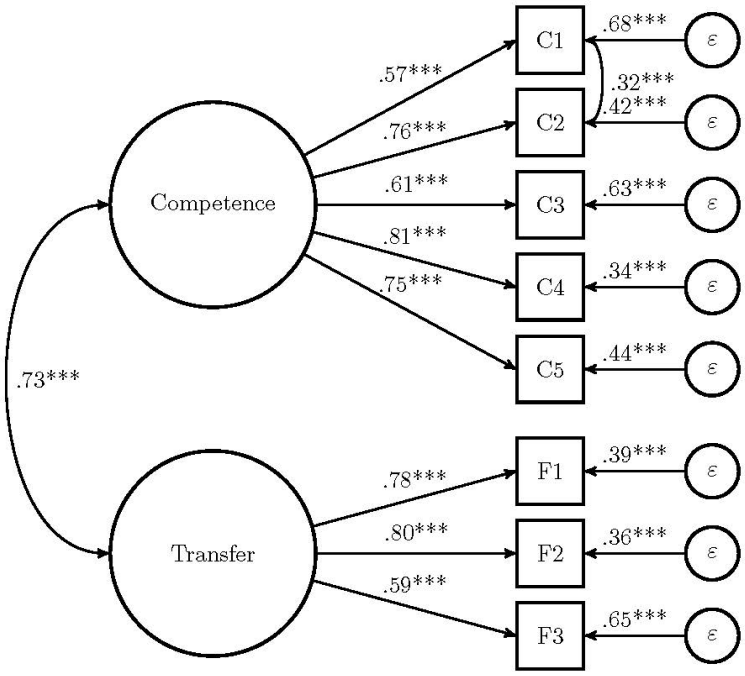

Figure 2. Results of confirmatory factor analysis for outcome scales, standardized coefficients. * $p < .05$. **. $p < .01$. *** $p < .001$.

4.2.3 Construct Related Measures

4.2.3.1 Bias variables

We tested for the influence of six bias variables which are presented as measures 17 to 22 in Table A7 in the OSF material at osf.io/m39ug. Based on Cohen's (1992) classification,

the effect sizes can be judged as small to medium. Only time expenditure showed a large association with the overextension scale ($r = .43$, $p < .001$), what might be seen as evidence for the validity of this scale because it can be assumed that participants who perceive the training as challenging need to spend more time with the course and the exam preparation than those who find it straightforward. Furthermore, results clearly contradict criticism on the meaningfulness of training evaluation results by denotations such as *happiness sheets* (Hölbling, 2007; also reported in Kirkpatrick, 1998). Even when neglecting the reciprocal association of mood and the assessment of training quality (i.e., mood might not only influence quality ratings but might also be a result of training quality), mood can only explain a small percentage of variance in the evaluation results.

4.2.3.2 Association With Corresponding Scales of Existing Evaluation Instruments

Results for scales of other established teaching evaluation instruments are presented as variables 7 to 10, 12, and 13 in Table A7 of the OSF material at osf.io/m39ug. All FIRE scales show large associations with their corresponding scales from other evaluation instruments (trainers' behavior and teaching competence (HILVE) $r = .67$, $p < .001$ and $r = .59$, $p < .001$ for teachers' support (HILVE); $r = .66$, $p < .001$ for the structure scales of TRIL and FIRE; $r = .66$, $p < .001$ for overextension scales of HILVE and FIRE; $r = .50$, $p < .001$ between the competence scale (FIRE) and quantitative learning as well as $r = .55$, $p < .001$ for competence and qualitative learning (both HILVE II). Significant correlations of FIRE scales with validation scales measuring one of the other facets were also observed, but with consistently smaller effect sizes. The self-constructed group validation measure and the FIRE group scale were moderately associated with each other ($r = .36$, $p < .001$). Note that the validation measure's low internal consistency of $\alpha = .57$ restricts the obtainable empirical correlation. Taken together, results show consistent positive associations for all FIRE scales with corresponding scales of other validated evaluation tools.

