Chapter 25: Conducting organizational based evaluations of coaching and mentoring programs

Siegfried Greif

Introduction

One of the basic questions of coaching and mentoring programs in organizations is how their effects can be evaluated. Without answering this question, those people in organizations who decide on the implementation or continuation of these programs will not be able to find out whether they are worth the investment of time and money. The primary aim of this chapter is to inform the reader about evaluation models and methods that meet high standards of quality and can be recommended for use in program evaluation studies.

What kinds of data can be used for evaluations that are to provide evidence for the effectiveness of Human Resources interventions in organizations? This question has prompted long-standing debates and research in the field of training. For coaching and mentoring, however, this is a relatively new question. Nevertheless, we can draw on solutions from training evaluation models to aid our understanding of what might work well in studies on coaching and mentoring.

The chapter starts with an overview of general program evaluation taxonomies. After this, specific evaluation models of coaching and mentoring and basic literature on quantitative measures will be presented. Qualitative evaluation methods will then be introduced that can also be applied and that are often seen as an alternative approach to quantitative methods. Following on from this, the challenging question of future evaluation research will be discussed, and finally, the last section will summarize the conclusions.

Standard evaluation concepts and taxonomies

The following section describes basic characteristics of “evaluation”, on which there is much consensus between practitioners and scientists. It shows how evaluations are used in organizational decisions on the implementation of coaching or mentoring programs. The last part of the section provides an introduction to the classical taxonomy of data by Kirkpatrick (1976), which has been applied widely in the evaluation of HR-interventions.

Definition and standards of evaluation

In their classic book, Rossi and Freeman (1993, p. 5) define evaluation research as the “systematic application of social research procedures in assessing the conceptualization and design, implementation, and utility of social intervention programs”. However, evaluation of interventions like coaching and mentoring is not restricted to scientific research methods. The general meaning of the term “evaluation” is the appraisal or assessment of interventions, including their preconditions, costs, processes, and especially their outcomes.

For many coaches' clients or mentors' protégés the coaching or mentoring experience is novel. A summary of surveys of coaching shows that they are usually satisfied or very satisfied with the intervention (Greif, 2008, p. 215) and sometimes they are even enthusiastic (Wasylyshyn, 2003). However, we also find skeptical practitioners who doubt the usefulness of mentoring or coaching programs (Bachmann & Spahn, 2004). “I don't need coaching” is a sentence that is often heard. It implies that those who require coaching are in some way disabled. Organizations contribute to
this kind of implicit discrimination when they coaching programs are started only with problematic cases ("coaching as the last chance before dismissal") and not with high potentials. Which employee when knowing this would voluntarily participate in a coaching program? To argue against coaching and mentoring is easy, since both are sensitive, complex and intangible services. If nobody except two people really know what has happened in the confidential coaching conversation, it is easy to spread negative rumors about it.

Coaching and mentoring, more so than other services, need rational evaluation in order to avoid prejudice and false information. Professional associations in both fields therefore strongly advocate scientific methods of evaluation and make use of evaluation studies for their marketing. In the discussion at the end of the chapter we will come back to examine the interesting relations between evaluation and marketing.

Personnel managers or other decision makers who select Human Resources (HR) interventions for their organization may evaluate their cost and benefit and ask for data on the economic return of the investment. Number-oriented deciders may ask for additional quantitative data, e.g. the satisfaction ratings of the coachees’ or mentees’, or improvement of leadership behavior assessed by leadership scales. However, in organizations we also find decision makers who are skeptical as to whether it is possible to evaluate all HR investments by means of economic data or quantitative measures. The majority of deciders probably follow the personal recommendations of their colleagues or of experts or counselors, based on practical experience or reports on the outcome of similar applications. The advantage of such qualitative information is their authenticity and credibility. The results of a recent survey (Stephan & Gross, 2011, p. 168) show that recommendations by word-of-mouth communication or opinions within the network are preferred by 58% of personnel managers when seeking a coach. Other types of information such as Internet searches (17%) or internal databases (5%) are used much less frequently.

The general purpose of evaluation that goes beyond word-of-mouth recommendations is to obtain an objective, reliable and valid information basis for decisions on investments. Scientific methods and evaluation studies that are performed and interpreted independently and without regard to preconceived opinions are often seen as the best information basis for making rational decisions. As Wottawa and Thierau (1998) emphasize, the practitioner or scientist who is conducting an evaluation study carries a high moral responsibility to follow the standards of quality of science in a transparent and checkable way. The standards apply when planning the design of the study and also to the selection of instruments. The realization, statistical data processing, interpretation and presentation of the results have to comply with these standards. The management is also meeting a clear corporate policy decision when it passes part of its influence over to an independent analysis. The results hereof are uncertain and may have important consequences.

Wottawa and Thierau (1998, p. 14) outline three basic characteristics of evaluations on which there seems to be much consensus:

1. “Evaluation” implies an appraisal of alternative measures and serves as a basis for planning and decision-making.
2. “Evaluation” is a goal and purpose oriented activity and aims primarily at a review or improvement of measures and decisions about measures.
3. “Evaluations” have to meet and be adapted to the current requirements of scientific techniques and methods.
Scriven (1980, 1996) has introduced the differentiation between *formative* and *summative evaluation*. Formative evaluation refers to the analysis of the preparation, implementation and processes of an evaluation object. Its purpose is often to shape and improve the quality of the measures, if possible in the course of the processes. Summative evaluation focuses on the results or outcomes of an intervention in relation to its goals. This chapter concentrates more on the latter, but as shown below, extended evaluation models in addition to outcome criteria also embrace the antecedents, processes and organizational context of coaching and mentoring.

**The classical taxonomy of Kirkpatrick**

Donald L. Kirkpatrick (1976) has published a taxonomy of different kinds of data that can be used in evaluating training programs in organizations as well as other HR-interventions like coaching and mentoring. It is so convincing that it is still regarded as a standard of quality. He differentiates between four levels of evaluation criteria or “segments” of the evaluation process:

1. *Reaction:* How do the participants feel about the program they attended? To what extent are they ‘satisfied customers’?
2. *Learning:* To what extent have the trainees learned the information and skills? To what extent have their attitudes been changed?
3. *Behavior:* To what extent has their job behavior changed as a result of attending the training program?
4. *Results:* To what extent have results been affected by the training program?” (Kirkpatrick, 1977, p. 9).

Examples of measures of reactions are questionnaires, e.g. on the satisfaction of the participants with the program or the trainer who ordered it. Kirkpatrick (1977) advocates anonymous surveys in order to get honest reactions. Such reaction evaluations are widely distributed in organizations. The results of such surveys are interesting, but normally the program goals go beyond making participants happy.

The relationship of data from the next level to program goals – learning – is normally more meaningful. In order to detect learning improvements, for example, the participants, following Kirkpatrick (1977), could be asked by interview or questionnaire what they have learned in the program. More objective evidence could be found by means of knowledge tests, administered before and after the intervention.

