Vocations and Learning (2014) 7:41–73
DOI 10.1007/s12186-013-9106-4

ORIGINAL PAPER

The Training Evaluation Inventory (TEI) - Evaluation of Training Design and Measurement of Training Outcomes for Predicting Training Success

Sandrina Ritzmann · Vera Hagemann · Annette Kluge

Received: 8 November 2012 / Accepted: 22 August 2013 / Published online: 3 September 2013

Abstract Training evaluation in research and organisational contexts is vital to ensure informed decisions regarding the value of training. The present study describes the development of a valid and reliable training evaluation inventory (TEI), as it does not exist so far. The objectives were a) to construct an instrument that is theoretically and empirically founded, but at the same time applicable within typical organisational constraints, and b) to include the assessment and perception of training design as a formative evaluation aspect. Based on previous research, ten scales were constructed, covering the training outcome dimensions subjective enjoyment, perceived usefulness, perceived difficulty, subjective knowledge gain, and attitude towards training, as well as the training design dimensions problem-based learning, activation, demonstration, application, and integration. Reliabilities of the scales were satisfactory. Data from two training studies show that the training outcome dimensions were related to external training outcome measures, underlining the validity of the TEI. Two survey samples were used to predict training outcomes based on training design. Demonstration, application, and integration emerged as the most important design dimensions. The TEI is applicable in both training research projects and in organisational contexts. It can be used for formative and summative training evaluation purposes.

Keywords Organisational training · Training evaluation · Training outcomes · Training design · Questionnaire development · Evaluation instrument

S. Ritzmann (✉)
Institute Humans in Complex Systems, University of Applied Sciences and Arts Northwestern Switzerland, Riggenbachstrasse 16, CH-4600 Olten, Switzerland
e-mail: sandrina.ritzmann@fhnw.ch

V. Hagemann · A. Kluge
Business and Organizational Psychology, University of Duisburg-Essen, Lotharstr. 65, D-47057 Duisburg, Germany

Training in Organisations

In the last few decades, training in organisations has been an important topic for scholars and practitioners alike. Technological progress, structural labour-market changes and an ageing workforce necessitate continuous professional education to warrant innovation and productivity (Billett 2008; European Centre for the Development of Vocational Training 2010).

Evaluation of Training and Levels of Training Outcomes

Given the importance of vocational education and professional training, there is an ongoing need to evaluate training to ensure that investments have the highest possible degree of efficiency. Evaluation, as the systematic collection of descriptive and judgmental information on training, is necessary for making informed decisions regarding the implementation, modification or value of organisational training (Goldstein and Ford 2002, p. 138). Important decisions that need to be made concern for example the choice of an external training provider or the course the revisions of an existing training programme should take. The starting point for most evaluation efforts is Kirkpatrick's hierarchical model of training outcomes (Kirkpatrick 1998), which provides a rough taxonomy for training criteria (Alliger and Janak 1989; Shelton and Alliger 1993). According to Kirkpatrick (1998), training can be evaluated at four outcome levels: 1) reactions, 2) learning and attitudes, 3) behaviour, and 4) organisational results. Ideally, comprehensive evaluation considers data on multiple levels (Tannenbaum and Woods 1992), but to save time and costs, organisations often restrict evaluation to the distribution of reactionnaires, primarily measuring whether or not participants enjoyed a particular programme (Blanchard et al. 2002; Twitchell et al. 2000). This fact has been widely criticised in the training literature (Blanchard et al. 2002; Shelton and Alliger 1993; Tannenbaum and Woods 1992), based on results showing that enjoyment of a training course does not necessarily lead to learning or transfer of behaviour (Alliger and Janak 1989; Alliger et al. 1997). However, it has also been emphasised that evaluation efforts should not be guided by an obligation to cover all evaluation levels, but rather by predefined objectives in the respective organisational context (Alliger and Janak 1989; Kraiger 2002), and that researchers should offer methods that are "practical, systematic, and perceived as feasible by trainers" (Twitchell et al. 2000, p. 104). We can thus identify a gap between cost-effective organisational evaluation practices and postulations of training scholars regarding the measurement of training effectiveness on multiple levels and related to training objectives. Against this background, *the first objective of the present article is to introduce an approach to training evaluation that is theoretically and empirically founded on the one hand, but meaningful to training decision makers and applicable within typical organisational constraints on the other hand.*

Impact of Training Design on Training Outcomes

The function of training is to facilitate learning through the design of adequate instructional events (Gagné et al. 2005). Correspondingly, a number of studies have investigated the impact of training design on training outcomes and have shown that the choice of training design has a relevant influence on training effectiveness (Blume

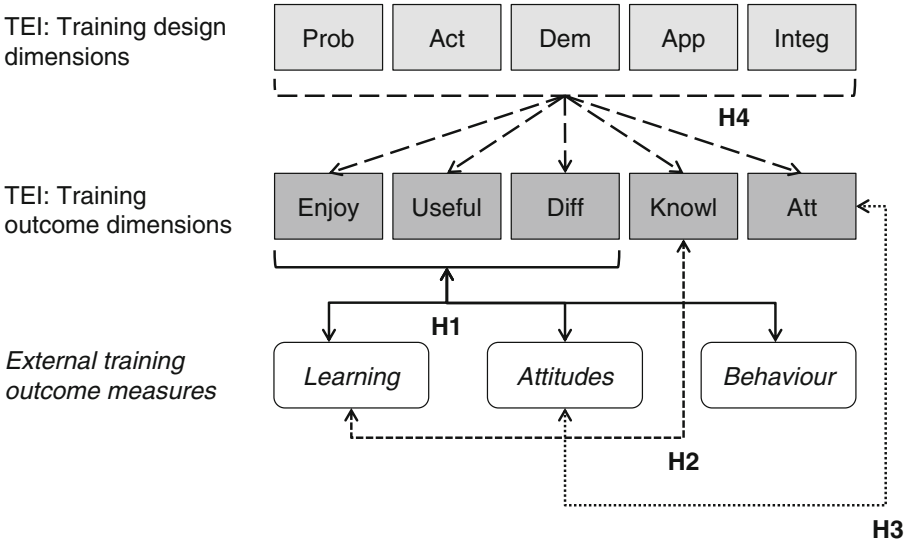
et al. 2010; Salas and Cannon-Bowers 2001). For example, the medium effect size of training found in a meta-analysis by Arthur and colleagues (Arthur et al. 2003) varied considerably not only depending on the criterion measured (reactions, learning, behaviour, results), but also on the training method used (lecture, simulation etc.). In a meta-analysis on the effectiveness of managerial training (Burke and Day 1986), different training methods were found to yield different effect sizes, and the magnitude of the effects also depended on the criteria used to operationalize training outcomes. Further meta-analyses focussed on particular training methods and showed that error training (Keith and Frese 2008), behavioural modelling (Taylor et al. 2005), or practice in training (Arthur et al. 1998) lead to effective training outcomes.

Despite the impact of training design on training outcomes, in the past, barely any training evaluation studies have included training design. The reason for this might lie in the fact that an outcome view of training validity has dominated evaluation designs (Goldstein and Ford 2002). This view is the result of a summative approach to evaluation with a focus on the effectiveness of completed interventions. Conversely, evaluating training design is a reflection of a formative approach to evaluation that aims at understanding why certain kinds of results were achieved (Goldstein and Ford 2002). *The second objective of this article is thus to include the assessment and perception of training design features as a formative evaluation aspect in our approach to training evaluation.*

The Training Evaluation Inventory: Defining Requirements

The initial motivation for our research was a joint project with an organisation with around 5'000 employees. The mandate was to advance a training programme targeting non-technical teamwork knowledge, skills, and attitudes (e.g. decision making, communication, or coordination) of employees. As the project demanded large-scale evaluations of training modules with several hundred participants, which could only be accomplished by applying a questionnaire directly after training, we decided to design a paper-and-pencil training evaluation inventory (TEI) that should fulfil a range of predefined requirements: 1) It should be *based on past empirical results and theoretical considerations* regarding training evaluation and design in order to be maximally informative with respect to the quality and effects of the evaluated training modules. 2) It should allow a *comparison of different training modules* with different content and objectives, designed for different target groups, in different organisations or organisational units. It should thus be generic and independent of training content. 3) We aimed to compare not only the outcomes of the training modules in question, but also their design and the impact of training design on training outcomes. The questionnaire should thus *measure training outcome and training design variables alike*. 4) It should also be *accepted as meaningful and feasible by members of organisations* (training decision makers, trainers, and trainees).

Given these requirements, the literature on training evaluation was reviewed in order to find suitable constructs that could be operationalized in a paper-and-pencil inventory to be filled out by participants immediately after training and tested by confirmatory factor analyses. Based on the literature review, we then formulated hypotheses regarding the constructs included in the questionnaire and their interrelations. This hypothesis-testing approach was adopted to assess the construct validity of the instrument (Nunnally 1978; see Fig. 1 for a graphical illustration of the hypotheses).



Notes. Prob = problem-based learning; Act = activation; Dem = demonstration; App = application; Integ = integration; Enjoy = subjective enjoyment; Useful = perceived usefulness; Diff = perceived difficulty; Knowl = subjective knowledge gain; Att = attitude towards training

Fig. 1 Graphical illustration of the hypotheses

Training Outcomes: Theoretical and Empirical Background

An important and influential framework for categorising training outcomes is Kirkpatrick's hierarchical model (Kirkpatrick 1998; Tannenbaum and Yukl 1992). The basic *reactions* level measure is enjoyment of training, but several studies have further differentiated reactions into enjoyment of training (affective reaction), usefulness of training (utility reaction), and perceived difficulty (Alliger et al. 1997; Warr and Bunce 1995). On the level of *learning and attitudes*, evaluations focus on the acquisition of declarative or procedural knowledge, but also on attitude changes. A more fine-grained classification for the learning and attitude level distinguishes cognitive (knowledge or cognitive strategies), skill-based (proceduralisation, automaticity etc.), and affective outcomes (attitudes, motivation, self-efficacy etc.; Kraiger et al. 1993). The *behavioural level* covers job-related behaviour and performance after training and indicates transfer of training to the job (Warr et al. 1999). Finally, *organisational results* such as reduced costs, an increase in production or sales, or a decrease in accident rates relate the training programme to organisational objectives.

Kirkpatrick's framework has inspired a great deal of research, but it has also undergone substantial criticism in the past two decades. The model's simplicity is appealing, but can also turn into a liability when its coarseness leads to overgeneralisations and misunderstandings (Alliger et al. 1997; Alliger and Janak 1989; Kraiger 2002). Holton (1996) furthermore stated that the four levels should rather be considered a heuristic taxonomy than a theoretically informed and experimentally confirmed model. Despite the critical appraisal of the framework, it remains the prevalent classification scheme in academic research and the most influential evaluation approach among practitioners (Aguinis and Kraiger 2009; Hochholdinger et al. 2008). It was thus established as point

of departure in the present study because of the large body of literature and research to it. This corresponds to the requirement defined for the inventory that it should be based on past empirical results and theoretical considerations. Moreover, its simplicity, making it readily understandable also by practitioners, made it feasible with regard to the requirement that the resulting inventory should be accepted by members of organisations. Based on a review of the literature, reactions as well as learning and attitudes were chosen for inclusion in the TEI. These two outcome levels are traditionally measured in a paper-and-pencil format, which again was a pre-defined requirement for the evaluation questionnaire.