4.2.3.3 Criterion Validity

All FIRE scales show medium to large associations with trainees' overall satisfaction (r ranging from .31 to .56, $p < .001$) and the school grade for the entire course (r ranging from .21 to .43, $p < .001$); see Table A7 in the OSF material at osf.io/m39ugfor detailed results. Thus, results support the valid prediction of both criteria investigated here.

4.2.3.4 Agreement Between Trainers and Trainees

To estimate the agreement between trainers and trainees, we applied random effects meta-analytic techniques with courses on study level. Effect sizes are Hedges' *g*s, which indicate the differences between the mean rating of all trainers and the mean rating of all participants from a specific course divided by the pooled standard deviation of both groups. For a detailed report on the results of the meta-analyses see Table A9 in the supplemental material. Negative *g*s are obtained if participant ratings are more positive than trainers' judgements. Results indicate that trainers do not judge themselves ($\hat{\theta} = -.14$, 95%-CI [-.57, .29]), the structure ($\hat{\theta} = -.46$, 95%-CI [-.93, .01]) or the group of participants ($\hat{\theta} = -.62$, 95%-CI [-1.26 .02]) consistently more positive or negative than trainees do. This does not necessarily imply that in each course trainers and trainees absolutely agree, but only that our results offer no evidence for general severity or leniency bias on the side of the trainers or trainees relative to each other. It should be stated, however, that power for these tests was low

and missing evidence for differences should not be interpreted as a proof of agreement on these scales. On the other scales, differences are medium to large or large (Cohen, 1992). Trainees perceive the training as considerably less overexerting ($\hat{\theta}$ = -.91, 95%-CI [-1.25, -.58]) and rated their own competence acquisition ($\hat{\theta}$ = -.63, 95%-CI [-1.04, -.22]) and transfer ($\hat{\theta}$ = -.66, 95%-CI [-1.11, -.22]) more positively. Trainers showed a severity tendency compared to the trainees. This tendency seems to be less pronounced on scales which can be easily influenced by themselves (trainer's behavior, structure). On scales that are stronger affected by participants' characteristics (overextension, group, competence, transfer), trainers showed a more pronounced severity tendency (or trainees a stronger leniency tendency respectively). Overall, the agreement is sufficiently high to assume a (at least to a certain degree) shared understanding of the content covered by the FIRE scales, but it is too low to rely on trainers' self-appraisals exclusively.

5 General Discussion

Given the enormous importance of leadership skills for officer-level firefighters, they must be trained to a very high quality. The present paper addresses the lack for an evaluation instrument for such trainings, resulting in a short and powerful tool for quality management in firefighter education. Altogether, evidence for several kinds of validity was collected. The exhaustive item construction based on qualitative analyses, especially results of the pilot described in Study 1, offered support for face and content validity: Participants explicitly stated the items covered all relevant aspects affecting the quality of trainings, and items were rated as appropriate and exhaustive. Subsequent item exclusions were carried out in close contact with subject-matter experts at IdF, where they affirmed face and content validity of the condensed items. Building on these results, the third study confirmed the proposed factor structure found in Study 2 using a completely different sample of trainees. Despite the moderate influence of some bias variables, Study 3 provides strong evidence for construct validity. Criterion validity was demonstrated for the prediction of participants' satisfaction and their overall course appraisal. Altogether, the FIRE questionnaire is highly capable of assessing rescue forces trainees' evaluations during leadership trainings. Covering the first two levels of Kirkpatricks (1979) evaluation framework, namely reaction and learning, trainers and responsible executives in firefighter education can get an impression about the success of their work using a short, time-efficient measure. Further, trainers can determine which areas to change, specifically with regard to the didactics used and the required level of exertion as well as by reviewing results on group behavior, perceived learning success, and transfer of knowledge.