Many programs aim at changing the behavior of the participants. This refers to the third level of the taxonomy. Simple methods of getting evidence here are open questions, asking the participants to describe what they eventually did differently after the program, or asking their bosses. A better but much more complex method mentioned by Kirkpatrick (1977) is observing and measuring participants’ behavior changes before and after the intervention.

Examples of data from the fourth level, the results, (mentioned by Kirkpatrick (1977, p. 9) are profits, return on investment, sales, production quality, quantity, schedules being met, costs, turnover, grievances and morale. The examples show that this level refers to the more long-term organizational outcomes of the program. These data are often very important and relate directly to central organizational goals.

Kirkpatrick’s essential message is that it is not sufficient to use only subjective reaction data after the intervention as evidence for its effectiveness. He demands the utilization of data from multiple sources and encourages pragmatic assessment methods. As Kirkpatrick (1977) discusses in his paper, it is relatively easy to find “evi-
idence” of changes for each level by comparing pre- and post-assessments. However, it is difficult to obtain “proof”, that the changes have been caused by the program. Many other possible factors that could have influenced the changes have to be eliminated. As the best solution, he advocates comparison of the program participants with randomized control groups (see also Grant’s chapter in this book). However, he argues, especially with regard to the evaluation of results that it is often impossible or impractical to prove that the changes have been caused by the program alone. He describes a nice example in which he helped a friend to convince his boss that it is impossible to prove “in dollars and cents that a certain leadership training program was achieving more benefits than it was costing” and that “evidence” in the case was good enough (Kirkpatrick, 1977, p. 12).

Kirkpatrick’s four-level evaluation taxonomy is still an international standard today. It has inspired many followers to raise the standard of quality for program evaluations in organizations. However, aspects of the model require improvement or modification. Kraiger, Ford and Salas (1993) propose a modification of the learning level. They differ between (1) cognitive outcomes (verbal knowledge, knowledge organization and cognitive strategies, (2) skill based outcomes (skill compilation and automaticity) and (3) affective (attitudinal and motivational) outcomes. Similarly, it might be recommendable to systematize the remaining “levels” of Kirkpatrick’s taxonomy. The “results” level in particular seems to be very heterogeneous.

Holton (1996) criticized more fundamentally the fact that Kirkpatrick and his followers called the concept a “model”. As he demonstrates, the system does not meet any of the standard criteria for scientific models (e.g. definition of constructs, assumptions about their relations, propositions and hypotheses, and predictions). He suggests calling it a “taxonomy of outcomes”. In his reply, Kirkpatrick (1996) accepts that his system is not a model in the conventional meaning and does not mind calling it a taxonomy. Holton (1996) recommends developing a more comprehensive evaluation model in the narrower meaning of the term that would contain preceding, intervening and outcome variables together with assumptions about their relations. The open question is whether his or other evaluation models can be generalized for use with all kinds of specific HR-interventions. In the following, two evaluation models are presented which refer more specifically to coaching and mentoring.

Program evaluation models and measures of coaching and mentoring outcomes

The role relationship between mentor and protégé is very different from that between coach and individual coachee. The mentor is often a senior in the occupational field and has not received any specific training. The coach has normally received a professional education, but does not necessarily come from the occupational field of the coachee. Additionally, the expected outcomes of both interventions differ. Mentoring aims at career development and building social networks. Individual coaching is broader in its scope. It can support the coachee in a variety of goals, for example improvement of leadership behavior, effective self- and stress management, team communication, conflict resolution, better work-life balance and in some cases also career developmental goals, to mention just a few. The following evaluation models mirror these differences. However, there are also similarities. In both, intimate supportive relationships are important that are based on mutual trust. The outcomes of coaching and mentoring on the reaction-level of Kirkpatrick’s taxonomy can be measured by satisfaction questionnaires. On the learning-level, increases in knowledge on social interactions are expected in both fields. Desired outcomes on
the behavior-level are self-efficacy and occupational performance improvement, and on the result-level eventually an increase in organizational productivity.

Both of the sub-chapters in the following section will commence with a description of the evaluation models. Then a summary of criteria and exemplary evaluation studies will be presented. The presentation focuses on the outcome measures of organizational programs that were implemented in individual and not group coaching and mentoring. The evaluation models mainly refer to quantitative measures. As mentioned in the introduction, qualitative evaluation methods also provide important, sometimes very convincing evidence of coaching outcomes. Exemplary methods will be described later in another sub-chapter.

**How to evaluate coaching programs**

Since the turn of the century, the number of reviews on coaching evaluation research and recommendations for future studies has increased substantially (Bachkirova & Kauffman, 2008; Cornett & Knight, 2009; De Meuse, Dai, & Lee, 2009; Ellam-Dyson & Palmer, 2008; Feldman & Lankau, 2005; Fillery-Travis & Lane, 2008; Grant, 2006, 2009; Grant, Passmore, Cavanagh, & Parker, 2010; Greif, 2007, 2011a; Künzli, 2006, 2009; Latham, 2007; Levenson, 2009; Linley, 2006; Passmore & Fillery-Travis, 2011; Passmore & Gibbes, 2007; Smither, Reilly, & London, 2001). A common theme is a plea for more and better research on the effectiveness of coaching that would provide a valuable foundation for the profession. In their review of leadership coaching studies Ely et al. (2010; Ely & Zaccaro, 2011) refer to the four-level taxonomy by Kirkpatrick (1976) and the supplement by Kraiger et al. (1993) as a framework. In their conclusions for future program implementations and evaluation studies they demand that there be more focus on the antecedents’ influence on coaching outcomes. They especially mention the organizational stakeholders and actors involved, their relationship, and the process of coaching. As many of the reviewers of the list above also emphasize, the effectiveness of the program depends much on the quality of the antecedents and processes.

The coaching evaluation model presented in figure 1 gives a broad overview of antecedents, organizational context and process variables that are assumed to influence proximal and distal outcomes. It is an extension of an earlier version, which summarized the coaching outcome research with a particular focus on studies that applied randomized control groups (Greif, 2007), a recent coaching research review by Grant, Passmore, Cavanagh and Parker (2010), and scales mentioned by Ely and Zaccaro (2011). Inspired by the mentoring model by Wanberg et al. (2003) that will be presented later, it has been extended to embrace antecedent and the context variables as well as a differentiation between proximal or short-term and distal or long-term outcomes. The figure contains many variables. Not all of them will be described in detail. The presentation concentrates instead on measures and scales with previously tested reliability and validity, and which have proved their worth in several outcome studies.

- Figure 1 about here -

Important antecedents of successful coaching interventions are acceptance of the program by organizational stakeholders and participation of people with high potential and not just so-called “problem cases”. Positive transfer climate, for instance, is a relevant organizational context variable. Examples of process variables are mutual respect and trust in the coaching relationship or skilled behaviors of the coach like activation of the client’s resources. We start with a description of the model and rec-
ommended measurement scales and instruments that can be applied both in the organizational and scientific evaluation projects.