Level 1: Reactions

Reactions are the most commonly collected training criteria in organisations (Bassi et al. 1996; Blanchard et al. 2002; Twitchell et al. 2000), but at the same time, they have been criticised by training scholars as insufficient to suggest learning or behavioural change (Tannenbaum and Yukl 1992). From a practitioner's point of view, evaluating reactions is appealing for several reasons. First, using standard reaction questionnaires is time- and cost-efficient (Hochholdinger et al. 2008). Second, training practitioners seem to lack expertise in how to conduct evaluations of other levels (Tannenbaum and Woods 1992; Twitchell et al. 2000). Third, collecting information on trainee reactions is important to win participants over to a training programme and influence more distant variables such as "word-of-mouth" advertising (Alliger et al. 1997).

From a scholarly point of view, the important question regarding reactions is to what extent they can be used as surrogate indicators of learning or behaviour change (Alliger et al. 1997, p. 343). When reactions are treated as a single, not further differentiated construct, meta-analytic studies have shown mixed results. Alliger and Janak (1989) found only very small correlations between reactions and learning or behaviour in their early meta-analysis, while a recent meta-analysis showed that reactions predicted post-training declarative and procedural knowledge (Sitzmann et al. 2008). Other studies looked at different facets of reactions. They found that affective reaction measures ("enjoyment") correlate only weakly with learning or behaviour transfer measures, while utility reactions ("usefulness") are more strongly related to immediate learning and transfer and seem to be as good a predictor of transfer as behaviour/skill demonstrations (Alliger and Janak 1989; Alliger et al. 1997; Warr and Bunce 1995; Warr et al. 1999). A very recent meta-analysis came to similar conclusions by showing that expectancy and instrumentality (Vroom 1964), which are motivational constructs that are closely related to utility judgments, were strong predictors of transfer of training in professional contexts (Gegenfurtner 2011). Another aspect of reactions that has been distinguished in the literature is perceived difficulty. It was shown to correlate significantly with self-reported competence, knowledge, self-reported use of training content on the job and perceived value of training (Warr et al. 1999).

Based on these results, we conclude that *reaction data can provide valuable information in training evaluation, especially when not only affective reactions, but also utility reactions and perceived difficulty are measured.* Taking also their high acceptance in organisations into account, enjoyment, perceived usefulness, and perceived difficulty

were thus included in the training evaluation inventory. Moreover, within the framework of our hypothesis-testing approach to construct validation (Nunnally 1978), the following hypothesis was formulated:

1. H1: The three reaction components enjoyment, perceived usefulness, and perceived difficulty as measured by the TEI are significantly and positively related to learning, attitudes, and behavioural transfer.

If indeed reactions can serve as a surrogate indicator for learning or attitude and behavioural change, then we should be able to confirm this hypothesis in a training study by applying the TEI and measures of learning, attitudes, and behavioural transfer.

Level 2: Learning and Attitudes

Learning as the acquisition of declarative or procedural knowledge, or a change in attitudes or values are common objectives of training (Warr et al. 1999), and effects on behaviour on the job or organisational results cannot be achieved without some form of cognitive, affective or skill-related change (Kraiger 2002). On level 2 of Kirkpatrick's model, learning has been measured considerably more often as an outcome of training than changes in attitudes, and the sample-weighted mean effect size *d* for studies measuring learning as a criterion of training effectiveness was shown to be medium to large (Arthur et al. 2003; Cohen 1992, describes *d* of 0.20, 0.50, and 0.80 as small, medium, and large effect sizes, respectively). Meta-analytic studies on the relationship between learning and behaviour or transfer found relationships ranging from small (Alliger et al. 1997) to moderate to high (Colquitt et al. 2000).

Learning and attitudes can be assessed through a variety of methods, e.g. questionnaires, exercises, or work samples (Kraiger 2002; Salas et al. 2006b). Usually, tests of learning and knowledge target the content discussed in training (e.g. declarative knowledge of health and safety regulations; Gegenfurtner 2012), and attitude questionnaires typically assess attitudes towards the attitude object discussed in training (e.g. attitudes toward disability and accessibility; Lewis 2009). As outlined above, it was established that the instrument we designed should be generic and content-independent. It was thus not feasible to include items measuring learning of and attitudes towards specific training content. Instead, we chose to measure self-reported, subjective knowledge gain and attitude towards the training module as a whole.

Self-assessments of knowledge have been defined as learners' estimates of how much they know (current knowledge level) or have learned (increase in knowledge level) about a domain (Sitzmann et al. 2010). In research on instructional communication, the use of students' self-assessments of knowledge has had a long tradition in terms of enabling the study of cognitive learning in teaching settings while generalising across subject areas (Chesebro and McCroskey 2000). Measures of students' perception of learning as used in this strand of research showed a moderate to strong correlation with students' test scores in an experimental teaching setting (Chesebro and McCroskey 2000). A recent meta-analysis by Sitzmann et al. (2010) also showed that self-assessments of knowledge correlated moderately with cognitive learning. Self-reported knowledge gain can thus serve as a proxy for learning, although the relationship is less than optimal and results have to be interpreted with caution. *We included subjective knowledge gain as an outcome in the TEI in order to*

gather information which is impossible to obtain through other channels in a generic training evaluation setting. We therefore aimed at reducing the criterion deficiency of the entire evaluation effort (Campbell and Lee 1988). The following hypothesis was formulated to enable us to assess the relationship of self-assessments of knowledge and objective learning measures within the hypothesis-testing approach to construct validity (Nunnally 1978):

2. H2: Subjective knowledge gain as measured by the TEI is significantly and positively related to objective learning measures as criteria of training outcome.

With regard to attitudes, we again chose a generic approach and measured attitudes towards the training modules as a whole, i.e. participants' opinion of the fact that training is offered. To the best of our knowledge, there are no previous studies which have investigated the relationship between general attitudes towards training on the one hand and attitudes towards specific training topics on the other hand. We argue that these two forms of attitudes are related. A central component of attitudes is evaluation, defined as the degree of favour or disfavour expressed towards the attitude object (Olson and Zanna 1993). As the degree of favour towards a training module is at least partly determined by the degree of favour towards its different topics, we assume that the *general attitude towards a training module is related to the attitudes towards specific training topics*, leading to the following hypothesis:

3. H3: Attitudes towards training as measured by the TEI are significantly and positively related to attitudes towards specific training topics.

In summary, we included subjective enjoyment, perceived usefulness, and perceived difficulty as components of reactions in the TEI. Additionally, subjective knowledge gain and attitude towards training as components of learning are measured. These five constructs are henceforth referred to as *training outcome dimensions*. In line with the hypothesis-testing approach to construct validity (Nunnally 1978), three hypotheses were formulated to test the relationship of the dimensions with other constructs.

Training Design: Theoretical and Empirical Background

Instructional and training design deals with the question of what it takes to help participants learn effectively in terms of instructional activities (Gagné et al. 2005). Such considerations are important because instructional factors such as learner-centred environments are positively related to transfer (Gegenfurtner 2011). Consequently, evaluating the design features of training is important to shed light on the reasons why certain training outcome effects were produced. This approach is in line with a formative conceptualisation of evaluation and provides the organisation with important feedback about how a training programme could be improved (Goldstein and Ford 2002). To enable the measurement of training design in the TEI, we had to take a normative approach by specifying generally valid and accepted instructional principles that could then be operationalized in a questionnaire. In an effort to find the essence of a wide range of recent instructional design theories, Merrill (2002) identified five underlying principles common to these theories, called the "five first

principles of instruction": problem-based learning, activation, demonstration, application, and integration. Descriptions of these principles are given in Table 1. Past research adapted the first principles in a questionnaire and showed significant relationships between learner ratings of the first principles of instruction and learning time, self-reported learning progress, satisfaction with the course, mastery of course objectives, and overall course quality (Frick et al. 2009). Based on these results, we chose to focus on the five first principles as evaluation dimensions of training design, asking for participants' subjective perception of their application. According to Merrill (2002, pp. 43–44), the first principles always hold under appropriate conditions, regardless of instructional practices or programmes. Learning by a given programme will be fostered in direct proportion to the implementation of the first principles. We adopted these propositions for the scales covering the five first principles of instruction, henceforth referred to as *training design dimensions*, which led us to the following hypothesis regarding the TEI within the hypothesis-testing approach to construct validation (Nunnally 1978):

4. H4: The training design dimensions serve as antecedents of training outcomes. Training design thus predicts training outcomes. The higher the ratings for training design, the better the training outcomes of participants as measured by the TEI.

Method

Questionnaire Development

Questionnaire development was conducted within a larger study in collaboration with a European Airline with roughly 5'000 employees and a German fire service with 700 employees. The training programmes which served as the data source were all variations

Table 1 Description of the "five first principles of instruction" (Merrill 2002)

Principle	Description
Problem-based learning	- Learning is facilitated when learners work on real-world cases or problems - Learners should be shown what they will be able to do after completion of training
Activation	- Learning is promoted when previous experience of the learners is activated or when the learners are provided with relevant experience that can serve as a foundation for new knowledge (or skills or attitudes)
Demonstration	- Learners should be demonstrated with what has to be learned, and not merely told, using multiple representations and suitable media - Demonstration should be consistent with learning goals and direct attention to the relevant information
Application	- Learning is facilitated when learners can practise their new knowledge consistently with the learning goals and receive (gradually diminishing) feedback
Integration	- Learning is promoted when learners can integrate their new knowledge into their existing knowledge and transfer it to everyday life - Integration is facilitated by discussions, creations of new and own ways to use the knowledge, or the possibility to demonstrate new knowledge and skills

of team training programmes, which seek to improve team coordination, teamwork skills, and team performance in order to reduce human error and increase safety (Helmreich et al. 1999; Salas et al. 2006a) in high reliability organisations (Weick and Sutcliffe 2007). Training content consisted of teamwork competencies such as communication skills, situation awareness or decision making (Salas et al. 1999).

The questionnaire development began in 2008 and was guided by the predefined requirements which the instrument should fulfil: It should be a) based on empirical results and theoretical considerations, b) generic and independent of training content, c) able to measure training outcome and training design variables alike, and d) accepted as meaningful and feasible by members of the organisation.