The newly developed evaluation instrument offers several practical benefits. First, we provide an option to measure the quality of firefighter trainings economically, meeting all central psychometric standards. Implementing the instrument within a course does not require further skills and only adds 10 minutes. Evaluation should take place directly after the training and under controlled conditions like in a seminar room. Besides that, scoring procedures are very straightforward. Each item value (7 for the most and 1 for the least preferable answer option) is assigned to the corresponding scale. Then, averages are calculated on participant level (scale values) and these, in turn, are aggregated on course level. Applying regular survey software, data can be easily collected via mobile devices and analyzed electronically, leading to further time efficiency.

Even though the instrument was developed in cooperation with a firefighter academy, we made sure that items do not contain fire service-specific content but cover relevant aspects of rescue forces trainings in general. We assume that underlying principles of good teaching are comparable between fire service, ambulance service and other rescue forces. Beyond that, the items might be applicable for leadership trainings in other high-reliability contexts such as police agencies or the military. However, the scales have been validated only in the firefighter context so far. Thus, the transfer into other areas of application requires further validation studies. Results suggest that it would be worth the effort: The diverging judgements of trainers and trainees on several FIRE scales in Study 3 underline that evaluations by the trainees add information which cannot be obtained by simply asking trainers.

5.1 Lessons Learned

On a theoretical level, the present studies add to the understanding of good teaching in firefighter education and its evaluation. By confirming the proposed factorial structure, we demonstrated that the multidimensional nature of student evaluations, as has been pointed out for college lectures (Ghedin and Aquario, 2008; Marsh, 1987), also holds true in vocational training contexts in rescue services. Our factors correspond very well to established dimensions of evaluation instruments for college lectures: The Students' Evaluation of Educational Quality questionnaire (SEEQ; Marsh, 1984) – one of the most widely used and empirically tested tools – describes nine dimensions of which seven are also covered by FIRE scales. Namely, *Learning*, *Group Interaction*, and *Workload/Difficulty* are completely analogous. Additionally, *Enthusiasm* and *Individual Rapport* form the Trainers' Behavior scale, *Organization* and *Breadth of Coverage* together form the FIRE Structure scale. Only the dimensions *Examination* and *Assignments* are not covered by FIRE scales, as they are not considered core aspects of vocational trainings for rescue forces.

Further implications for the understanding of latent factors in the quality of vocational trainings can be derived from the intercorrelations of FIRE scales. A high intercorrelation and similar association patterns with other variables indicate a close relationship between the FIRE factors of trainers' behavior and structure. Both factors are tightly linked to trainers' actions. In contrast, overextension and group also depend on participant characteristics. These patterns are in line with the multifactorial model of course quality (Rindermann and Schofield, 2001), which classifies structure and several behavioral aspects of teaching activities as a combined *trainer* factor. High intercorrelations of acquired competence and transfer can be interpreted in two different ways: One possible explanation is that if a trainee learns something during a training at IdF, the likelihood of him/her being able to transfer the acquired competence into on the job behavior is high. This interpretation is supported by the fact that trainings have a clear vocational focus, and much time is spent with mission exercises. Alternatively, one may assume that trainees have limited abilities to predict transfer of acquired knowledge into their on-the-job behavior, and therefore trainees derive the appraisal of transfer from a single higher-level mental concept of perceived training success. However, contradicting this interpretation, confirmatory factor analysis clearly confirmed two latent factors, which explain response behavior. Additionally, trainees for leadership positions already have considerable professional experience as firefighters and are therefore experts in judging transferability of course contents. Thus, the chosen approach of basing the FIRE scales on a qualitative exploration of good teaching in the target area combined with a

literature search, an explorative, and a confirmatory study has led to a sound evaluation instrument.

5.2 Limitations and Future Research

The FIRE scales were developed in close dialogue with the largest state-run academy for fire service forces in Germany; tested samples did not differ in any known way from the target population of trainees at German firefighter academies. Yet, empirical evidence for the validity of our scales has been collected in fire service context exclusively. Consequently, to use the scales in other contexts, there should at least be another expert rating on face and content validity for the planned context. Generally, experiences should be collected with the administration and interpretation of FIRE scales in other institutions and by evaluation coordinators not part of the development team in order to test its robustness in different contexts.