Companies evaluate the efficiency of their structures, processes and output by means of comparing multiple measures with the best practices of other companies. They do this by *benchmarking* and investing money and time in order to improve their performance in comparison with their competitors (Boxwell, 1994). In order to compare the improvement in investments and outcomes of coaching and mentoring programs, it would be necessary to use the same or at least similar standard benchmark criteria and measurement scales. This would also be of great advantage to research. Only if different research studies apply comparable scales will it be possible to derive general conclusions on the efficiency of coaching. We will therefore pay special attention in the following passages to criteria and scales that have approved reliability and validity, have been applied frequently in evaluation studies, and can be recommended as future standard benchmark measures both for research and practice.

**Proximal outcomes**

Typically, the short-term outcomes of coaching programs are evaluated by the *satisfaction ratings* of the coachees (Peterson & Kraiger, 2003). The use of such scales is not restricted to coaching evaluation. They are therefore classified in figure 1 as *general measures*. Five-point Likert scales (1 = not satisfied to 5 = very satisfied) are common. Satisfaction ratings are reaction-level data and subject to the so called “Hello-good-bye-effect”, a positive feeling of gratefulness or leniency or a way of saying thank you. Positive satisfaction ratings do not imply that the coaching has produced concrete results. They can be interpreted as a standard feedback-reaction from the customers. They are useful for marketing and to motivate other potential coachees (Ely & Zaccaro, 2011).

It is somewhat disappointing that in many organizations coaching programs are evaluated merely by means of simple ratings and seldom rigorously (Fillery-Travis & Lane, 2008). One example is the coaching implementation conducted in the world wide leading ERP-Software company SAP in Germany (Grafe & Kronig, 2011). The outcomes were evaluated by means of four satisfaction ratings (satisfaction with the coaching results, satisfaction with the coaching process, willingness to participate again in coaching and willingness to recommend the coaching). The company won the second place of the German Coach Federation's coaching award 2010 for being a best practice project, not because of their evaluation model or very good satisfaction ratings, but because the project is a model for high quality standards in the integration of the program into their personal development measures, as well as for the selection of qualified coaches according to high standards of the profession. It is an especially good model for successful bottom-up implementation strategies: the employees for example need not ask their supervisor for permission if they want coaching. A continuous rise in demand shows the acceptance of the strategy. The jury that decided on the award therefore evaluated the implementation by means of the program antecedents and the management of the process of implementation.

It does not make much sense for companies to construct homemade satisfaction scales for the outcome evaluation of HR-programs. *Benchmark-comparison* of coaching programs in different organizations or between different types of HR-interventions are only possible if the outcomes are measured by standard scales that have been applied in many different organizations and evaluation studies.

An example of a diffused and very simple rating of customer satisfaction that has proved practical is the *Net Promoter Score® (NPS)* (Reichheld, 2003). The coachee would have to answer only one, nevertheless very revealing, question: “How likely is
An open-ended question follows, where the person may elaborate on or explain their rating. Customers who give a rating of 9 and 10 are classified as “Promoters”, and those with ratings of 0 to 6 as “Detractors”. The promoter score is calculated by subtracting the percentage of the Detractors from that of the Promoters. It can vary between -100 and +100. The answers to the open questions can be used to improve the services. One of the advantages of the NPS is its simplicity. Nobody could argue that this scale generates too much work for the company. Due to the long eleven-point scale it is possible to identify enthusiastic customers: something that would be impossible with five-point scales.

A more sophisticated standard satisfaction measurement instrument, and one which has been applied in numerous production and service branches, is the American Customer Satisfaction Index (ACSI) (Fornell, Johnson, Anderson, Cha, & Bryant, 1996). It uses ten-point scales. The index is based on weightings of the importance of standard quality attributes of the service and comparison with competitors. It is not purely a low level measure of emotional reactions, since it also predicts customer behavior and correlates with economic return. Notably, up to the present day there have been no published studies that have applied either the NPS or ACSI in the field of coaching.

The second group of frequently used outcome criteria consists of subjective goal attainment or goal satisfaction ratings. Again, we often find simple five-point ratings or open responses with estimated percentages. Such ratings may also be classified as reaction level data if they are based merely on the spontaneous impressions and feelings of the coachee. However, if they are assessed by means of more refined methods they can reach the behavior level.

Spence (2007) recommends a refined method of Goal Attainment Scaling (GAS) to reduce subjective distortions or biases. He obtains a score by asking the respondents to rate their success for each goal on a five-point Likert scale (e.g. 1 = 0-20% successful, to 5 = 80-100% successful). A direct estimation of the percentages is also possible. To obtain a mean value, these ratings are summed up and divided by the total number of goals. In addition, he recommends assessing the ratings of difficulty of reaching the goal for each goal (e.g. 1 = very easy, to 4 = very difficult). The goal attainment score can now be calculated by the following formula (Spence, 2007):

\[
\text{Goal Attainment} = \frac{(\text{Goal 1 Difficulty} \times \text{Attainment}) + (\text{Goal 2 Difficulty} \times \text{Attainment}) + (\ldots)}{\text{Total Number of Goals}}
\]

In this formula the attainment of difficult goals is weighted more highly than that of easy goals. Spence recommends assessing such scores at different times (e.g. at least at the beginning and the end of the coaching). The scoring approach has been applied especially by the researchers of the Coaching Psychology Unit of the University, Australia.

Similarly, in the research at the University of Osnabrück, Germany, it is standard to assess ratings of goal attainment and difficulty (or probability of reaching the goals). Before the ratings, the coaches are asked to define their goals and to describe sub goals and possible steps by means of open-ended questions. In addition, Likert scales ratings are used that refer to the importance of the goals, goal attainment satisfaction, determination to reach the goals, disappointment if the goals cannot be reached, concreteness of the steps planned, persistence in pursuing the goals, favorable or unfavorable context conditions, and finally the chances of reaching the
goals. The differentiated ratings provide valuable information for the subsequent coaching. It is also taken into account that goal and other attributes may change. If the initial goals change during the intervention process or if the client realizes that it is impossible to reach the goal 100%, goal attainment satisfaction seems to be a more valid rating than the percentage of goal attainment.