Development included three main phases. In the first phase, items for the five training outcome and the five training design dimensions were formulated in the form of statements (e.g. "Contents were illustrated with concrete examples"). As for the answer format, a five-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") was chosen. The aim of this phase was to develop a short and concise instrument with as small a number of items as possible. This resulted in a first version of the questionnaire with 26 items. The number of items per scale was (*number in brackets*): subjective enjoyment (*3*), perceived usefulness (*3*), perceived difficulty (*5*), subjective knowledge gain (*2*), attitude towards training (*3*), problem-based learning (*2*), activation (*2*), demonstration (*3*), application (*1*), and integration (*2*). The first version was tested in a pre-study in four different team training modules with differing contents (55 participants in total; 68.5 % male). The mean age was 33.1 years (*SD*=9.3). Analysis of the reliability of the scales covering the dimensions (internal consistency: Cronbach's α) showed that three scales had a consistency below .6 (see Table 2), which is the lowest value considered as acceptable in the literature for low-stakes instruments designed for programmatic decision-making that do not have direct consequences for individuals, such as in this case (Wasserman and Bracken 2003). Furthermore, training observations by the authors showed that the existing items did not cover every relevant aspect of the different training design dimensions (e.g. the aspect of training objectives in the demonstration dimension, or the feedback aspect in the application dimension). Minor changes were also made to the training outcome dimensions. Although the questionnaire was indeed concise, with completion times rarely exceeding 10 min, it was decided that inclusiveness was more important.

Thus, in the second phase, 20 additional items in eight scales were formulated and one redundant item was removed. Two surveys were then conducted using the extended scales, with 482 and 470 training participants, respectively. Both surveys are presented in more detail below. The reliabilities of the extended scales were deemed acceptable, and the feedback from organizational stakeholders was positive with regard to the applicability of the questionnaire and to the conclusions that could be drawn based on the results. The whole questionnaire in its extended form was subsequently used in two training study samples, also explained in more detail below. Reliability values for all samples and the mean overall reliability are displayed in Table 2. The complete questionnaire can be found in the appendix.

In the third and final phase, the data from the second survey was subjected to two confirmatory factor analyses[1] in order to determine the factor structure of the questionnaire (the final models are described in the results section). A Missing Value

[1] Statistical software used: SPSS Amos 19

Table 2 Descriptive statistics and reliability (internal consistency: Cronbach's α) of all TEI dimensions for the pre-study and samples S1, S2, TrS1, and TrS2

	Pre-study (*N*=55)		Sample *S1* (*N*=482)		Sample *S2* (*N*=470)		Sample *TrS1* (*N*=81)		Sample *TrS2* (*N*=46)		Mean reliability[a] (excluding pre-study)
	M (SD)	*r*	*M (SD)*	*r*	*M (SD)*	*r*	*M (SD)*	*r*	*M (SD)*	*r*	
Training outcome dimensions											
Subjective enjoyment	4.30 (.64)	.85	4.11 (.64)	.82	4.17 (.61)	.81	3.93 (.69)	.85	3.85 (.61)	.76	.81
Perceived usefulness	4.11 (.79)	.88	4.04 (.75)	.87	4.09 (.74)	.90	3.82 (.81)	.88	3.43 (.81)	.91	.89
Perceived difficulty	4.30 (.57)	.79	4.52 (.45)	.70	4.42 (.50)	.63	4.46 (.54)	.81	3.89 (.54)	.75	.73
Subjective knowledge gain	3.67 (.70)	.54	3.73 (.73)	.63	3.67 (.73)	.80	3.48 (.75)	.72	3.31 (.73)	.83	.75
Attitude towards training	4.05 (.68)	.77	4.00 (.75)	.81	3.92 (.75)	.81	3.73 (.67)	.73	3.56 (.77)	.87	.81
Training design dimensions											
Problem-based learning	3.63 (.82)	.58	3.75 (.64)	.71	4.00 (.58)	.75	4.08 (.55)	.75	–	–	.74
Activation	3.81 (1.02)	.86	3.87 (.70)	.81	3.89 (.63)	.77	3.76 (.83)	.85	3.70 (.60)	.80	.81
Demonstration	4.21 (.67)	.68	4.23 (.53)	.74	4.41 (.45)	.76	4.47 (.47)	.81	3.96 (.46)	.74	.76
Application	3.61 (1.02)	*single item*	3.00 (.95)	.81	3.25 (.82)	.82	3.26 (.72)	.70	3.04 (.67)	.74	.77
Integration	3.96 (.75)	.52	3.84 (.60)	.60	3.88 (.60)	.74	3.83 (.57)	.66	3.64 (.50)	.62	.66

Scales ranged from 1 to 5; [a] Calculated using Fisher's Z transformation; *S1* survey sample 1; *S2* survey sample 2; *TrS1* training study sample 1; *TrS2* training study sample 2

Analysis with the software SPSS Statistics 20 showed that values in the dataset were missing completely at random (Little's MCAR test: χ^2=6624.7, *df*=6,629, *p*=.51). Thus, only complete cases (*N*=245) were used to avoid problems arising from data imputation procedures (Tabachnick and Fidell 2007). The scales of the two parts of the questionnaire were analysed separately because they cover conceptually distinct aspects. Data were transformed into a correlation-covariance matrix within SPSS and imported into Amos. Model convergence was an iterative process and the final models were overidentified. The method of estimation applied was the maximum likelihood (ML) method, because it is also suitable for small samples. To determine the goodness of fit, the model fit indicators Comparative Fit Index (CFI), Tucker Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA), and Standardized Root Mean Residual (SRMR) were used. A model fit of CFI and TLI above .80 is considered acceptable; above .90 the fit is good. A RMSEA weaker than .08 is acceptable, while a value of .06 and lower is good. SMRS values below .08 are good. Regarding Chi^2, a value between 1 and 3 should follow from dividing the value by the degrees of freedom (for detailed description of model fit indicators see Hu and Bentler 1999). Hu and Bentler (1999, p. 27) recommend examining the combination of the two indicators TLI/CFI and SRMR in particular.

Data Collection

To test the hypotheses, the TEI was used in four samples between 2008 and 2010, of which two were large survey samples and two were drawn from quasi-experimental training studies. In the surveys, solely the TEI was implemented on a large scale basis, while in the smaller scale training studies, further variables in addition to the TEI were measured. The training study samples served to test hypotheses 1 to 3, which are concerned with the interrelations of the TEI training outcome dimensions and their relationships with other outcome measures such as objective learning. Hypothesis 4, introducing training design as an antecedent of training outcomes, was tested using a merged dataset including the survey samples and one training study sample. The second training study sample was not part of the merged dataset. Due to the setting of the study, no data had been collected on the training design dimension of problem-based learning. Hence, the data could not be used to test hypothesis 4.

Survey Samples

Data for the first survey sample (S1; see Table 3 for descriptive statistics) were collected from April to December 2008. The objective of this survey was to gain a complete picture of the status quo of all team training activities conducted by the collaborating airline regarding subjective perception of their training design, reactions of participants, and learning and attitudes of participants. Based on the results, modifications of certain training activities were planned (for results concerning cockpit and cabin crews, see Ritzmann et al. 2009; Ritzmann 2012). Thirteen team training modules were included, and data of *N*=482 course participants could be collected. Professional groups included were pilots, cabin crew members, maintenance technicians, and air traffic controllers. The course length varied from half a day

Table 3 Frequencies and descriptive statistics regarding profession, gender, and age from samples TrS1, TrS2, S1, and S2

	Sample TrS1 (*N*=81)		Sample TrS2 (*N*=46)	Sample S1 (*N*=482)	Sample S2 (*N*=470)	
	Training condition 1	'Training condition 2			Training module 2009	Training module 2010
Profession	Cabin Crew: 38	Cabin Crew: 43	Fire Fighters: 46	Pilots: 146 Cabin Crew: 259 Maintenance: 58 ATC: 19	Cabin Crew: 190	Cabin Crew: 196 Pilots: 84
Gender	Male: 6 Female: 32 Missing: -	Male: 4 Female: 36 Missing: 3	Male: 43 Female: 3 Missing: -	Male: 256 Female: 220 Missing: 6	Male: 44 Female: 147 Missing: 4	Male:130 Female: 147 Missing: 6
Mean Age[a] (*SD*)	26.1 *(6.0)*	23.2 *(3.4)*	≤ 20: 1 21–30: 14 31–40: 16 41–50: 9 > 50: 6	30.6 (*10.2*)	33.9 *(11.1)*	40.9 *(9.6)*

[a] Age categories for sample TrS2

to three days. Questionnaires were administered by course instructors and regularly collected by the authors or sent to them per mail.

Data for the second survey sample (S2; see Table 3 for descriptive statistics) were collected in October and November 2009 as well as from April to July 2010 with participants of the annual refresher team training 2009 (*N*=190 cabin crew members) and 2010 (*N*=196 cabin crew members, *N*=84 pilots), respectively. The objective of this second survey was to compare the training module of 2009 with the training module of 2010 regarding the training design and training outcome dimensions covered in the questionnaire. The two training modules differed in their format and training design, with the training module of 2010 being more problem-oriented and linking non-technical skills training with more technical safety skills with the aim of creating an integrated training experience (Ritzmann et al. 2011). In both data collection periods, trainers were asked to distribute questionnaires in three of four courses per week. Completed questionnaires were collected and handed to the authors. Training lasted for half a day.

Training Study Samples

The first training study sample (TrS1) stems from a study carried out from June to September 2010. Four classes of newly hired junior flight attendants (*N*=81) in their four-week initial cabin crew member course served as participants. Data were collected from their one-day initial team training module. Questionnaires were distributed after training and collected directly from the participants after completion. The sample is described in further detail in Table 3. The aim of the study was to compare the effects of two different team training conditions (attitude-oriented vs. competency-based training). The training conditions and results are described elsewhere (Ritzmann 2012). The trainer was held constant for all four groups to minimise the influence of instructor style on trainee reactions (Sitzmann et al. 2008). In

addition to the administration of the TEI, teamwork attitudes, knowledge, and behavioural intentions were measured. Data were collected before training in order to establish a baseline (T0), directly after training (T1), and again 8 weeks later (T2). Between training and data collection 8 weeks later, junior flight attendants completed a phase of introductory flights (for more information, see Ritzmann 2012).

Data for the second training study sample (TrS2) were collected in July 2010 in collaboration with a German fire service. Participants, all fully qualified fire fighters (N=46), received a half-day team training module in groups of 10 to 15 people. Questionnaires were distributed one day after training and collected directly from the participants after completion. Descriptive statistics of the sample can be found in Table 3. The aim of the study was to evaluate the acceptance and success of team training within the fire service. For the results, see Hagemann (2011) and Hagemann et al. (2012). Again, the trainer was held constant for all four groups in order to minimise the influence of instructor style (Sitzmann et al. 2008). Due to the different professional category, the wording of questionnaire items was adapted if necessary (e.g. "I know the importance of the different *crew resource management topics* in various situations" vs. "I know the importance of different *team competencies* in various situations"). Due to time constraints, no problem-based learning could be implemented within the training and the corresponding training design dimension was not included in the questionnaire for this sample. In addition to the administration of the TEI, knowledge as well as attitudes towards leadership, debriefing, appraisal of stress and human fallibility were measured (for more information, see Hagemann 2011). Data were collected before training in order to establish a baseline (T0) and one day after training (T1). An additional experimental variation with an impact on knowledge and attitudes (a form of team debriefing) took place before the second post-test T2. Hence, data from T2 were not used, because they were influenced by the second experimental variation as well (see Hagemann 2011).