Users should also be aware of the fact that the questionnaire was developed for trainings that prepare for leadership positions in which the trainee will supervise up to 21 firefighters (one platoon). These positions require limited strategic competencies regarding the coordination of tactical units as a German fire service platoon consists of two teams and one squad. Therefore, FIRE scales presented here could be well applied in trainings for ranks that are associated with comparable tactical and strategical duties. Other ranks may require training aspects not covered by these scales. Thus, in currently ongoing studies, we test adaptions of the FIRE scales in trainings at a basic level as well as for higher, more strategic positions and crisis management groups.

Due to the low percentage of female respondents in our sample, we cannot draw any conclusions about potential gender effects in the response process. We were neither able to investigate potential effects of trainees' gender on their judgments nor whether male and female trainers are evaluated equally. However, our sample is representative for the population of fire service forces and trainers in this field. More woman working in the respective positions are a requirement for reliable investigations of these questions. Besides, investigations on other potentially relevant facets of response behavior in teaching evaluations in this context, such as survey mode or aspects of social exchange (see Thielsch, Forthmann & Brinkmöller, 2018), might be worth investigating.

Furthermore, having a tool for the first levels of the Kirkpatrick model at hand, future research should address levels three and four (behavior and results). While the instrument introduced here relies on self-ratings of learning success, further studies should investigate their relationship with other data sources like exam results and performance in the field. In this context, larger data sets are required to investigate co-variation of both kinds of measures on the course level, as being able to explain between-course variance is of central importance in this context.

Finally, we need to point out that all of our studies used the German FIRE version. Future research should check the English version for appropriateness of the content in other fire service systems and test the English translation with the same procedures as described in Study 3 before use. Other language versions of the FIRE measures are highly welcome as well.

5.3 Conclusion

This paper provides the first validated evaluation questionnaire for trainings of firefighters aspiring to leadership positions. The FIRE scales are ready for use and represent an economic, reliable and valid way to measure the quality of trainings for rescue forces. Providing trainers and executives in firefighter education and beyond with such a tool, we hope to contribute to the ongoing improvement of public emergency infrastructure.

6 Acknowledgements

The authors declare no conflict of interest.

The authors thank the IdF NRW, in particular Berthold Penkert, Thomas Löchteken, Yannick Ngatchou, Stephanie Vöge and Matthias Wegener for their support in conducting this research. Special thanks to Stephanie Babiel, who put a lot of effort in conducting Study 1 and 2, and Mona Beverborg for her contributions to Study 3. Also, the authors thank Celeste Brennecka, Guido Hertel, and Heinz Holling for their valuable suggestions.

7 References

AG-B III – LFV NRW, AGBF NRW, WFV NRW, & IdF NRW (2007). Lernziele für die Ausbildung zum Gruppenführer in der Berufsfeuerwehr [Learning Objectives for the Training of Group Leaders in Professional Fire Departments]. Retrieved from http://www.idf.nrw.de/service/downloads/pdf/biii_lzk_2008.pdf.

AG-F III – LFV NRW, AGBF NRW, WFV NRW, & IdF NRW (2005). Gruppenführer-Ausbildung und Truppmann- / Truppführer- Aus- und Fortbildung der Freiwilligen Feuerwehren in Nordrhein-Westfalen [Group Leader Training and Troop Member/ Troop Leader Training for Voluntary Fire Departments in North Rhine-Westphalia]. Retrieved from http://www.idf.nrw.de/service/downloads/pdf/ rderlass_im_20051221_lernziele.pdf.

AG-F IV – LFV NRW, AGBF NRW, WFV NRW, & IdF NRW (2007). Lernziele für die Ausbildung zum Zugführer (Freiwillige Feuerwehr) [Learning Objectives for Training of Platoon Leaders (Voluntary Fire Department)]. Retrieved from http://www.idf.nrw.de/ service/downloads/pdf/lernziele_zugfuehrer_20070925.pdf.