Additional scales showing effects of coaching in several outcome studies using random or not-random control groups (Ely, et al., 2010; Grant, 2009; Grant, et al., 2010; Greif, 2007, 2011a) are: affect scales, for example PANAS (Watson, Clark, & Tellegen, 1988) or similar instruments, The Depression, Anxiety and Stress Scale (DASS; Lovibond & Lovibond, 1995), and the scales of Psychological Well-being (Ryff & Keyes, 1995) or Cognitive Hardiness (Nowack, 1990). The results of several studies support a reduction in negative affect, anxiety, stress and depression after coaching as well as an improvement in general well-being and health. Some studies found an increase in hardiness of self-esteem or hope. The standard scales that are used here are reliable and valid and have proven their worth in many different fields of application. If they are used in parallel in training programs that aim at similar goals, then benchmark-comparison between coaching and training interventions is possible. Normally we would expect that the outcomes are significantly better for coaching since it is more individualized and intensive. Contrary to expectation, the necessary expenditure of time and the investment costs of coaching can be lower than the costs of training. This applies if the expense of providing substitutes for the participants in the company's training programs are high, as has been demonstrated in the cost analysis of a case study (Greif & Scheidewig, 1998).

If we consider the hypothetical short-term general outcomes of the coaches, we would expect that satisfaction, general well-being and self-esteem would improve if the coaching were rated as a success by the coachees and organization. However, to date there are no published studies that assess proximal outcomes of this sort for the coach.

Figure 1 also presents outcomes that are specific to coaching (coachee, coach and organization). Goal oriented coaching, for example, is expected to result specifically in an improvement of goal clarity or concreteness that can be assessed by goal attainment scaling. In contrast to such elementary ratings, Runde (2004) has validated a holistic questionnaire instrument (S-C-Eval, German version only). Confirmatory factor analysis confirmed three basic quality dimensions of coaching: Structure-, Process- and Result-Quality by means of retrospective evaluation of the coachees. The short but reliable scales have so far been used for coaching program evaluation in the German police force (Runde, Bastians, & Weiss, 2005), individual leadership coaching (Sass, 2006), a randomly controlled coaching program with business administration and law students (Schmidt & Thamm, 2008), and in an adapted version also by a mentoring program (Klien, 2011). It would be recommendable to translate and test it in other countries.

A questionnaire scale that seems particularly suitable for coaching evaluation is the Insight Scale by Grant, Franklin and Langford (2002). It assesses the perceived clarity of the participant's own feelings. The items of the scale are similar to the items of the Emotional Clarity subscale of the Trait Meta-Mood Scale (TMMS, Mayer, Salovey, & Caruso, 2004; Salovey, Mayer, Goldman, Turvey, & Palfai, 1995). Emotional clarity is an important component of Emotional Intelligence. According to several studies emotional insight improves after coaching (Greif & Berg, 2011).

According to Grant (2006, pp. 153 f.), coaching not only means facilitating the coachees in setting goals, developing plans of action, and beginning to act. It also
implies strengthening their self-monitoring and self-awareness of their performance when evaluating the results. A synonym term that is often used in the field is self-reflection. Self-reflection is seen as a basis for modifying coachees’ actions in order to better reach their goals. Grant refers to the self-regulation theory of Carver and Scheier (1998), who regard cycles of such processes as the core of all goal-directed actions. Ely and Zaccaro (2011) mention self-awareness as an important outcome of coaching, but demand more rigorous measurements. There have been several approaches to developing scales for assessing self-awareness as a positive outcome of coaching. However, an increase is not always positive. Self-awareness or self-reflection may be confounded with rumination and correlate with depression (Greif & Berg, 2011). Therefore, in our integrative process theory of coaching (Greif, 2008, 2010), we differentiate between circular rumination and result-oriented problem- and self-reflection. Further, we assume that in the majority of cases coaching stimulates merely those reflections that are salient to the specific coaching intervention for a short period of time. A coaching session that focuses on goal clarification, for example, will intensify goal reflections. In contrast, a self-management coaching will result in more reflections on self-management. We have constructed validated specific scales that assess result-oriented problem- and self-reflection of goals and self-management and are not biased by rumination. Several randomly controlled studies have shown the expected specific short-term improvements (Greif & Berg, 2011). We therefore recommend the application of these scales in evaluation studies.

Figure 1 lists improvements in self-efficacy beliefs as hypothetical proximal results in coaching. The concept and basic theory was developed by Bandura (1977). He assumes that self-efficacy expectations (e.g. “I am able to perform the desired behavior”) are a central source of motivation and are, among other variables, good predictors of behavior changes. It is possible to differ between global self-efficacy beliefs (for example “If I want, I can solve all problems”), situation oriented (“I am very good at communicating with people”) and task-specific beliefs (“I am very good at leading discussions in my team”). Global scales of self-efficacy have been applied in several studies and have often shown the expected increases (Ely, et al., 2010; Greif, 2007). They belong to the group of general measures. We assume that more specific self-efficacy beliefs, which relate to the themes of coaching, will result in even stronger effects. Webers (2008) has supported this hypothesis in her evaluation study.

One example of a high quality randomized control trial in the field of coaching was a project completed by a team at Sydney University, Australia (Cavanagh, 2010). The study involved 270 participants (lawyers and hospital managers). They were assigned to three groups at random: training with coaching, training only, and waiting list control. Assessments by means of a set of standard scales were performed before, directly after and one year after the intervention. In the outcome evaluation the team applied Goal Attainment Scaling and standard scales as mentioned above, as well as qualitative interviews. A very interesting specific outcome scale that they developed was a questionnaire concerning Perspective Taking Capacity (PTC). They assume that this capacity is required when managing complex organizational changes. In order to manage such changes successfully, it is advantageous to perceive and integrate the different perspectives of the people involved in them. The results of the study confirm the assumption that coaching strongly improves this capacity.

Figure 1 mentions performance improvements as proximal outcomes. Without doubt, the assessment of changes on the individual behavior-level that relate to the goals of the coaching program would be very important evidence. However, there are only very few studies that have applied behavior observation methods and “hard” perfor-
mance measures (for example of improvements in productivity or in meeting performance goals), when assessing coaching outcomes (Ely, et al., 2010; Ely & Zaccaro, 2011; Levenson, 2009). One example is a study by Sue-Chan and Latham (2004). They developed and applied a team behavior observation method and used MBA-grades for appraising performance.

Often, 360-degree-ratings of performance by different sources, especially supervisors, colleagues, subordinates and the persons themselves, are employed as substitutes for laborious behavior observation systems. An example of best practice implementation of coaching combined with before-after 360-degree feedback assessments has been described by Kaufel and his colleagues (Kaufel, 2009; Kaufel, Scherer, Scherm, & Sauer, 2006). However, the feedback-ratings often disagree (especially among subordinates). The methodological aspect of the problem is that the disagreement reduces interrater reliabilities of such measures substantially (Conway & Huffcutt, 1997). The psychological problem is low acceptance of inconsistent negative feedback. According to a meta-analysis of Kluger and DeNisi (1996) most 360-degree feedback appraisal systems are effective, but more than a third are ineffective and lower subsequent performance. DeNisi and Kluger (2000) recommend individual coaching in order to help the recipients interpret the differences in the feedback, cope with negative feedback, and develop a strategy for performance improvement (see also Luthans & Peterson, 2003). In summary, 360-degree feedback assessments are possible - but somewhat knotty - outcome measures, since these measures elicit positive or negative reactions depending on their differences and direction. If in addition the quality of the coaching has an influence on the quality of the resulting measurement of the performance changes, the values of the control groups without coaching are not comparable.