Measures

In all four samples, the TEI was used to collect data on training design and training outcomes as perceived by the participants. The descriptive statistics and reliabilities of the scales can be found in Table 2. All items of the TEI had to be answered on a five-point Likert scale ranging from "strongly disagree" to "strongly agree". Except for perceived difficulty, where a higher score indicates less difficulty, higher scale scores represent higher values assigned to the underlying dimension.

Additional measures of external training outcomes (henceforth referred to as external training outcome measures) were used in the training studies to test hypotheses 1 to 3. They are described in more detail in Table 4.

Results

Confirmatory Factor Analyses

The results of the confirmatory factor analyses are displayed in Figs. 2 and 3. The factor loadings of all items were above .40 (except one) and significant. The indicators showed

Table 4 Description of external training outcome measures used in the training study samples TrS1 and TrS2

Measure	Description	Sample item	Format
Sample TrS1 (Cabin Crew)			
Teamwork attitudes	- 9 items adapted from existing attitude questionnaires (e.g. Operating Room Management Attitudes Questionnaire (ORMAQ); Yule et al. 2004) - 2 items constructed for the study	*I am ashamed when I make a mistake in front of other team members*	Five-point Likert scale (strongly disagree to strongly agree)
Knowledge	- 7 knowledge questions tailored to training content	*Please indicate whether the following statement is correct or not: 'Stress is the result of the appraisal of a situation and arises when you feel you don't have enough resources to manage the demands'.*	1 open question 1 short open question 5 true-false questions
Behavioural intentions	- 3 short scenarios in written form, requiring open answers in own words - 1 scenario where the most adequate reaction had to be chosen from a given selection	*During boarding, you notice that the passenger on seat 29C is sweating and breathing heavily. He looks off-colour. What do you do?*	3 open questions 1 multiple-choice question
Sample TrS2 (Fire Fighters)			
Knowledge	- 13 knowledge questions tailored to training content	*Please specify why feedback is important for teamwork.*	7 open questions 5 true-false questions 1 mapping task
Attitudes			
Leadership	- 5 items adapted from ORMAQ (Sexton et al. 2000; Yule et al. 2004)	*Senior staff should encourage questions from junior staff during missions:*	Five-point Likert scale (strongly disagree to strongly agree)
Debriefing	- 2 items adapted from ORMAQ (Sexton et al. 2000)	*A regular debriefing of procedures and decisions after a mission is an important part of teamwork.*	
Appraisal of Stress	- 2 items adapted from Crewmember Attitudes Questionnaire (CAQ; McDonnald and Shadow 2003), 1 item adapted from ORMAQ (Sexton et al. 2000)	*Personal problems can adversely affect my performance*	
Human Fallibility	- 2 items adapted from ORMAQ (Sexton et al. 2000)	*I am more likely to make errors in tense or hostile situations.*	

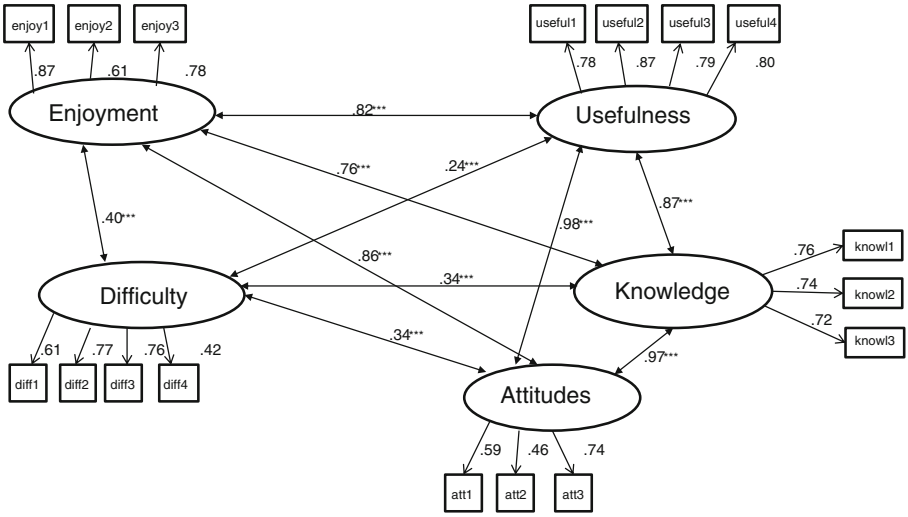


Fig. 2 Results of the confirmatory factor analysis of the training outcome dimensions

acceptable model fit for the five training outcome dimensions (χ^2=261.02, *df*=101, *p*=.000, χ^2/df=2.58, CFI=.97, TLI=.95, RMSEA=.059, CI=.050–.067, SRMR=.055). The indicators also showed acceptable model fit for the five training design dimensions (χ^2=708.22, *df*=302, *p*=.000, χ^2/df=2.35, CFI=.92, TLI=.92, RMSEA=.054, CI=.049–.059, SRMR=.062). The final questionnaire consisted of 45 items.

According to Hair et al. (2006), a construct shows convergent validity if variance-extracted measures exceed the 50 % level and Cronbach's α is larger than .70. Reliability estimates show that the average variance extracted (AVE) regarding reported enjoyment (58 %), perceived usefulness (66 %), perceived difficulty (46 %), learning knowledge (59 %), and learning attitudes (60 %) almost reached or exceeded the 50 % level. Also the AVE regarding problem-based learning (46 %),

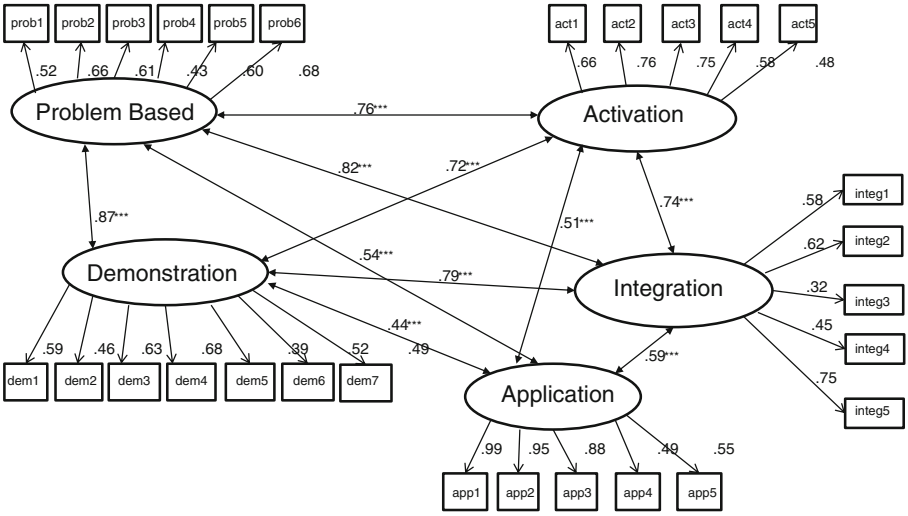


Fig. 3 Results of the confirmatory factor analysis of the training design dimensions

activation (49 %), demonstration (44 %), application (64 %), and integration (45 %) almost reached or exceeded the 50 % level. We did not exclude more variable indicators for increasing AVE due to training relevant information of the items. The values of Cronbach's α are displayed in Table 2 and exceed α=.70, except for one scale (training design dimension "integration"). These values are also acceptable for low-stakes instruments designed for programmatic decision-making (Wasserman and Bracken 2003). Thus, the theoretically developed structure of the two parts of the inventory measuring training outcomes and training design could be supported.

Testing of the Hypotheses

Following our hypothesis-testing approach to assess construct validity (Nunnally 1978), the four hypotheses formulated for relations of the TEI dimensions with other measures and with each other were tested after the confirmatory factor analyses.

Correlational Analyses of Training Outcome Dimensions and External Training Outcome Measures

Correlational analyses involving the training outcome dimensions of the TEI and external training outcome measures implemented in the training study samples were performed on the respective datasets to test hypotheses 1 to 3. The descriptive statistics and the correlations of the variables involved are displayed in Table 5.

First-order partial correlations were calculated between the training outcome dimensions and the external outcome measures after training (T1 and T2), controlling for the baseline level at T0 to isolate effects of training independent of prior attitudes, knowledge, or behavioural intentions (see Warr et al. 1999, for a discussion of this procedure). Table 5 shows the first-order partial correlations of the five training outcome dimensions with the training effectiveness measures employed in the two training studies. A range of significant positive relationships with small to large effect sizes can be reported, with r=.10 being a small, r=.30 being a medium and r=.50 being a large effect (Cohen 1992).

To test hypothesis 1, correlations involving the three reaction components subjective enjoyment, perceived usefulness and perceived difficulty were analysed first. Subjective enjoyment showed significant correlations with knowledge of training content. Generally speaking, participants who enjoyed the training course more had more knowledge of its content. In sample TrS1, comprising flight attendants, the effect was not significant immediately after training at T1, but was significant 8 weeks later at T2. In sample TrS2, comprising fire fighters, subjective enjoyment was significantly related to knowledge immediately after training at T1. Thus, the more participants enjoyed the training, the more of the training content they remembered.

Perceived usefulness showed a significant correlation with attitudes towards teamwork and knowledge of training content. The relationship with attitudes was found in sample TrS1, comprising flight attendants. The more useful trainees rated the training, the more positive were their attitudes towards teamwork immediately after training (T1) and also 8 weeks later (T2). The relationship between perceived usefulness and knowledge could be observed in sample TrS2, comprising fire fighters, meaning that the more useful these participants rated the training, the more knowledge of training content they showed after training.

