Rezension zu:

Kirkpatrick, Donald L./Kirkpatrick, James D.: Evaluating Training Programs. The Four Levels. 3rd ed. San Francisco u.a.: Berrett-Koehler 2006. 379 Seiten, 42,95 $ (Hardcover), ISBN: 978-1-576-75348-4

Wolfgang Beywl
Universität Bern, Koordinationsstelle für Weiterbildung

Die Auswirkungen von Trainings hat Donald Kirkpatrick erstmals 1959 als Ertrag seiner Dissertation viergegliedert. Die Essenz des zum 3. Mal überarbeiteten Bestsellers hat sich beinahe auf dem ganzen Globus verbreitet. Die Four Levels bilden das Fundament vieler Evaluationssysteme in Unternehmen, tausender Artikel, Bücher oder akademischer Qualifizierungsarbeiten. Erstmalig sind Vater (Donald) und Sohn (James) gemeinsam Herausgeber. Letzterer hat insbesondere das Kapitel zur Verbindung der «Four Levels» mit der «Balanced Scorecard» beigesteuert.

Der Titel macht den Geltungsbereich des Buches deutlich: Es behandelt vorrangig «training», was «development» einschließt – dabei fokussiert auf Führungs- und Managementtraining. Es geht weniger um «education», also Bildung und Erziehung, die auch von materiellen Verwendungszwecken (weitgehend) frei sein können. Donald Kirkpatrick meint zwar im Vorwort, dass die präsentierten Konzepte auch auf «academic courses» übertragbar seien, doch scheint mir dies auf Angebote der «corporate universities» beschränkt, die einige der 16 präsentierten Fallbeispiele beisteuern.

Neben den würdigenden Geleitworten zweier Unternehmensführer und Donald Kirkpatricks eigenem Vorwort sowie einem hilfreichen Stichwortverzeichnis besteht das Werk aus zwei großen Teilen: Theorie (Kapitel 1 bis 11) und Fallstudien (Kapitel 12 bis 27).

Teil Eins («Concepts, Principles, Guidelines, and Techniques») verortet zunächst die Evaluationsfunktion als abschließenden von zehn didaktischen Schritten, in denen Trainingsprogramme geplant und umgesetzt werden (empirische Bedarfsanalyse, Zielklärung usw.). Die Notwendigkeit für Evaluation – und den Motor, das Buch zu schreiben – sieht Kirkpatrick in der ständigen Bedrohung der Trainingsfunktion im Unternehmen durch Downsizing oder Outsourcing, da ihr oft Belege für den Erfolg fehlten. Er nennt die drei klassischen Evaluationszwecke: Rechenschaftslegung, Entscheidungsbasierung und Verbesserung.

Die Kapitel 3 bis 7 – 50 klar geschriebene Seiten – sind der theoretische Kern des Buches. Auswirkungen von Trainings können demnach auf vier Ebenen verortet werden:

(1) Reaktion («reaction») der Teilnehmenden auf das Angebot als Ausmaß der Kundenzufriedenheit;
(2) Lernen («learning») der Teilnehmenden im Lernfeld, d.h. Zuwachs an Wissen, Veränderung von Einstellungen und Erwerb von Fertigkeiten («skills») gemäß der Programmziele;
(3) Verhalten («behavior») der Teilnehmenden im Funktionsfeld, also Anwendung des Gelernten ‚on the job';
(4) Systemresultate («results») im Unternehmen, z.B. bezüglich Mitarbeitendenfluktuation, Umsatz oder Gewinn.

Die *Planung* der Programme soll ausgehen von den gewünschten Systemresultaten über das Verhalten und Lernen bis zu den Reaktionen. Für jedes der Four Levels sind optimalerweise durch Messgrößen («metrics») operationalisierte Erfolgskriterien («standards») zu setzen, und zwar

vorgängig zum Training. So entsteht für die Beurteilung ein gesicherter Referenzrahmen. Die *Evaluation* von Programmen erfolgt in umgekehrter Reihenfolge, von den Reaktionen bis zu den Systemresultaten. Im Idealfall soll die Evaluation keines der Four Levels auslassen:

– Keinesfalls dürfen tiefere Levels ausgelassen werden, da sie für das Unternehmen weniger wichtig zu sein scheinen. Das Ausbleiben (oder auch das Eintreten) von gewünschten Systemresultaten/Verhaltensweisen kann auf Veränderungen/Probleme in der Umwelt des Unternehmens oder in den Transferbedingungen zurückgehen. Das Programm kann (in Bezug auf Zufriedenheit und Lernen) dennoch erfolgreich sein (was bei Auslassen der ersten Levels unerkannt bliebe).
– In der Praxis werden höhere Levels wegen fehlender Messinstrumente, Zeit oder Geld oft ausgelassen. Je höher das Level, desto aufwändiger die Messung und desto differenzierter das Design (Pre-Post-Test/Kontrollgruppen usw.), um auch in spannungsreichen betrieblichen Kontexten überzeugende Evaluationsergebnisse zu erzielen.