As an alternative for outcome studies of leadership coaching we would recommend reliable and validated leadership behavior scales. A well-established example is the MLQ-scale by Bass and Aviolo (1990). It measures transformational leadership and is therefore particularly suitable for evaluation studies on coaching leaders managing organizational changes. As will be discussed below, a review of meta-analysis studies on leadership development interventions (Avolio, Avey, & Quisenberry, 2010) shows that in most cases leadership development programs improve the economic benefit of the organization.

It is a complex question as to which proximal improvements belong to the learning-level and which to the behavior-level of Kirkpatrick’s taxonomy, since learning can be defined by behavior changes. The allocation of variables therefore depends on the background theory. We assume that skills and traits are hierarchical superordinate constructs that imply concrete behavior changes. We would therefore prefer to classify them as a distinct level above behavior changes and have added this in figure 1. Examples are communication or leadership skills. Improvement of cognitive flexibility, as mentioned by Ely and Zaccaro (2011), would also be classified here. Emotional insight, self-reflection and Perspective Taking Capacity could also be grouped here. It might even make sense to integrate learning and behavior change to form a combined category.

Figure 1 summarizes a short list of specific changes of coaches and the organization. However, this is speculative since studies that apply such measures are missing in the field (Ely, et al., 2010; Ely & Zaccaro, 2011).

**Distal outcomes**

Coaching research has until now mostly been restricted to proximal outcomes. There are only a few studies that cover longer periods of time, for example that of Finn, Ma-
son and Griffin (2006) who assessed self-ratings of changes directly after the coaching and after three months. They found that the ratings of openness to new behaviors and developmental planning showed proximal changes and additional distal improvements. In the field experiment mentioned before, Cavanagh (2010) collected long-term changes after one year. He observed expected increases after coaching, but also unexpected long-term improvements in the waiting list control groups. These effects could be attributed to interventions by the organizations that were not anticipated in the planning of the study. Outside the laboratory it is difficult to establish control groups that are not subject to other influences that may have an impact on the outcome measures.

In figure 1, as examples of hypothetical distal coachee outcomes, we imported several variables from the mentoring model (see below, figure 2) such as promotion, compensation and career satisfaction. Life satisfaction is a distal outcome that is known not only in mentoring but also in coaching outcome studies. The Life Satisfaction scale by Diener, Emmons, Larsen and Griffin (1985) is a standard scale that has been used in several coaching studies.

Examples of long-term coach outcomes that are assumed are income, occupational success and professional credibility. The latter is again, in accordance with Sue-Chan and Latham’s (2004) study, an antecedent of coaching outcomes. Kets de Vries (2010) provokes the profession by assuming that coaching satisfies the rescuer needs of the coaches. A mild need tendency may in his opinion be OK. However, he found that many coaches show problematic rescuer syndrome scale values and behaviors in his questionnaire.

There is a lack of studies that apply measures, which evaluate the long-term outcomes for the organization (Ely, et al., 2010), such as organizational climate, productivity and efficiency. This is a severe deficit. However, one exception seems to exist. Several studies estimate the economic Return on Investment (ROI) of coaching. The most cited estimation has been published in a report by the consulting firm MetrixGlobal and Cylient by McGovern et al. (2001). It amounts to a ROI after coaching of an incredible 545%. M.C. Anderson (2004) similarly found a value of 689% ROI-improvement through coaching. D.L. and M.C Anderson (2005) show how to estimate the ROI for a “coaching that counts” (see also M. C. Anderson, 2011). The simple formula applied here is in principle OK (ROI% = estimated financial benefit from coaching – costs of coaching / costs of coaching x 100). However, a problem is that the coaching benefit is estimated directly by executives and other experts of the organization, who are involved in the coaching program. Typical are direct estimations of the proportion of the total economic profit of a company that are attributed to coaching. Grant et al. (2010) criticize the estimations by using the following example: the overall monetary benefits for the company are large (e.g. US$ 10 million). The executives attribute a high proportion to the coaching (lets say 50%). The coaches’ fee is low (for example 5,000 US$). If we insert these values into the formula, the result would be a ROI of 99,000%

The big question is whether any expert is able to make a fair estimation of the proportion of the ROI that is caused by coaching. The overall ROI of a company is influenced by many complex factors. There may be cases where it is possible to isolate coaching clearly as an initial cause: for example a brilliant solution that later increased the profit of a company may have emerged in a coaching conversation. Also, a problem reflection could have resulted in efficient measures that prevented great economic risks. However, even if coaching may have been the spark of the solution
here, much more has to be done and many more people have to be involved to turn an idea into money.

A more refined and cautious estimation approach of ROI has been applied in the field of leadership training by Avolio, Avey and Quisenberry (2010), that could be used as a model for future research in the field of coaching. The authors have used the methodology of Cascio and Boudreau (Cascio & Boudreau, 2008). The model evaluates HR-interventions over multiple points and therefore allows the endurance and sustainability of their effects to be estimated. The authors have chosen more conservative estimations for transforming performance into dollar benefits. Notably, the basic estimated economic value of the interventions is based on a meta-analysis of 133 studies which had compared the effects with randomly and not randomly controlled groups (Avolio, Reichard, Hannah, Walumbwa, & Chan, 2009). Instead of using subjective estimates, they analyze the strength of the statistical relationship between the interventions and subsequent improvement in leadership effectiveness, so called “effect sizes”. The results show that the expected ROI from leadership development interventions – with a conservative confidence interval of 95% – range from a low negative return to the organization ($460,588) to highly positive returns ($5,811,600 or over 200%). Its magnitude depends on the type and duration of the intervention and the management level. On average, the authors expect a substantial positive dollar return for the organization, but the range is very wide and in some cases even extends into the negative range. We might expect that leadership coaching similarly contributes to leadership effectiveness. Since comparable studies and meta-analyses are missing, this has remained until now a merely speculative assumption.

As in the model by Wanberg et al. (2003), we assume that economic outcomes and other changes on the organizational level are normally distal outcomes (on the result-level of Kirkpatrick’s taxonomy). In our evaluation model the proximal changes (e.g. individual improvements in skills, leadership behavior and style, more efficient self-management, better team leadership, and incorporation of more perspectives in complex change management tasks or conflict management) are intermediate variables that are assumed to have a possible distal impact on economic return and other criteria of organizational outcomes. In the last section of the chapter we will return to the subject.