Table 5 Correlations of the training outcome and training design dimensions with external training outcome measures of the training study samples TrS1 and TrS2 (pairwise N is given below correlation coefficients)

			Training outcome dimensions					Training design dimensions				
		M (SD)	Subjective Enjoyment	Perceived Usefulness	Perceived Difficulty	Subjective Knowledge Gain	Attitude towards Training	Problem-based learning	Activation	Demonstration	Application	Integration
Sample TrS1: Junior Flight Attendants	Attitude towards teamwork (scale: 1–5)											
	T1 (*T0 held constant)*	4.05 (.30)	.089	.246*	.112	.156	.258*	.181	.114	.060	.226+	.107
			76	75	77	74	77	74	70	74	69	74
	T2 (*T0 held constant)*	4.01 (.30)	.098	.255*	.148	.155	.255*	.085	−.043	.216+	−.069	−.012
			73	73	74	71	74	71	69	70	68	73
	Knowledge of training content (0–16 pts.)											
	T1 (*T0 held constant)*	7.91 (2.91)	−.050	.025	.198^{+}	.048	.042	.179	.078	.160	.250*	.143
			72	72	73	69	73	68	64	69	64	70
	T2 (*T0 held constant)*	6.44 (2.24)	.248*	.196	.271*	.290*	.247*	.347**	.262*	.443**	.187	.284*
			69	69	70	66	70	67	63	66	63	68
	Behavioural intentions (−3.5–12.5 pts.)											
	T1 (*T0 held constant)*	4.47 (1.53)	.022	.042	.166	.110	.204	.176	−.034	.212	.180	.364**
			56	55	56	53	56	54	50	54	49	53
	T2 (*T0 held constant)*	4.05 (1.64)	.010	−.140	−.089	−.059	−.051	.008	−.091	.088	.003	.153
			60	59	60	57	60	58	54	58	54	58
Sample TrS2: Fire Fighters	Knowledge (0–60 pts.)											
	T1 (*T0 held constant)*	35.41 (8.94)	.430*	.518**	.369+	.408*	.546**	–	.488*	.489*	.364+	.449*
			28	27	27	28	28		28	27	28	28
	Attitude (scale: 0–4)											
		3.06 (.37)	.427*	.365+	.124	.208	.292	–	.317	.384+	.632**	.240

Table 5 (continued)

	M (SD)	Training outcome dimensions					Training design dimensions				
		Subjective Enjoyment	Perceived Usefulness	Perceived Difficulty	Subjective Knowledge Gain	Attitude towards Training	Problem-based learning	Activation	Demonstration	Application	Integration
Leadership T1 (*T0 held constant)*		28	27	27	28	28		28	27	28	28
Debriefing T1 (*T0 held constant)*	3.32 (.55)	.447*	.293	.066	.267	.253	–	.136	.239	.299	.373+
		28	27	27	28	28		28	27	28	28
Appraisal Stress T1 (*T0 held constant)*	2.77 (.61)	.332+	.249	−.142	.142	.302	–	.347+	.284	.441*	.239
		28	27	27	28	28		28	27	28	28
Human Fallibility T1 (*T0 held constant)*	2.40 (.62)	.462*	.380+	.116	.269	.396*	–	.368+	.658**	.417*	.450*
		28	27	27	28	28		28	27	28	28

T0 baseline before training; *T1* measurement immediately after training; *T2* measurement 7.5 weeks after training; $^{+}$ $p<.10$, * $p<.05$, ** $p<.01$ (two-tailed)

Perceived difficulty, the third reaction component, correlated significantly with knowledge of training content. The relationship could be observed at T2 in the flight attendant sample TrS1, meaning that the easier trainees found the training, the more of it they remembered after 8 weeks. No significant correlations of perceived difficulty and other outcome measures could be found in the fire fighters sample TrS2.

To sum up, we found significant relationships between reactions, as measured with the TEI, and learning and attitudes after training, as measured with other instruments. Subjective enjoyment showed the highest number of significant correlations, and perceived difficulty the lowest. These results are in line with our first hypothesis. However, contrary to our expectations, we did not find any significant relationships between reactions and behavioural intentions. Furthermore, the correlations that emerged were not entirely consistent between the samples. We thus consider hypothesis 1 as partially confirmed.

Hypothesis 2 stated that subjective knowledge gain as measured by the TEI is related to objective learning measures. In support of the hypothesis, significant correlations of subjective knowledge gain and knowledge of training content could be found in both samples, TrS1 and TrS2, although in the flight attendant sample TrS1, the relationship was only significant at T2, 8 weeks after training, and not at T1. In terms of the discriminant validity of the dimension, it can be noted that it did not significantly correlate with any other external outcome criteria (i.e. attitudes and behavioural intentions). In summary, these results show strong, albeit incomplete support for hypothesis 2.

Hypothesis 3 was tested by examining the relationships between attitudes towards training as measured by the TEI and attitudes towards the training topics covered in the two training studies. In sample TrS1, with flight attendants as participants, attitudes towards training indeed correlated significantly with attitudes towards teamwork at both times of measurement (T1 and T2). The more positive the attitude towards training was as a whole, the more positive were the specific attitudes towards teamwork. Looking at sample TrS2, comprising fire fighters, attitude towards training correlated significantly with attitudes towards human fallibility. Although the remaining attitude measures were not significantly related to the TEI attitude dimension, two of three correlations showed medium effect sizes around r=.30. Regarding the discriminant validity of the dimension, significant correlations with knowledge of training content could be observed in both samples. The discriminant support is thus less strong than for subjective knowledge gain. Summing up, the evidence for hypothesis 3, it was confirmed in sample TrS1 and partially confirmed in sample TrS2, leading to an overall partial confirmation.

Multiple Regression Analyses of Training Design Dimensions as Antecedents of Training Outcomes

Multiple regression analyses were performed to test hypothesis 4, which assumes that the training design dimensions serve as antecedents of training outcomes. To accomplish this, it was decided to combine the dimension scores as well as demographic information on sex, age, and prior experience with team training from all suitable samples in one dataset. Data of sample TrS2 were not included, as it did not contain information on the problem-based learning dimension, which had been excluded from the questionnaire for this particular study. The combined dataset consisted of

N=1,041 respondents (43.1 % male). The mean age was 33.5 years (SD=11.1). Approximately three quarters of respondents (75.9 %) had prior experience with team training. Intercorrelations of the TEI dimensions and age were calculated. Due to the large sample size, almost all of the 55 correlations were significant but the correlations showed no multi-collinearity. All the correlations between the TEI dimensions were significant at p<.01 (ranging from r=.13 to r=.82) with the exception of application and perceived difficulty (r=.02, *ns*). Age resulted in eight out of ten significant correlations with the dimensions, but all of them showed a small magnitude (ranging from r=−.15 to r=.21).

Five hierarchical multiple regression analyses on the combined dataset were then carried out using the demographic variables and the five training design dimensions as predictor variables and the training outcome dimensions as criteria. Demographic variables were entered in the regression to control for "third variable" effects (Field 2005). Sex was coded as 0 (male participants, 45.4 % of the sample) and 1 (female participants). Prior experience with team training was coded as 0 (no prior experience, 27.5 % of the sample) and 1.

Subjective Enjoyment

Sex, age, and prior experience with team training were able to account for 6.1 % of the variance in enjoyment of training (p<.001, f^2=.06). The training design dimensions accounted for another 26 % of variance, leading to a total explained variance of 32.2 % (p<.001, f^2=.47; see Table 6). Prior experience in team training and the training design dimensions demonstration, application, and integration emerged as significant predictor variables in the model. As the positive beta values show, more prior experience as well as more demonstration, application, and integration led to more enjoyment of training.

Perceived Usefulness

With regard to perceived usefulness, the demographic variables entered in the first step accounted for 2.4 % of the variance. This small effect was significant (p<.01, f^2=.02). The training design dimensions were able to explain a further 24.2 % of variance, resulting in 26.6 % of explained variance in the final model (p<.001, f^2=.36; see Table 6). Looking at the significant predictor variables, the positive beta value of sex indicates that male respondents seemed to find the training more useful for their job, as males were coded as zero. Furthermore, higher values of problem-based learning, demonstration, application and integration resulted in higher perceived usefulness.

Perceived Difficulty

Sex, age, and prior experience with team training accounted for 2.6 % of the variation in perceived difficulty of training, again a small but significant effect (p<.001, f^2=.03). An additional 15.5 % was explained by the training design dimensions, leading to 18.2 % of explained variance (p<.001, f^2=.22; see Table 6). The significant demographic predictor variables in the model show that female respondents and younger participants in our sample perceived training programmes to be less difficult. With regard to training design, activation, demonstration, and integration made training easier for participants.

Table 6 Multiple regressions of demographic variables and training design dimensions on training outcome dimensions

	Subjective enjoyment (N=697)[1]		Perceived usefulness (N=701)[2]		Perceived difficulty (N=687)[3]		Subjective knowledge gain (N=702)[4]		Attitude towards training (N=692)[5]	
	β	*Sig.*	β	*Sig.*	β	*Sig.*	β	*Sig.*	β	*Sig.*
Step 1										
Constant										
Sex	−.014	.73	−.070	.08	.095	.02*	−.033	.41	−.089	.03*
Age	.134	.00**	.082	.08	−.101	.03*	.112	.01*	.126	.01**
Team training experience	.144	.00**	.052	.24	−.007	.88	.011	.80	−.034	.45
Step 2										
Constant										
Sex	−.052	.12	−.107	.00**	.076	.04*	−.069	.05*	−.129	.00**
Age	.053	.17	.014	.74	−.121	.01*	.030	.46	.045	.26
Team training experience	.145	.00**	.051	.19	.015	.73	.001	.99	−.034	.37
Problem-based learning	.060	.15	.133	.00**	−.034	.46	.116	.01**	.082	.06
Activation	.079	.06	.033	.43	.093	.04*	.081	.06	.053	.21
Demonstration	.238	.00**	.231	.00**	.323	.00**	.151	.00**	.231	.00**
Application	.139	.00**	.116	.00**	−.142	.00**	.153	.00**	.140	.00**
Integration	.161	.00**	.153	.00**	.130	.00**	.170	.00**	.166	.00**

[1] R^2 =.061 for Step 1 (p<.001); ΔR^2 =.260 for Step 2 (p<.001); [2] R^2 =.024 for Step 1 (p<.01); ΔR^2 =.242 for Step 2 (p<.001); [3] R^2 =.026 for Step 1 (p<.001); ΔR^2 =.155 for Step 2 (p<.001); [4] R^2 =.018 for Step 1 (p<.01); ΔR^2 =.249 for Step 2 (p<.001); [5] R^2 =.026 for Step 1 (p<.001); ΔR^2 =.259 for Step 2 (p<.001)

* p<.05; **p<.01 (two-tailed)

However, respondents perceived training with higher application ratings to be more difficult, as evidenced by the negative beta value.

Subjective Knowledge Gain

Demographic variables only explained 1.8 % of variation in subjective knowledge gain, but despite its small size, the effect was significant (p<.01, f^2=.02). The training design dimensions entered in the second step of the regression accounted for a further 24.9 % of variance, resulting in 26.7 % of overall variance explained (p<.001, f^2=.36; see Table 6). Sex proved to be a significant demographic predictor, with male respondents in our sample reporting more knowledge gain than female respondents. Regarding the training design dimensions, higher values in problem-based learning, demonstration, application, and integration predict higher subjective knowledge gain.