Bandura, A. (1977), *Social learning theory*, Prentice-Hall, Oxford, England.

Beywl, W. (Ed.). (2003). "*Selected comments to the standards for evaluation of the German evaluation society - English edition –"*, available at http://www.degeval.de/images/ stories/Publikationen/Selected_Comments_German_Evaluation_Standards_030409.pdf

Blanchard, P. N., and Thacker, J. W. (2010), *Effective training: Systems, strategies, and practices* (4th ed.), Pearson Education, Upper Saddle River, N.J..

Bliese, P. (2016), multilevel: Multilevel Functions. (Version 2.6) [Computer software], available at https://CRAN.R-project.org/package=multilevel

Cattell, R. B. (1966). The scree test for the number of factors. Multivariate behavioral research, 1(2), 245-276. DOI: 10.1207/s15327906mbr0102_10

Cohen, J. (1992), "A power primer", *Psychological Bulletin*, Vol. 112, No. 1, pp. 155-159.

Costello, A. B., and Osborne, J. W. (2005), "Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis", *Practical Assessment, Research & Education,* Vol. 10, No. 7, pp. 1–9.

Epskamp, S. (2014), "semPlot: Unified visualizations of structural equation models", *Structural Equation Modeling,* Volume 22, No. 3, pp. 474–483.

Evers, A. (2001), "The revised Dutch rating system for test quality", *International Journal of Testing,* Vol. 1, No, 2, pp. 155-182.

Feuerwehr-Dienstvorschrift (FwDV) 3: Einheiten im Lösch- und Hilfeleistungseinsatz [German fire service regulation 3: units in extinguishing and assistance incidents] (2008).

Gagne, P., and Hancock, G. R. (2006), "Measurement model quality, sample size, and solution propriety in confirmatory factor models", *Multivariate Behavioral Research,* Vol. 41, No. 1, pp. 65-83.

Ghedin, E., and Aquario, D. (2008), "Moving towards multidimensional evaluation of teaching in higher education: A study across four faculties", *Higher Education,* Vol. 56, No. 5, pp. 583-597.

Gläßer, E., Gollwitzer, M., Kranz, D., Meiniger, C., Schlotz, W., Schnell, T. and Voß, A. (2002), „Trierer Inventar zur Lehrevaluation [Trier inventory of teaching evaluation; Measurement instrument]", available at https://www.zpid.de/pub/tests/ PT_9004523_TRIL_weibl_Doz_Fragebogen.pdf

Hölbling, G. (2007). *Handlungshilfen für Bildungsberater: Qualitätssicherung betrieblicher Weiterbildung* [Guidance for eduactional consultants: quality management in vocational training]. Bielefeld, Germany: Bertelsmann.

Holling, H. (1999), "Evaluation eines komplexen Fortbildungsprogramms zur Steigerung der beruflichen Kompetenz [Evaluation of a complex profressional training for the enhancement of job skills]", in H. Holling and G. Gediga (Eds.), *Evaluationsforschung* [*Evaluation research*], Hogrefe, Göttingen, pp. 1-33.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30(2), 179-185. DOI: 10.1007/BF02289447

IdF NRW [State Fire Service Institute North Rhine-Westphalia] (2012), *IdF NRW Ausbildungsplan - Musterstundenplan B III 2012 [IdF NRW exemplary syllabus for group leaders in professional fire departments 2012]*, available at http://www.idf.nrw.de/ausbildung/katalog/dokumente/musterausbildungsplan_biii.pdf

IdF NRW [State Fire Service Institute North Rhine-Westphalia] (2013). IdF NRW Ausbildungsplan –Musterplan F III [IdF NRW exemplary syllabus for group leaders in voluntary fire departments], available at http://www.idf.nrw.de/ ausbildung/katalog/dokumente/musterausbildungsplan_fiii_26.pdf

Jäger, R. (2004), „Konstruktion einer Ratingskala mit Smilies als symbolische marken [Construction of a rating scale with smilies as symbolic labels]", *Diagnostica,* Vol. 50, No. 1, pp. 31-38.