Der Evaluationszyklus verläuft auf jedem der vier Levels weitgehend gleich: Erhebungsgegenstand bestimmen – Erhebungsinstrument entwickeln – Erhebung durchführen, dabei Störgrößen minimieren – Ergebnisse mit gesetzten Erfolgskriterien vergleichen und angemessene Maßnahmen einleiten – Ergebnisse kommunizieren.

Level 1 – Reaktion – ist für Kirkpatrick grundlegend für den Trainingserfolg: „Positive reaction may not ensure learning but negative reaction almost certainly reduces the possibility of its occurring“ (S. 22). „If training is going to be effective, it is important that trainees react favorably to it“ (S. 27). Die Messung der Kundenzufriedenheit geschieht mittels kurzer Feedbackbögen («reaction sheets»). Diese können im Detail zwar sehr unterschiedlich gestaltet sein (das Buch enthält ein gutes Dutzend Beispiele), entsprechen dabei allerdings weitgehend folgendem Muster:

– es gibt überwiegend geschlossene (Einschätzungs-)Fragen, eventuell auch einzelne offene Fragen, insbesondere nach Verbesserungsvorschlägen;
– Standardthemen sind die subjektive Wichtigkeit des Trainingsinhalts/-themas, die Beurteilung der Trainierenden sowie der Ausstattung/der Lernunterlagen.

Kirkpatricks Tipp (auch für Level 2 gültig): Erhebe noch während des Trainings, um 100 Prozent Rücklauf zu erhalten.

Level 2 – Lernen – ist aus Kirkpatricks Sicht eine ebenso wichtige («without learning, no change in behavior will occur») wie einfach zu bewältigende Aufgabe: Die zu messenden Lernresultate können aus den Programmzielen abgeleitet werden oder sind in ihnen bereits operationalisiert. Die Erhebungen sind oft Teil des Trainings selbst: Die Basis wird durch Lerneingangstests gemessen. Abschlusstests und Prüfungen erheben den Lernstand zum Ende des Programms. Die Differenz kennzeichnet den Lernzuwachs. Optimalerweise werden diese Messungen auch an einer Kontrollgruppe vorgenommen. In der Praxis wird man aufgrund von Zeit- und Budgetrestriktionen immer wieder auf eines oder mehrere dieser Elemente verzichten.

Level 3 – Verhalten – ist weitaus schwieriger zu messen: Die Teilnehmenden haben den Kursraum verlassen und agieren – meist weit verstreut – in ihren alltäglichen Arbeitsumgebungen. Wenn dort das gewünschte Verhalten auftritt, muss genügend Zeit seit dem Lernereignis vergangen sein. Die Umsetzungsbedingungen – dazu zählen insbesondere die Unterstützung der Vorgesetzten und darüber hinaus Anreizsysteme – müssen stimmen. Im Idealfall wird das Verhalten im Transferfeld aus mehreren Perspektiven erhoben: durch die Weitergebildeten selbst, durch Vorgesetzte, durch Untergebene und vielleicht durch speziell geschulte Beobachtungspersonen. Dies ist wie vieles eine Frage der Relation von (Evaluations-) Aufwand und Ertrag. Oft wird man sich auf Selbstauskünfte der Programmteilnehmenden beschränken müssen. Die Fallbeispiele des zweiten Teils enthalten Beispiele für solche Varianten.