As has been shown before in a review of evaluation studies in the field of classic group dynamic trainings (Greif, 1976), we should not assume that all outcomes are fully predictable. Coaching is a broadband intervention with many unpredictable adaptations and changes in the process. We should not be surprised if some of the improvements predicted cannot be found. As a consequence, evaluation studies should embrace a broadband set of outcome measures. If unexpected effects are to be expected in evaluation studies, we can apply statistical tests in a different way. Instead of testing the statistical significance of all the single variables, the test model estimates the probability of the distribution frequency of effects in our set of outcome variables. The test should show that the frequency of significant improvements is the result of more than chance.

**How to evaluate mentoring programs**

In mentoring evaluation studies the discussion about providing evidence for the economic return of organizational investments seems to be less intensive in comparison with the field of coaching. There are three possible reasons that might explain the difference. Firstly, as mentioned in the introduction, coaching is performed by psychological experts in confidential conversations and provokes negative, sometimes even discriminatory reactions in some people. Secondly, mentoring is a voluntary activity
and an honorary job often performed by occupational successful mentors. Thirdly, the program costs of mentoring are very low in contrast to coaching. This may therefore be the reason why almost nobody complains that organizational mentoring programs are often evaluated simply by means of protégé satisfaction surveys. However, as we can deduce from more sophisticated evaluation studies, it is useful to analyze the intervention by means of additional criteria.

This section will commence with a description of a comprehensive evaluation model of formal mentoring programs. Secondly, measures of the model's variables and some exemplary studies will be presented. Many of the measures behind the variables in the model are the same or similar to those presented in the coaching field. This allows the following presentation to be abbreviated by naming and explaining only those variables of the model that have not been described above. Also, for a detailed review of the literature on measures applied in studies on the efficacy of mentoring we can refer to chapter 26 by Tong and Kram in this handbook.

Figure 2 shows a modified version of the mentoring model by Wanberg et al. (2003), who tried to summarize over 90 studies of mentoring outcomes. Since mentoring focuses on facilitation of long-term career development, longitudinal studies are very important and cover longer periods than typical coaching studies.

We modified the original figure by Wanberg et al. (2003, p. 92) and organized it in a similar structure to our model in figure 1 (the original contains arrows symbolizing hypothetical relations between the variables). Wanberg et al. (2003) do not specify antecedent mentor and protégé characteristics in their model. A longitudinal study by Blickle, Schneider, Meurs and Perrewé (2010) analyses the perceived barriers to mentoring. They found that protégé characteristics such as low socioeconomic origin and a dispositionally low positive affect reduce participation in mentoring. This reduces the chances of positive outcomes for mentoring programs. Prior positive experiences with mentoring predict positive outcomes. A possible explanation is a better participant knowledge about strategies concerning how to obtain mentoring support and how to surmount possible barriers. These variables could be added to the model as well as further variables from our coaching model in figure 1.

Program antecedents that refer to the quality of the planned mentoring concept are specified very precisely in the model. This mirrors the fact that evaluation studies show that the outcome depends on the quality of the program. It seems to be a waste of time to implement mentoring programs without detailed advance design of meeting manuals and of plans as to how to ensure the quality of the mentoring meetings. The mentors should be prompted, instead of merely talking extensively about their life experience, to focus in the meetings on the needs of the protégé. In figure 2, we have added the costs and time expenditure of the program, since we would expect that if they were very low, then they would indicate low quality and effort.

According to the evaluation research summarized by Wanberg et al. (2003), the organizational context, especially a favorable organizational culture and support for the mentoring program, also influences the outcomes. Blickle et al. (2010) found in their study that a low culture of organizational learning and embeddedness (e.g. low encouragement of relationship building), is a barrier to success. In figure 2, we append “transfer climate”, a technical term from the field of Organizational Psychology that indicates the organizational support of the program.

The model presented in figure 2 includes variables that belong to the mentoring process, such as the relationship (intimacy, interpersonal perception, conflict and complementary nature of interactions). In the coaching process, a positive relationship is also referred to by attributes that are similar in parts. The different terminology mir-
rors a more professional role relationship in contrast to mentoring. We included process factors in the coaching evaluation model that implicate aspects of specialized professional behavior of the coach, which we would not expect in a mentor.

Wanberg et al. (2003) name three variables that characterize the mentoring received: frequency of the meetings (number and duration of the conversations), scope (breadth of subjects and mentoring functions, e.g. learning about other parts of the organization, how to get recognition, or career strategies), and strength of influence (of the protégé by information that meets his or her needs in contrast to superficial ideas). Mentoring relations may fail if only one or two meetings are held and if the scope is narrow or if the content of the conversations has no relevance for the protégé.

According to the model, the proximal outcomes are protégé changes and satisfaction with mentor and mentoring program. A collection of short questionnaire forms and interview questions has been published by Saito (2001). It can be applied in the evaluation of youth mentoring programs and may be adapted to other target groups. Wanberg et al. (2003) refer to the classification by Kraiger, et al. (1993) mentioned above, who differentiated between the “learning” level of Kirkpatrick’s taxonomy and also distinguish between cognitive, skill-based and affective learning. Tong and Kram mention research in chapter 26 of this handbook, which supports the theory that mentoring can improve performance and short-term development of personal skills or relational skills learning. Since mentoring is about developing networks, Wanberg et al. (2003) add social networks to their model. As mentioned in chapter 26 of this handbook, developmental networks are a focus of contemporary mentoring research.

Karcher et al. (2006) propose a model of the causal relationship between proximal outcomes (for example increased social support or improvement) and distal results (for example gains in achievement). They assume that their relationship is mediated by enabling intermediate outcomes (for example increased self-esteem). Typical focal distal outcomes that have been assessed in many mentoring evaluation studies are career success indices, especially promotion and compensation (see chapter 26 for a review of the literature). In the model they are classified as extrinsic outcomes. Blickle et al. (2010) in their longitudinal study assess yearly gross income, reported by the protégé as an outcome measure. Similar to the coaching model, figure 2 lists improvements in the job and life-satisfaction of the protégés and specifically career satisfaction, which are classified here as intrinsic outcomes.

Since the mentor is a volunteer and is not normally compensated for the mentoring activity, an obvious question is what might his or her alternative short- and long-term outcomes be. Wanberg et al. (2003) assume that, similar to informal mentoring, the benefit is personal satisfaction. Qualitative research supports the view that mentors perceive feelings of self-satisfaction, accomplishment and renewed meaning in their working lives, or the personal loyalty of the protégés. In addition, we might assume that, similar to coaching, mentoring could satisfy the rescuer needs of a proportion of the mentors.

Assumed distal organizational outcomes of formal mentoring programs include improvements in organizational commitment, loyalty, and retention, organizational communication, managerial succession, productivity and performance, and perceived justice.
Outcome variables such as promotion, retention, managerial succession and perceived justice are specific to mentoring, but there are similarities to our evaluation model of coaching. It is therefore possible to apply the same standard scales of satisfaction, goal attainment, negative affect, and adapted versions of general and specific self-efficacy and result-oriented problem- and self-reflection, as Klien (2011) shows in her evaluation study (control group not randomized) of a mentoring program for university students. Again, the advantage is that if we use standard scales, we are able to benchmark and compare the outcome with other mentoring or coaching program implementations.