Kirkpatrick, D. L. (1979), "Techniques for evaluating training programs", *Training and Development Journal,* Vol. 14, No. 6, pp. 78-92. Available at:

http://iptde.boisestate.edu/
FileDepository.nsf/bf25ab0f47ba5dd785256499006b15a4/693b43c6386707fc872578150059c1f3/$FILE/Kirkpatrick_79.pdf

Kirkpatrick, D. L. (1998), *Evaluating training programs - the four levels* (2nd ed.). Berrett-Koehler Publishers, San Francisco, CA.

Marsh, H. W. (1984), "Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility", *Journal of Educational Psychology*, Vol. 76, No. 5, pp. 707-754.

Marsh, H. W. (1987), "Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research", *International Journal of Educational Research,* Vol. 11, No. 3, pp. 253-388.

Marsh, H. W., and Roche, L. A. (1997), "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility", *American Psychologist*, Vol. 52, No. 11, pp. 1187-1197.

Mayring, P. (2000), "Qualitative content analysis", *Forum: Qualitative social research 1*, Art. 20.

McDonald, R. P. (1999), *Test theory: A unified treatment*, Lawrence Erlbaum Associates, Mahwah, NJ.

Meyer, D., Dimitriadou, E., Kornik, K., Weingessel, A., and Leisch, F. (2015), „*E1071: misc functions of the department of statistics, probability theory group (formerly: e1071)", TU Wien* (Version 1.6-7) [Computer software], TU Wien, Vienna, Austria.

R Core Team. (2016), R: A language and environment for statistical computing (Version 3.3.2) [Computer software], R Foundation for Statistical Computing, Vienna, Austria.

Revelle, W. (2016), psych: Procedures for Personality and Psychological Research (Version 1.6.12) [Computer software], Northwestern University, Evanston, Il.

Rindermann, H. (2001). *Lehrevaluation: Einführung und Überblick zu Forschung und Praxis der Lehrverantstaltungsevaluation an Hochschulen. mit einem Beitrag zur Evaluation computerbasierten Unterrichts* (1st ed.), Empirische Pädagogik e.V, Landau.

Rindermann, H. and Schofield, N. (2001), „Generalizability of Multidimensional Student Ratings of University Instruction Across Courses and Teachers", *Research in Higher Education*, Vol. 42, No. 4, pp. 377-399.

Rosseel, Y. (2012), "Lavaan: An R package for structural equation modeling", *Journal of Statistical Software,* Vol. 48, No. 2.

Schermelleh-Engel, K., Moosbrugger, H., and Müller, H. (2003), „Evaluating the fit of structural equation models: Tests of significance and descriptive goodness-of-fit measures", *Methods of Psychological Research,* Vol. 8, No. 2, pp. 23-74. Available at: https://www.dgps.de/ fachgruppen/methoden/mpr-online/issue20/art2/mpr130_13.pdf

semTools Contributors (2016), semTools: Useful tools for structural equation modeling (Version 0.4-14) [Computer software], available at: https://CRAN.R-project.org/package=semTools

Spiel, C. (2001), "Der differentielle Einfluß von Biasvariablen auf studentische Lehrveranstaltungsbewertungen [The differential influence of bias variables on

students‘ course evaluations]”, in U. Engel (Eds.), *Hochschul-Ranking. Zur Qualitätsbewertung von Studium und Lehre [College rankings: on the quality appraisal of studying and teaching],* Campus, Frankfurt am Main/New York, NY, pp. 61 –82.

Thielsch, M. T., Brinkmöller, B. & Forthmann, B. (2018). Reasons for responding in student evaluation of teaching. *Studies in Educational Evaluation, 56,* 189-196. https://doi.org/10.1016/j.stueduc.2017.11.008

Viechtbauer, W. (2010), “Conducting meta-analyses in R with the metafor package”, *Journal of Statistical Software,* Vol. 36, No. 3, pp. 1-48.