Level 4 – Systemresultate – ist, je nach Entwicklungsstand des betrieblichen Rechnungs- und Berichtswesens, unterschiedlich schwer oder leicht zu messen. Es geht um ‚harte‘ («tangible») Daten wie Einhaltungsgrade von Qualitätsstandards, Produktivitätszuwachs, Senkung der Mitarbeitendenfluktuation oder Zuwachs an Unternehmensgewinn oder Aktienwert. Es geht auch um ‚weiche‘ (oft «intangible») Daten zum Unternehmensklima oder zur Innovations- oder Zukunftsfähigkeit. Im Idealfall lassen sich die Trainingsinvestitionen mit den daraus resultierenden Erträgen verrechnen zum «Return on Investment – ROI» (was in zwei der Fallbeispiele – mit erstaunlich guten Quoten – demonstriert wird). Erforderlich sind – wie schon für Level 2 – umfas-

sende empirisch-methodische und auch statistische Kompetenzen. Die Hauptklippe liegt in der Zurechenbarkeit der Wirkungen: In welchem Maße ist der jeweilige Erfolg auf das Training, in welchem auf andere Faktoren (Technologiesprung, Konjunkturbelebung) rückführbar? Kontrollgruppen – so wünschenswert sie sein mögen – werden hier selten zu bilden sein.

Teil Zwei («Case Studies of Implementation») enthält 16 überwiegend von Trainingsverantwortlichen in Unternehmen, Trainingsanbietenden sowie Evaluationsfachleuten beigesteuerte, oft aktuelle Fallstudien, welche die Anwendungsvarianten der Four Levels demonstrieren. Alle Beispiele stammen aus Unternehmen oder anderen hierarchisch geführten Organisationen, meist aus den USA. Einen Schwerpunkt bilden Führungstrainings, oft zwischen einem und drei Tagen Dauer. Benachbart sind die drei Beispiele zur Einführung neuer Mitarbeitenden, zur Nachwuchsplanung und zu einer Coaching-Ausbildung. Zwei Fälle betreffen die Einführung komplexer DV-Systeme. Weitere sind eine Standard-Office-Schulung sowie ein Verkaufstraining (dazu auch mehrere Instrumente im Teil Eins). Zwei Beispiele sind themenunspezifisch und stellen Standardinstrumente für Trainings aus verschiedensten Bereichen vor. Besonders bemerkenswert ist das letzte Fallbeispiel, in dem es um die Evaluation eines weitgehend individualisierten, selbstgesteuerten Trainingsprogramms in Zusammenhang mit einem unmittelbar auf Produktivitätssteigerung angelegten Informationssystem bei einem führenden Anbieter für Netzwerklösungen geht (über alle Four Levels hinweg).

Das methodische Anspruchsniveau bietet eine große Bandbreite: Vom auf Level 1 beschränkten Feedbackbogen bis hin zu komplexen, multimethodisch und multiperspektivisch angelegten Designs.

Kirkpatrick lädt mehrfach zum „Ausleihen" der vorgestellten Instrumente und Evaluationslösungen ein. Gleichzeitig weist er darauf hin, dass immer eine Anpassung auf das zu evaluierende Programm, auf das jeweilige Unternehmen und die aktuellen technologischen und ökonomischen Bedingungen vorzunehmen ist. Für alle, die vor einer Evaluationsaufgabe in der betrieblichen Praxis stehen, bieten die Praxisbeispiele einen Fundus an Anregungen und methodischen Vorbildern.

Diskussion

An diesem Buch darf man als Anbieter von Trainings oder als der im Unternehmen für Training Verantwortliche nicht vorbei, schon weil alle Kolleginnen und Kollegen die Four Levels im Munde führen und die Viergliederung zu den absoluten Basics der betrieblichen Weiterbildung gehört. Beeindruckend ist die Erfahrungsbasierung des Ansatzes, bei fast vollständigem Verzicht auf Belege aus der Forschung oder auf Querbezüge zu Standardliteratur zur Weiterbildungsdidaktik oder Evaluation. Kirkpatrick will und kann mit seinem Ansatz allein stehen. Er entfaltet dabei ein originäres Evaluationsverständnis: Zyklizität der Studien, Anpassung an das jeweilige Trainingsprogramm und Gebundenheit an den Kontext. Dies findet sich auch in seiner Beschreibung von Evaluation als Wissenschaft und Kunst: „As a science it is organized knowledge – concepts, theories, principles, and techniques. And as an art it is application of the organized knowledge to realities in a situation, usually with blend or compromise, to obtain desired practical results" (S. 74).

Sein Pragmatismus ist eindrucksvoll. So soll Training «PIE» sein: «Practical – Interesting – Enjoyable», mit Referenz an den deutschen Reformpädagogen Peter Petersen, bei dem es Hand, Haupt und Herz heißt. Auch für das Managen von Wandel hat Kirkpatrick Merkwörter parat: Empathie, Kommunikation, Partizipation (vgl. S. 81). Sätze wie „Something beats nothing" oder „(…) be satisfied with evidence, because proof is usually impossible to get" kennzeichnen seine Grundhaltung.