Qualitative evaluation
The evaluation models and research literature presented above concentrate on variables that have been assessed by quantitative methods. However, in the field of coaching, there is a debate about their value. Many researchers prefer qualitative methods (Fatzer, 2008) and some favor a combination (Greif, 2011b). As mentioned in the introduction, qualitative evaluations and reports of successful programs can have an important influence on the decision to start coaching or mentoring programs. They provide concrete observations that are more comprehensible and convincing than scale numbers.

In the following section a short summary of recommendable qualitative evaluation methods will be presented that can be applied to both coaching and mentoring. The section will start with a short introduction to phenomenological analysis and its difference to methods of grounded theory development. Secondly, narrative interview methods will be introduced that can be used to protocol rich stories of experiences. Then, a standard evaluation method will be described that combines qualitative interviews, structure analysis and quantitative ratings. Finally, recent evaluation approaches will be referred to that apply linguistic micro-analysis to transcripts of interactions between coach and coachee.

Interpretive phenomenological analysis and grounded theory development
Qualitative phenomenological analysis has a long tradition in the Social Sciences. It goes back to the German philosopher and phenomenologist Edmund Husserl (1859-1938). He claimed that it is possible to grasp the essence of phenomena by means of intuitive but “objective” contemplation, the so-called “Wesensschau” (perception of the essential meaning). Recently, Jonathan Smith (1996) introduced a modern method of Interpretive Phenomenological Analysis (IPA). It has since become popular as an approach to analyzing and interpreting narratives of individual experiences from interviews and also diaries or focus groups (Smith, Folwers, & Larkin, 2009). After transcription of the narratives, the interpreter reads the text several times. He or she makes notes and a record of the words of the participants and his or her interpretations of the text, focusing on the perspectives of the participants. Repeating patterns that emerge in a bottom-up exploration and interpretation process are called “themes”. Finally, the clustering of the themes is organized in a hierarchical table.

Gyllensten and Palmer (2007) give an example of the application of IPA to the analysis of nine interviews on workplace coaching experiences. They identified “the coaching relationship” as a main theme with three subthemes: valuable coaching relationship, trust and transparency. Other themes are “working toward goals” and “improving performance”. They conclude that it is important that coaches are aware of and work with these themes.
The method of *Grounded Theory* (GT) is also a bottom-up approach that can be applied when analyzing qualitative interview data. The first version was published by Glaser and Strauss (1967). The basic idea was to design an alternative research strategy to mainstream top-down hypothesis-testing science and to develop more local theories that are “grounded” in qualitative data. Similar in part to IPA, the interview data are analyzed and categorized bottom-up. However, the theoretical categories from the beginning of the analysis are checked between different participants. The research strategy also permits additional studies to be performed that review, refine or even reform the emerging theory. Later, Glaser and Strauss designed different versions of GT. For an overview we refer to Coolican (2009, p. 231).

Wilkins (2000) interviewed twenty-two coaches by telephone. In her GT analysis a theoretical process model emerged. It characterizes the essentials of the coaching process as an interaction between coach and coachee, seeking to develop the coachees to their fullest potential. Passmore (2010) also analyzed transcripts of semi-structured interviews by means of a GT approach. He identified multiple key behaviors and attributes of executive coaches, for example confidentiality, being collaborative, setting take-away tasks, stimulating problem-solving, helping develop alternative perspectives, that use of a variety of focusing tools and “self as a tool”.

**Narrative interviews**

Understandable descriptions of experience with concrete cases or successful program implementations are useful for substantiating the practical value of coaching or mentoring. Narrative interviews are methods that can be applied in order to assess and analyze the experiences of coaches or protégés and the managers who implemented the programs. Gold, Devins, & Johnson (2003) discuss the applicability of narrative interviews with managers in the field of mentoring. They claim that they are more appropriate than questionnaire evaluation in small businesses and industry. They provide illuminating stories with rich context information which help to convince the managers of the value of the intervention.

König (2005) designed a *Construct Interview*, which is a further interesting example of qualitative coaching evaluation, since here the interviewer not only asks the coach to describe his or her experiences with a coaching case, but also to explicate his or her subjective assumptions about the coaching processes and outcome. Only four to six open core-questions are determined in the interview guideline (for example the start and development of the coaching process, outcomes, and factors that have led to them). The interviews should not be performed in a standardized form, but rather in a way that provides open space for a reconstruction of experiences.

**Change Explorer interview**

The *Change Explorer* is a theory-based evaluation method that combines semi-standardized interview and rating scales. The purpose is to assess and reflect the subjective reconstruction of the evaluation models of the interview partners. This is used as a basis for making concrete improvement suggestions. It assesses the personal ratings of the success or failure of an intervention or change and goal attainment of its sub-goals. The interviewees are requested to describe perceived outcomes as concretely as possible. In addition, they are asked to evaluate the methods or other measures used in the coaching sessions and to explicate the subjective causes of all specified outcomes.

Originally, the methods were applied widely as an instrument for the evaluation and improvement of organizational change processes (Greif & Seeberg, 2007; Jacobs, Keegan, Christe-Zeyse, Seeberg, & Runde, 2006). Adapted versions can be used in the field of coaching (Greif, 2008, S. 239 ff.; Greif, Runde, & Seeberg, 2005; Rheber-
A download of the guideline as to how to apply it as a self-evaluation and reflection tool between coach and coachee in the closing session of an individual coaching session, is accessible (Greif, Seeberg, & Santaniello, 2011). The feedback protocols of the interviews show that the common reflection, stimulated by the method, facilitates a deep common understanding between coachee and coach of the intangible processes, and possible causes and outcomes of coaching. Moreover, it helps the coachee and coach to explicate and communicate concrete results, and also how coaching works, to other people. We therefore recommend it as a standard qualitative evaluation method for coaching and mentoring.

**Linguistic analysis**
Graf, Aksu and Rettinger (2010) have published an exemplary qualitative linguistic analysis of transcripts from authentic coaching narratives. They analyze and interpret two short excerpts from the dialogues between coach and coachee in which the coach explains his or her coaching concept. Characteristic for the method of conversation analysis is a detailed transcript of all verbal and para-verbal interactions. Subsequently, the micro-process is described and abnormalities are identified (e.g. hesitant talking, vague formulations or dominant behavior by the coach, when he or she started to explain the coaching concept). The conversations are interpreted theoretically as common negotiations and constructions of the identities of the coach and coachee. The practical goal is to stimulate a critical exchange between practitioners and scientists from different disciplines on authentic micro-processes of coaching and, in the long run, to develop common standards for the evaluation of coaching narratives.