Attitude Towards Training

Sex, age, and prior experience with training accounted for 2.6 % of the variance in attitude towards training (p<.001, f^2=.03). An additional 25.9 % could be explained

by the training design dimensions. The final model thus accounted for 28.6 % of the variation in attitude towards training ($p<.001$, $f^2=.40$; see Table 6). The significant demographic predictors show an influence of sex (male respondents in the sample had a more positive attitude towards training) and age (older participants showed a more positive attitude towards training). The design dimensions demonstration, application, and integration were also linked to more positive attitudes.

To sum up, training design dimensions explained around 25 % of the variance in the training outcome dimensions when the variance explained by the demographic variables was held constant, with the exception of perceived difficulty. Given the wide variety of possible factors that have an influence on training outcomes beside training design (e.g. factors relating to the trainee such as motivation, or factors stemming from the organizational context; see Cannon-Bowers et al. 1995 for more details), this can be considered a rather high value. *The most important training design dimensions predicting the reaction, learning, and attitude dimensions were demonstration, application, and integration.* Generally speaking, when these instructional design principles were implemented, more favourable reactions, a higher subjective knowledge gain, and more positive attitudes towards training were achieved. The only exception was the effect of application on perceived difficulty, with higher values in application corresponding to a subjectively more difficult training for participants. In terms of the remaining two design dimensions, problem-based learning had a positive effect on subjective knowledge gain and perceived usefulness, and activation reduced perceived difficulty of training. All models showed medium to large effect sizes (Cohen 1992). Overall, these results confirm hypothesis 4. The higher the ratings for the training design were, the more positive were the training outcomes, although application apparently made training subjectively more difficult.

Post-hoc Analyses

To broaden the scope of hypothesis 4, we explored the correlations between training design dimensions and external training outcome measures in a post-hoc analysis. This procedure lowers the risk of confirming hypothesis on the basis of common method variance, as both the training design and the training outcome dimensions were measured in a single questionnaire (Podsakoff et al. 2003).

The correlations between training design and external training outcomes and their significance can be found in Table 5. To summarise the results for the flight attendant sample TrS1, the training design dimensions mainly had an impact on knowledge after 8 weeks, but in the case of application also on knowledge directly after training. Participants who perceived the training to be well designed were thus able to memorise and remember more of the discussed topics. Furthermore, the possibility to integrate what was learned into their own knowledge was positively related to more successful behavioural intentions. Integration is thus the only dimension in the TEI that showed an impact on behavioural intentions in the flight attendant sample.

Summarising the results for the fire fighter sample TrS2, three of the four training design dimensions showed a positive correlation with knowledge, mirroring the results of sample TrS1. Moreover, a positive impact of training design on attitudes, especially on attitudes towards human fallibility, could be observed. Overall, we thus

found support for the notion that a well-planned training design with a focus on instructional principles has a positive effect on training outcomes, also when external outcome measures were used. The effect mainly emerged for the knowledge measures, and to a lesser extent for attitudes and behaviour.

Discussion

As stated in the introduction, the objective of this article was to introduce an approach to training evaluation that is theoretically and empirically founded, but at the same time meaningful to training decision makers and applicable within typical organisational constraints that do not permit to use more specific evaluation measures. Furthermore, with the inclusion of training design dimensions, the TEI supports formative evaluation and the exploration of design-related questions.

The results showed that the reliabilities of the training outcome dimensions (subjective enjoyment, perceived usefulness, perceived difficulty, subjective knowledge gain, and attitude towards training) and the training design dimensions (problem-based learning, activation, demonstration, application, and integration) of the TEI were satisfactory, with all values deemed acceptable for low-stakes instruments designed for programmatic decision-making with only minor or indirect consequences for individual examinees (Wasserman and Bracken 2003, p. 55). It can be argued that once a test meets a reliability criterion set for its area of application, the benefits of further increasing reliability are limited because of the risk of producing a narrow scale with compromised validity (Clark and Watson 1995; cited after Wasserman and Bracken 2003, p. 55). The items of the TEI were also subjected to two confirmatory factor analyses. The CFA supported the theoretically developed structure of the two parts of the inventory measuring training outcomes (five factors) and training design (five factors).

The first hypothesis that was tested concerned the three reaction dimensions included in the TEI and their relationship with other measures of training outcomes. As hypothesised, they showed a range of significant correlations with knowledge of training content and attitudes towards teamwork skills. Our results are thus in line with studies and meta-analyses showing substantial relationships between reactions and other training outcomes (Alliger et al. 1997; Sitzmann et al. 2008; Warr et al. 1999). Although reactions have been described as neither necessary (Kauffeld 2010) nor sufficient (Tannenbaum and Yukl 1992) for subsequent changes in attitudes, knowledge, or behaviour, they do possess informational value, based on our results and other published work in this area. Furthermore, the aspect of reactions ensures face validity from the perspective of training practitioners. Despite the encouraging results, the hypothesis could only be partially confirmed because the correlations in the data differed between the two samples used to test our assumptions. One reason for this might be that differing external measures of knowledge and attitudes were used. Future research to meet this limitation should ideally use the same instruments along with the TEI in different samples and across different types of training. However, the latter is difficult to realise, as the outcome measures have to be adapted to the content of training. An innovative method to overcome this problem could be to use meta-analytic techniques (see e.g. Cooper and Hedges 2009) to aggregate the

results from different samples and to seek generalisation regarding the relationship of the TEI with other training outcome measures. Moreover, possible moderator variables such as training motivation (Colquitt et al. 2000) or learning styles (Gully and Chen 2010) that are known to influence training outcomes should be assessed.

The second hypothesis stated that the TEI dimension subjective knowledge gain would be significantly related to objective learning measures. This was confirmed in both samples, although in the flight attendant sample, the correlation between subjective knowledge gain and knowledge of training content was only significant after a transfer period of 8 weeks. One possible explanation for this finding is that immediately after the training course all participants had a high level of knowledge, leading to a ceiling effect masking the true relationship. An indicator for the high discriminant validity of the scale is that it was not related to any other training outcome (attitudes or behavioural intentions). Our results thus show that subjective knowledge gain can be useful as a proxy for knowledge of training content when the use of knowledge tests is not possible or different courses with different knowledge content have to be compared. Our data are in line with research in the area of instructional communication, where the use of self-assessments of knowledge has had a long tradition (Chesebro and McCroskey 2000), and with meta-analytic results showing moderate correlations of self-reports of knowledge and cognitive learning (Sitzmann et al. 2010).

In the third hypothesis, we assumed that the general attitude towards training as measured in the TEI would be significantly correlated with the more specific attitudes towards the training content. We indeed observed significant relationships, and even though in the fire fighter sample, this was only true for one of four specific attitude scales, two of the other scales showed medium effect size correlations as well. The discriminant validity of the attitude dimension was lower than the discriminant validity of the subjective knowledge gain dimension; we observed significant correlations of attitude towards training and knowledge of training content in both samples. This result might be due to the fact that attitudes towards training are formed in an evaluative process (Olson and Zanna 1993), which is positively influenced by high knowledge gain (the more participants learn, the more favourable their attitude). More research regarding the inclusion of attitude towards training in evaluation is needed, but our results are promising and show that attitudes as part of the learning level of evaluation can be included in a generic training evaluation instrument.

Our fourth hypothesis stated that training design following sound instructional principles is an antecedent of positive training outcomes, and was confirmed by hierarchical regression analyses. Around 25 % of variance in the training outcome dimensions could be explained by training design, with demonstration, application, and integration being the most important design dimensions. The final models all showed medium to large effects (Cohen 1992). Our results are thus in line with previous studies showing a relationship between training design and the outcomes of instruction (Frick et al. 2009). Furthermore, significant relationships of training design dimensions, knowledge of training content, and specific attitudes towards teamwork could be observed in a correlational post-hoc analysis. We regard these results as an important indication that the outcomes of the regression analyses were not generated by common method variance (Podsakoff et al. 2003). However, it

would be valuable in future studies to run similar hierarchical regression analyses with training outcome data that was not collected using the TEI. To sum up the results regarding the hypotheses, they were all completely or partially confirmed and support the construct validity of the TEI.

Concerning the generic, context-independent nature of the TEI, it can be argued that this approach risks being superficial because training outcome measures should match what is being learned to assure the relevance of results (Goldstein and Ford 2002; Kraiger et al. 1993). We agree with this notion. However, when choosing a methodology for training evaluation, available resources, the intended purpose of the evaluation, and the needs of the intended audience have to be considered (Aguinis and Kraiger 2009). The TEI is not suitable to answer questions such as whether or not trainees are able to apply a learned technique in practice, to name an example. However, the TEI can be used to answer questions such as whether or not trainees found training useful or have gained knowledge through training. Additionally, training design aspects can be appraised. Thus, the TEI should not replace more outcome-specific, tailored evaluation measures, but be a reliable and valid complementary tool for contexts with restraints that make a generic questionnaire the most feasible option.

Implications for Further Research

First, using the TEI and other paper-and-pencil outcome measures, the problem of common method variance has to be considered. To mitigate consistency effects as a source of common method variance, the TEI contains a number of items high enough to prevent participants from deliberately "tuning" their answers to be consistent (Podsakoff et al. 2003). Additionally, we used external training criteria to test hypotheses 1 to 3. These criteria had a different format and were presented as separate questionnaires or tests to ensure that they were perceived as distinct from the TEI (Podsakoff et al. 2003). We can thus assume that the influence of common method variance was minimised.

Second, no higher-order evaluation with regard to feasibility of the instrument was done with practitioners, although organizational stakeholders were asked for feedback regarding the applicability of the questionnaire. Therefore, further research is needed to investigate the subjective acceptance of the questionnaire by trainees and to identify the possible application of the questionnaire in all kinds of seminars and trainings for trainees with various cognitive qualifications.

Third, the data analysed in this paper stems from different team training contexts. It would be beneficial to conduct further studies using the TEI in other training areas. Currently, our research group applies the TEI in a range of contexts, for example in the evaluation of blended learning or of undergraduate statistical courses. This data can serve to further validate the TEI in the future.

Fourth, the TEI does not explicitly include the transfer motivation of trainees, although two items within the dimensions of attitude towards training and integration consider aspects of the intention to transfer the training content to the workplace (att1, int5, see Appendix). Previous work has shown the influence of transfer motivation as a moderator variable on training outcomes such as work effectiveness or knowledge retention (Gegenfurtner et al. 2009; Gegenfurtner 2011). As a result of this research, a

context-independent measure of motivation to transfer including three scales covering autonomous (internalised) motivation to transfer, controlled (external) motivation to transfer, and intention to transfer (willingness to engage in transfer actions) has recently been proposed (Gegenfurtner 2012). These scales could be used as a valuable addition to the TEI to gauge whether transfer issues have to be addressed in training to support the later application of knowledge in practice (zu Knyphausen-Aufsess et al. 2009).