Der Text ist leicht verständlich, ohne Schnörkel und entbehrliche Fachtermini. Er ist eine Meisterleistung in der Anwendung des Pareto-Prinzips: Mit einem schmalen Text erledige ca. 80 Prozent der anstehenden Aufgaben in guter Qualität; überlass den Rest hoch spezialisierten Fachleuten. Man kann hier lernen, wie man einen Lehrbuch-Bestseller schreibt.

Folgende Vorbehalte möchte ich anmelden:

Der erste betrifft Kirkpatricks uneingeschränktes Plädoyer dafür, die Zufriedenheitswerte (Level 1) zu maximieren (es finden sich in allen Fallstudien hohe Zufriedenheitswerte). Ich meine es gibt Fälle, gerade in der Weiterbildung von Führungskräften, von OrganisationsentwicklerInnen, TrainerInnen oder Coaches, die personales Lernen und Reifen erfordern, die auch Pha-

sen, mit sich selbst, der Lerngruppe und den Lernbegleitenden ‚unglücklichen Lernens' («unhappy learning») erfordern. Aber vielleicht zählen diese Fälle zu den oben angesprochenen 20 Prozent für Spezialisten.

Der zweite Vorbehalt betrifft diejenigen, welche Kirkpatricks Strickmuster auf andere Bereiche anwenden, für das es nicht entwickelt ist (seine Wendung, es sei auch für «academic courses» geeignet, empfinde ich als zweckoptimistisch). Es gilt für Trainings, die in «corporate settings» angewendet werden, in denen der Beitrag des Lernens für die Zielerreichung der Organisation (Level 4) außerhalb jeden Zweifels nicht nur eine legitime Anforderung, sondern darüber hinaus auch die letztlich entscheidende Kriteriendimension ist. Dies ist außerhalb der Unternehmen und unternehmensartig geführten Organisationen nicht gültig, z.B. für Kindertageseinrichtungen, Schulen, Hochschulen, Volkshochschulen und andere öffentlich finanzierte Bildungsanbieter. Hier gilt die Anforderung demokratischer Aushandlung der Bildungsziele und Festlegung der Bewertungskriterien. Es gibt dort Multi-Stakeholder-Settings und Auseinandersetzungen darüber, was die zentralen Evaluationsfragestellungen sind, und was die richtigen und wichtigsten Ziele des Programms sein *sollen,* was andere/ergänzende Evaluationsansätze erfordert.

Dies leitet zum dritten Vorbehalt über: In ausnahmslos allen Fallbeispielen des Buches wird (meist außerordentlicher) Erfolg vermeldet. Dies setzt zum einen voraus, dass die Programme gut geplant sind (kurz im Einleitungskapitel angesprochen), insbesondere dass ihre Ziele klar definiert und möglichst operational formuliert sind. Ist diese Voraussetzung nicht gegeben (was auch in betrieblichen Settings vorkommt), greift der Evaluationsansatz von Kirkpatrick ab Level 2 buchstäblich ins Leere. Diese Voraussetzung hoher didaktischer Qualität der Programme sollte Kirkpatrick deutlicher ansprechen (bei mehreren Fallbeispielen wird diese eindrucksvoll deutlich). Zum anderen habe ich den Verdacht, dass die Kultur des Profit-Bereichs oft, wenn nicht Erfolg, so doch zumindest Erfolgsmeldungen *erzwingt*. Eingeständnisse von Lücken oder Fehlern sind riskant für die Akteure. Hingegen gibt es im Non-Profit-Bereich eine (vielleicht vielfach überzogene) Kultur der Skepsis bezüglich der Machbarkeit von Lern- und darüber hinausgehenden Erfolgen.

Es macht jedenfalls Sinn, Analogien zwischen den originären und anderen möglichen Anwendungsbereichen der Four Levels zu ziehen und das Denken kann auch dort die Perspektiven erweitern. So wenig, wie das Buch für die Evaluation der betrieblichen Weiterbildung verzichtbar ist, so wenig reicht es allein aus, um in anderen Feldern der (Weiter-) Bildung dem Stand von Wissenschaft und Kunst entsprechend zu evaluieren.