**Challenging questions of evaluation research**
Coaching and mentoring are complex and intangible services. The outcomes can be very concrete, but it is difficult for customers to understand how they are co-constructed in conversations. As Schneider and Bowen (1995, p. 93) recommend, it is essential to help the customers clarify his or her co-production role. “Providing customers with realistic service previews (RSPs) can help them to make informed decisions about whether they want a co-production role and how they could perform in it.” Schneider and Bowen (1995, p. 95) emphasize that such information enhances the quality of the service and also customer esteem and loyalty. Evaluation models and methods contain the relevant information that can be utilized here. As a conclusion, practitioners of coaching and mentoring are encouraged to employ evaluation models as guidelines as to how to inform their customers. Knowledge of the results of qualitative and quantitative evaluation studies can be used to help customers to develop realistic expectations regarding the expected value of the intervention. This would be a profitable marketing investment.

A large variety of quantitative and qualitative methods can be applied when planning and conducting program evaluation studies. The discovery of the results of coaching or mentoring programs depends on the selection of adequate outcome measures and qualitative assessment methods. When planning an evaluation project, how can we select the relevant variables and evaluation methods pragmatically? The first step is to decide on an evaluation model, which embraces the variables that are relevant according to the theoretical assumptions of the evaluators and the demands of the organization. The second step is to answer three questions that will help us to select

---

1 The program, based on this evaluation, won the award “Diversity as an Opportunity” of the Federal Commissioner for Migration and Integration of Germany.
adequate methods and outcome criteria. (1) Are reliable and valid standard qualitative and qualitative methods available? (2) Are the scales and methods specific and sensitive enough to assess the effects of our concrete intervention? (3) Will the results be comprehensive, relevant and convincing for the clients?

Once again, we recommend selecting reliable and valid standard quantitative measurement instruments and combining them with standard qualitative methods wherever possible. Most of them are very short and there are not many reasons that speak against them. Why develop homemade scales of goal attainment, stress, well-being or leadership behavior, if tested and tried scales exist and we have neither time nor money to test the reliability and validity of our homemade scales before we apply them in our study? For benchmarking we need methods with tested quality. In addition, we recommend using standard qualitative methods for qualitative benchmarking.

However, even where similar measures have been applied in randomly controlled evaluation studies, the results are not consistent (Grant, et al., 2010). One study shows this result while the next shows another, and other studies even show no statistically significant results. As mentioned above, the outcomes are not fully predictable. Possible reasons are heterogeneity of the goals and context conditions or differences between the concepts and quality of the coaching or mentoring interventions and programs. Therefore, it is advantageous to expect unexpected outcomes. It would be wise to allow for a broadband of criteria and to include qualitative methods that are open for exploring new evidence bottom-up. Another conclusion is not only to evaluate the outcomes, but at the same time also to compare the quality of the program and the interventions, as listed in our evaluation models.

The high road for improving low outcome is to use evaluation studies as feedback and to improve the quality of the interventions until the results are in the desired range. This is not only true for practical evaluations, but also for scientific research studies, where the quality of the intervention is often low. Sue-Chan and Latham’s (2004) study for example, the “education” of the “coaches” took less than half a day and the coaching was very short. I like the study because it is one of the few that assesses and provides evidence that coaching improves observable behavior change and the final grades of MBA-education, but actually, for me it still remains a miracle that such a “mini-coaching” can do that.

The second question is more intriguing. Discovery often requires the development of specific measurement instruments. For example, in order to detect nuclear radiation, specific measurement instruments had to be developed. Without a Geiger counter people exposed to radiation cannot sense the magnitude and believe that they are safe. I expect that one of the reasons why we do not detect consistent outcomes of coaching and mentoring is that we have not invented the requisite methods that are specifically sensitive to discovering all relevant outcomes. Methods applied in today’s generation of coaching and mentoring evaluation research are in my view a bit like poking around with a rod in the fog. One conclusion, as mentioned above, is to develop methods that are able to assess such specific outcomes. Simple examples are scales of specific efficacy beliefs that relate specifically to the interventions or to specific result-oriented problem- and self-reflections (Greif & Berg, 2011). The Perspective Taking Capacity scale (Cavanagh, 2010) is an example of a more complex methodological innovation. In the future, we should encourage and invent further evaluation methods that are adapted to the fields of coaching and mentoring.

The answer to the third question from a practical perspective is perhaps the most important. As in the nice example of Kirkpatrick mentioned above who helped his friend
to convince his boss, it is not an easy task to sell an evaluation study to customers. To test whether our evaluation model and the methods selected are comprehensible, relevant and convincing to our customers, we may try to explain the concept to a practitioner friend and ask for feedback. If he or she is not convinced, it would be better to improve the message and/or the measures!

Critics challenge coaching by demanding proof of its utility. An evaluation of the Return on Investments (ROI) is often judged as the ultimate measure of success of organizational coaching programs (Grant, et al., 2010, p. 142). As discussed above, existing studies estimating the ROI are open to criticism since they are based on assailable subjective estimations. They would not convince intelligent critics. In the Change Explorer interviews described above, we routinely ask whether the coaching resulted in measurable economic outcomes. Only in a few exceptions were the professional coaches able to explicate concrete examples.

ROI or other economic criteria are not the most important outcomes and are often even irrelevant measures of the success of coaching and mentoring. Like Kirkpatrick and his friend, we have to explain this to the customers. As the evaluation model shows, coaching is a broadband intervention, which can be adapted to many different purposes and goals. It can lead to a broad variety of very useful results. Examples are improvements in individual behavior and performance, development of skills and potentials, stress reduction, a better work-life balance, well-being and life-satisfaction, and organizational commitment and productivity. The open question is whether these important changes have a direct or long-term effect on the economic return on the organization’s investment. Mentoring is more specific. It can promote the career and raise individual income, but it can also improve a broad range of further outcomes. This broadness and variety of outcomes is the major return of investing time and money in coaching and mentoring programs.

Conclusions
The general purpose of the evaluation of HR-investments is to obtain an objective, reliable and valid information basis for decision-making. Investment decisions should not rely on subjective evaluations alone: data from multiple methods and different levels have to be included. This chapter presents comprehensive evaluation models of organizational coaching and mentoring programs, which show how to evaluate relevant antecedents, organizational context, processes, and the expected short- and long-term outcomes of successful program implementations. These models provide a basis for planning and managing evaluation studies as well as for program improvement.

This chapter reviews and discusses a variety of quantitative measures, including estimations of economic return, which can be recommended for the evaluation of programs in organizations and in scientific research. Standard scales that meet scientific criteria and have proven their value in several evaluation studies are highly advisable. It is favorable to combine them with qualitative methods in order to incorporate rich context information and case histories. These models and methods can then be used to inform customers as to how complex and intangible coaching and mentoring services work, as well as how they can be adequately evaluated and optimized.
References