Finally, we held specific assumptions concerning the relationships between training outcomes (reactions, subjective knowledge gain, and attitude towards training) and external training outcome measures, which led to three hypotheses. On the other hand, we formulated only one broad hypothesis regarding training design, namely that the training design dimensions would predict training outcomes. In the hierarchical regression analyses, all design dimensions were entered at the same time. This rather exploratory approach was taken because so far, training research does not allow specific assumptions to be derived in this respect either from a theoretical or an empirical point of view. Although a number of general models concerning training process variables and training outcomes exist, such as the model by Cannon-Bowers et al. (1995) or the meta-analysis by Colquitt et al. (2000), they unfortunately only show the differentiated interplay of several person-related and design variables and their effect on training outcomes on a very broad level. It would be more elegant for future research to make assumptions in advance about the specific contribution of each training design dimension. Approaches could be to a) propose differential contributions of the training design dimensions in different organisational training contexts and regarding different training objectives and transfer requirements, or to b) experimentally vary the relative weight of each training design dimension and to investigate their impact with regard to a specific training objective.

Practical Implications

As stated above, the TEI is scientifically sound, but also highly feasible for applications in an organisational context to evaluate a wide range of training programmes under common organisational constraints. Researchers concerned with the development and effectiveness of training can gather data in a structured way, while practitioners can use the TEI to make training-related decisions, for example on the modification or (dis-) continuation of training programmes. In the study by Ritzmann and colleagues (Ritzmann et al. 2009; Ritzmann 2012) resulting in survey sample 1, TEI results were used to identify deficits in an existing training programme and to plan appropriate modifications. Likewise, a vocational training for improving safety behaviour in a large German steelworks was abandoned due to the evaluation results gained with the TEI. External training providers offering training services to companies could use the TEI to formatively check the quality of their product and whether it is perceived as intended.

To conclude, the TEI is based on past findings from training evaluation research and can be considered as theoretically and empirically founded. At the same time, the inventory offers a practical and systematic structure and is generic and independent of training content.

Acknowledgments The work presented in this article was partly funded by the Swiss Federal Commission for Technology and Innovation, Switzerland (project no. 9140.1 PFES-ES).

Appendix: Scales and Items of the TEI

This appendix presents the items and scales of the TEI in German and English. Square brackets indicate areas where the items have to be adapted to the specific course type (eg. *training, seminar, workshop,...)* and its content.

Training Outcome Dimensions

Subjective enjoyment

Item	German	English
enjoy 1	Das [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] hat mir generell gefallen.	Overall, I liked the [*DESCRIPTION OF COURSE TYPE—e.g. training*].
enjoy 2	Die Lernatmosphäre war angenehm.	The learning atmosphere was agreeable.
enjoy 3	Das Lernen hat Spass gemacht.	The learning was fun.

Perceived usefulness

Item	German	English
useful 1	Ich finde das [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] nützlich für meinen Beruf.	I find the [*DESCRIPTION OF COURSE TYPE—e.g. training*] useful for my job.
useful 2	Zeit in dieses [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] zu investieren war sinnvoll.	Investing time in this [*DESCRIPTION OF COURSE TYPE—e.g. training*] was useful.
useful 3	Ich kann die Inhalte des [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] in meinem Beruf anwenden.	I can apply the content of this [*DESCRIPTION OF COURSE TYPE—e.g. training*] in my job.
useful 4	Ich ziehe persönlichen Nutzen aus diesem [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*].	I derive personal use from this [*DESCRIPTION OF COURSE TYPE—e.g. training*].

Perceived difficulty

Item	German	English
diff 1	Die Inhalte waren verständlich.	The contents were comprehensible.
diff 2	Die Sprache (Fremd- und Fachwörter) war verständlich.	The language (foreign words and technical terms) was comprehensible.
diff 3	Ich bin thematisch im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] mitgekommen.	I kept up thematically in [*DESCRIPTION OF COURSE TYPE—e.g. training*].
diff 4	Die Zeit war ausreichend für die bearbeiteten Themen.	The time was sufficient for the themes covered.

Subjective knowledge gain

Item	German	English
knowl 1	Ich habe den Eindruck, mein Wissen hat sich langfristig erweitert.	I have the impression that my knowledge has expanded on a long-term basis.
knowl 2	Ich werde mir die neuen Themen gut merken können.	I will be able to remember the new themes well.
knowl 3	Ich denke, ich werde auch einige Zeit nach dem [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] noch berichten können, was ich gelernt habe.	I think that I will still be able to report what I learned some time after the [*DESCRIPTION OF COURSE TYPE—e.g. training*].

Attitude towards training

Item	German	English
att 1	Ich werde das Gelernte im beruflichen Alltag anwenden.	I will apply what I learned to my day-to-day work.
att 2	Ich finde es gut, dass [*THEMA/INHALT—z.B. Teamarbeit*] vermittelt bzw. besprochen wurden.	I find it good that [*THEME/CONTENT e.g. teamwork*] were imparted and/or discussed.
att 3	Ich würde dieses [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] meinen Kollegen empfehlen.	I would recommend this [*DESCRIPTION OF COURSE TYPE—e.g. training*] to my colleagues.

Training Design Dimensions

Problem-based learning

Item	German	English
prob 1	Es wurden zuerst Probleme thematisiert und durch ihre Bearbeitung habe ich somit die Themen gelernt.	First of all, problems were addressed, and by working on them I consequently learned the themes.
prob 2	Ich konnte echte Probleme bearbeiten und darin die zuvor gelernten Inhalte vertiefen.	I was able to work on real problems and therein consolidate the previously learned contents.
prob 3	Es wurden wahre Vorfälle aus [*ARBEITSFELD—z.B. Fliegerei*] vorgestellt und durch das selbständige Entwickeln von Lösungen habe ich die Themen vertieft.	Real incidents from [*FIELD OF WORK—e.g. aviation*] were presented, and through the independent development of solutions, I consolidated the themes.
prob 4	Es wurden Situationen aus dem Arbeitsalltag vorgestellt und ich musste herausfinden, was das mit [*THEMA—z.B. Teamarbeitsfähigkeiten*] zu tun hat.	Situations from day-to-day work were presented and I had to find out how they were linked to [*THEME—e.g. teamwork abilities*].
prob 5	Es wurden problematische Situationen aus dem Arbeitsalltag vorgestellt und ich musste herausfinden, mit welchen Fähigkeiten/welchem Wissen die Situation hätte verbessert werden können.	Problematic situations from day-to-day work were presented and I had to find out through which abilities/knowledge the situation could have been improved.

prob 6	Im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] wurden Situationen vorgestellt, die wir dann auf [*THEMENASPEKTE—z.B. Teamarbeits-Aspekte*] hin genauer betrachteten.	In the [*DESCRIPTION OF COURSE TYPE—e.g. training*], situations were presented which we then considered in more detail in terms of [*THEMATIC ASPECTS—e.g. aspects of teamwork*].

Activation

Item	German	English
act 1	Ich konnte mein eigenes Wissen in das [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] mit einbringen.	I was able to bring my own knowledge into the [*DESCRIPTION OF COURSE TYPE—e.g. training*].
act 2	Auf meine bisherigen Erfahrungen zu den behandelten Themen ist eingegangen worden.	My previous experiences regarding the themes covered were addressed.
act 3	Ich konnte mein berufliches Wissen zu [*THEMA—z.B. Teamarbeitsfähigkeiten*] in das Seminar/Training mit einbringen.	I was able to bring my own professional knowledge on [*THEME—e.g. teamwork abilities*] into the seminar/training.
act 4	Ich konnte meine bisherigen beruflichen Erlebnisse einbringen.	I was able to bring in my previous professional experiences.
act 5	Der Trainer hat mich aufgefordert mein Wissen und meine Erlebnisse zu [*THEMA—z.B. Teamarbeitsfähigkeiten*] aus dem Berufsalltag einzubringen.	The trainer invited me to bring in my own knowledge and experiences from my day-to-day work regarding [*THEME—e.g. teamwork abilities*].

Demonstration

Item	German	English
dem 1	Inhalte wurden mit konkreten Beispielen erläutert.	Contents were illustrated with concrete examples.
dem 2	Die Lernziele waren mir bekannt.	I was aware of the learning objectives.
dem 3	Die Lernziele wurden erreicht.	The learning objectives were achieved.
dem 4	Der Trainer machte deutlich, welches die zentralen Punkte der besprochenen Themen waren.	The trainer made it clear what the central points of the discussed themes were.
dem 5	Am Anfang des [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] wurden die Ziele bekannt gegeben, die mit dem [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] erreicht werden sollten.	At the beginning of the [*DESCRIPTION OF COURSE TYPE—e.g. training*], the objectives which were to be achieved through the [*DESCRIPTION OF COURSE TYPE—e.g. training*] were announced.
dem 6	Die eingesetzten Medien (PPT, Video, Poster, etc.) waren hilfreich für mein Verständnis.	The media employed (PPT, video, posters etc.) were helpful for my understanding.
dem 7	Die eingesetzten Medien (PPT, Video, Poster, etc.) waren geeignet die Inhalte zu präsentieren.	The media employed (PPT, video, posters etc.) were suitable for presenting the contents.

Application

Item	German	English
app 1	Ich konnte das Gelernte im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] üben.	I was able to practise what I had learned in [*DESCRIPTION OF COURSE TYPE—e.g. training*].
app 2	Ich habe im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] Feedback zu meinem Verhalten/meiner Leistung bekommen.	In the [*DESCRIPTION OF COURSE TYPE—e.g. training*] I received feedback on my behavior/my performance.
app 3	Ich konnte das Feedback umsetzen und im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] an meinem Verhalten/meiner Leistung arbeiten.	I was able to implement the feedback and work on my behavior/my performance in the [*DESCRIPTION OF COURSETYPE—e.g. training*].
app 4	Im [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] hatte ich die Möglichkeit, Dinge, die ich später in der Arbeit umsetzen soll, schon einmal auszuprobieren.	In the [*DESCRIPTION OF COURSE TYPE—e.g. training*], I had the opportunity to try out things which I should later implement in my work.
app 5	Das Feedback aus dem [*BEZEICHNUNG AUSBILDUNGSGEFÄSS—z.B. Training*] hilft mir, weiter am Gelernten zu arbeiten.	The feedback from the [*DESCRIPTION OF COURSE TYPE—e.g. training*] helps me to work further on what I learned.

Integration

Item	German	English
int 1	Inhalte wurden in Diskussionen vertieft.	Contents were consolidated in discussions.
int 2	Ich hatte Gelegenheit das Gelernte zu reflektieren.	I had the opportunity to reflect on what I had learned.
int 3	In der Diskussion habe ich erfahren, dass Kollegen andere Sichtweisen zu dem behandelten Thema haben.	In the discussion, I discovered that colleagues have different views on the theme covered.
int 4	Ich kenne die Wichtigkeit der einzelnen Themen für unterschiedliche Situationen.	I know the importance of the individual themes for different situations.
int 5	Mir ist klar geworden wie ich die behandelten Inhalte im Arbeitsalltag anwenden kann	It became clear to me how I can apply the themes covered in day-to-day work.

References

Aguinis, H., & Kraiger, K. (2009). Benefits of training and development for individuals and teams, organizations, and society. *Annual Review of Psychology, 60*, 451–474.

Alliger, G. M., & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: thirty years later. *Personnel Psychology, 42*, 331–342.

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

Arthur, W., Bennett, W., Stanush, P. L., & McNelly, T. L. (1998). Factors that influence skill decay and retention: a quantitative review and analysis. *Human Performance, 11*(1), 57–101.

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

Bassi, L. J., Benson, G., & Cheney, S. (1996). The top ten trends. *Training and Development, 50*(11), 28–42.

Billett, S. (2008). Welcome to the new journal. *Vocations and Learning, 1*, 1–5.

Blanchard, P. N., Thacker, J. W., & Way, S. A. (2002). Training evaluation: perspectives and evidence from Canada. *International Journal of Training and Development, 4*(4), 295–304.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: a meta-analytic review. *Journal of Management, 36*(4), 1065–1105.

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of Applied Psychology, 71*(2), 232–245.

Campbell, D. J., & Lee, C. (1988). Self-Appraisal in performance evaluation: development versus evaluation. *The Academy of Management Review, 13*(2), 302–314.

Cannon-Bowers, J. A., Salas, E., Tannenbaum, S. I., & Mathieu, J. E. (1995). Toward theoretically based principles of training effectiveness: a model and initial empirical investigation. *Military Psychology, 7*(3), 141–164.

Chesebro, J. L., & McCroskey, J. C. (2000). The relationship between students' reports of learning and their actual recall of lecture material: a validity test. *Communication Education, 49*, 297–301.

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

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: a meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*(5), 678–707.

Cooper, H., & Hedges, L. V. (2009). Research synthesis as a scientific process. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 3–16). New York: Russell Sage.

European Centre for the Development of Vocational Training. (2010). *Employer-provided vocational training in Europe. Evaluation and interpretation of the third continuing vocational training survey.* Luxembourg: Publications Office of the European Union.

Field, A. (2005). *Discovering statistics using SPSS*. London: Sage Publications Ltd.

Frick, T., Chadha, R., Watson, C., Wang, Y., & Green, P. (2009). College student perceptions of teaching and learning quality. *Educational Technology Research and Development, 57*, 705–720.

Gagné, R. M., Wagner, W. W., Golas, K. C., & Keller, J. M. (2005). *Principles of instructional design* (5th ed.). Belmont: Wadsworth.

Gegenfurtner, A. (2011). Motivation and transfer in professional training: a meta-analysis of the moderating effects of knowledge type, instruction, and assessment conditions. *Educational Research Review, 6*, 153–168.

Gegenfurtner, A. (2012). Dimensions of motivation to transfer: a longitudinal analysis of their influence on retention, transfer, and attitude change. *Vocations and Learning*. doi:10.1007/s12186-012-9084-y.

Gegenfurtner, A., Veermans, K., Festner, D., & Gruber, H. (2009). Motivation to transfer training: an integrative literature review. *Human Resource Development Review, 8*, 403–423.

Goldstein, I., & Ford, J. K. (2002). *Training in organizations* (4th ed.). Belmont: Wadsworth.

Gully, S., & Chen, G. (2010). Individual differences, attribute-treatment interactions, and training outcomes. In S. W. Kozlowsky & E. Salas (Eds.), *Learning, training, and development in organizations* (pp. 3–64). New York: Routledge.

Hagemann, V. (2011). *Trainingsentwicklung für High Responsibility Teams [Training development for High Responsibility Teams]*. Lengerich: Pabst Science Publishers.

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

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis* (6th ed.). Upper Saddle River: Prentice Hall.

Helmreich, R. L., Merritt, A. C., & Wilhelm, J. A. (1999). The evolution of crew resource management training in commercial aviation. *The International Journal of Aviation Psychology, 9*(1), 19–32.

Hochholdinger, S., Rowold, J., & Schaper, N. (2008). *Ansätze zur Trainings- und Transferevaluation [Approaches to the evaluation of training and transfer]. In Evaluation und Transfersicherung betrieblicher Trainings [Evaluation and assurance of transfer of vocational training]* (pp. 30–53). Göttingen: Hogrefe.

Holton, E. F. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly, 7*, 5–21.

Hu, & Bentler. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55.

Kauffeld, S. (2010). *Nachhaltige Weiterbildung [Sustainable further education]*. Berlin: Springer Verlag.

Keith, N., & Frese, M. (2008). Effectiveness of error management training: a meta-analysis. *Journal of Applied Psychology, 93*(1), 59–69.

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

Kraiger, K. (2002). Decision-based evaluation. In K. Kraiger (Ed.), *Creating, implementing, and managing effective training and development* (pp. 331–375). San Francisco: Jossey-Bass.

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

Lewis, J. L. (2009). Student attitudes toward impairment and accessibility: an evaluation of awareness training for urban planning students. *Vocations and Learning, 2*, 109–125.

McDonnald, L. M., & Shadow, L. W. (2003). *Precursor for error: An analysis of wildland fire crew leaders' attitudes about organizational culture and safety.* Presentation to the 3rd International Wildland Fire 2003 Conference, Australasian Fire Authorities Council, Sydney, Australia.

Merrill, M. D. (2002). First principles of instruction. *Educational Technology Research and Development, 50*(3), 43–59.

Nunnally, J. C. (1978). *Psychometric theory.* New York: McGraw-Hill.

Olson, J. M., & Zanna, M. P. (1993). Attitudes and attitude change. *Annual Review of Psychology, 44*, 117–154.

Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: a critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*(5), 879–903.

Ritzmann, S. (2012). *Entwicklung und Evaluation von Crew Resource Management Training für Flight Attendants [Development and evaluation of crew resource management training for flight attendants]*. Lengerich: Pabst Science Publishers.

Ritzmann, S., Kluge, A., & Hagemann, V. (2009). Crew Resource Management für Kabinenbesatzungen: Ein konzeptbasierter Ansatz. [Crew Resource Management for cabin crews: A conceptual approach.]. In M. Grandt & A. Bauch (Eds.), *Kooperative Arbeitsprozesse [Cooperative work processes] (DGLR-Report 2009–02)* (pp. 261–277). Bonn: Deutsche Gesellschaft für Luft- und Raumfahrt e.V.

Ritzmann, S., Kluge, A., Hagemann, V., & Tanner, M. (2011). Integrating safety and crew resource management (CRM) aspects in the recurrent training of cabin crew members. *Aviation Psychology and Applied Human Factors, 1*(1), 45–51.

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

Salas, E., Prince, C., Bowers, C. A., Stout, R. J., Oser, R. L., & Cannon-Bowers, J. A. (1999). A methodology for enhancing crew resource management training. *Human Factors, 41*, 161–172.

Salas, E., Wilson, K. A., Burke, C. S., & Wightman, D. (2006a). Does crew resource management work? An update, an extension, and some critical needs. *Human Factors, 48*(2), 392–412.

Salas, E., Wilson, K. A., Priest, H. A., & Guthrie, J. W. (2006b). Design, delivery, and evaluation of training systems. In G. Salvendy (Ed.), *Handbook of human factors and ergonomics* (pp. 472–512). Hoboken: Wiley.

Sexton, J. B., Helmreich, R. L., Glenn, D., Wilhelm, J. A., & Merritt, A. C. (2000). Operating Room Management Attitudes Questionnaire (ORMAQ) *The University of Texas at Austin, Human Factors Research Project.*

Shelton, A., & Alliger, G. M. (1993). Who's afraid of level 4 evaluation? A practical approach. *Training and Development, 47*, 43–46.

Sitzmann, T., Brown, K. G., Casper, W. J., Ely, K., & Zimmerman, R. D. (2008). A review and meta-analysis of the nomological network of trainee reactions. *Journal of Applied Psychology, 93*(2), 280–295.

Sitzmann, T., Ely, K., Brown, K. G., & Bauer, K. N. (2010). Self-assessment of knowledge: a cognitive learning or affective measure. *Academy of Management Learning & Education, 9*, 169–191.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using multivariate statistics*. Boston: Pearson Education.

Tannenbaum, S. I., & Woods, S. B. (1992). Determining a strategy for evaluating training: operating within organizational constraints. *Human Resource Planning, 15*, 63–81.

Tannenbaum, S. I., & Yukl, G. (1992). Training and development in work organizations. *Annual Review of Psychology, 43*(1), 399–441.

Taylor, P. J., Russ-Eft, D. F., & Chang, D. W. (2005). A meta-analytic review of behavior modeling training. *Journal of Applied Psychology, 90*(4), 692–709.

Twitchell, S., Holton, E. F., & Trott, J. W. (2000). Technical training evaluation practices in the United States. *Performance Improvement Quarterly, 13*, 84–109.

Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.

Warr, P., & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology, 48*(2), 347–375.

Warr, P., Allan, C., & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*, 351–375.

Wasserman, J. D., & Bracken, B. A. (2003). Psychometric characteristics of assessment procedures. In J. R. Graham & J. A. Naglieri (Eds.), *Handbook of psychology: Assessment psychology (Volume 10)* (pp. 43–66). Hoboken: Wiley.

Weick, K. E., & Sutcliffe, K. M. (2007). *Managing the unexpected* (2nd ed.). San Francisco: Jossey-Bass.

Yule, S., Flin, R., Paterson-Brown, S., & Maran, N. (2004). Surgeons' attitudes to teamwork and safety. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting* (pp. 2045–2049).

zu Knyphausen-Aufsess, D., Smukalla, M., & Abt, M. (2009). Towards a new training transfer portfolio: a review of training-related studies in the last decade. *Zeitschrift für Personalforschung [German Journal of Research in Human Resource Management], 23*, 288–311.

Sandrina Ritzmann is a research associate at the University of Applied Sciences Northwestern Switzerland (FHNW) and at the Center for Adaptive Security Research and Applications (CASRA), Switzerland. Her research areas are human factors in aviation security and safety with a focus on training, competency assessment, and teamwork. She obtained her doctorate in psychology (crew resource management training for flight attendants) at the University of Duisburg-Essen, Germany, in 2012.

Vera Hagemann is a research associate within the field of human factors, crew resource management, teamwork, and training design as well as consumer behaviour at the University of Duisburg-Essen, Germany. She obtained her doctorate in psychology (training development for high responsibility teams) at the University of Duisburg-Essen, Germany, in 2011.

Annette Kluge is a professor for business and organisational psychology at the University of Duisburg-Essen, Germany. Her research areas are skill acquisition and maintenance in complex environments and violations within organizational contexts. She obtained her doctorate in ergonomics and vocational training at the University of Kassel, Germany, in 1994.