

Programmevaluation zur Ermittlung des Wertbeitrags betrieblicher Trainingsmaßnahmen

Dr. René Fahr und Anthea Friedrich, beide Seminar für Allgemeine Betriebswirtschaftslehre
und Personalwirtschaftslehre
Wirtschafts- und Sozialwissenschaftliche Fakultät
Universität zu Köln

Albertus-Magnus-Platz
50923 Köln

Kontakt:

Dr. René Fahr

Tel: 49-(0)-221-470-6312

Fax: 49-(0)-221-470-5078

email: rene.fahr@uni-koeln.de

Programmevaluation zur Ermittlung des Wertbeitrags betrieblicher Trainingsmaßnahmen

Februar 2006

Zusammenfassung

Der Beitrag liefert einen Überblick über verschiedene Methoden zur Messung des Erfolgs betrieblicher Trainingsmaßnahmen mit einem besonderen Augenmerk auf einem Vergleich der Ausgestaltung der Ergebnisevaluation. Hierzu erfolgt eine Darstellung der Programmevaluation in der betriebswirtschaftlichen, ökonometrischen und psychologischen Literatur. Der Beitrag verdeutlicht die Notwendigkeit eines einheitlichen und interdisziplinären Konzeptes zur Evaluation von Trainingsmaßnahmen, um der Voreingenommenheit der Praxis gegenüber einer quantitativen Evaluation zu begegnen.

Schlüsselbegriffe: Programmevaluation, Humankapital, Personalentwicklung, Training

Abstract

We overview different methods to measure the returns to training and compare them with respect to the design of the outcome evaluation. Specifically, we describe different approaches of program-evaluation in the management, econometric and psychological literature. This contribution shows the necessity to come up with a uniform and interdisciplinary approach to evaluate firm-based training. By this it might be possible to overcome the practitioners' resentments towards quantitative evaluation approaches.

Keywords: Program-evaluation, Human Capital, Human Resource Development, Training

1. Einleitung

Die zunehmend wertorientierte Sichtweise in den Personalabteilungen führt dazu, dass das Verlangen nach konkreter und fundierter Information über den Erfolg und Nutzen von Trainingsmaßnahmen zunimmt. Gerade die Personalentwicklung hatte die Maximierung der Anzahl der Trainingstage lange Zeit als primäres Ziel. Der Kostendruck in Unternehmen und die Umgestaltung von Personalentwicklungsabteilungen zu Cost- und Profitcenter führt dazu, dass der Wertbeitrag der Personalentwicklungsmaßnahmen quantifiziert werden muss und damit die Ergebnisevaluation an Bedeutung gewinnt.

Dies erfordert jedoch den Nachweis eines kausalen Zusammenhangs zwischen der Trainingsmaßnahme und der Ergebnisvariablen. Die Ermittlung eines solchen kausalen Zusammenhangs hat sich die Forschung zur ökonomischen Programmevaluation unter Verwendung (mikro-) ökonometrischer Modelle als Ziel gesetzt hat. Mit wenigen Ausnahmen werden die Erkenntnisse in der ökonomischen Programmevaluation, die zum Beispiel zur Beurteilung der Effektivität von Maßnahmen der aktiven Arbeitsmarktpolitik eingesetzt werden (Lechner 1998) in der betriebswirtschaftlichen Literatur wenig¹ und der unternehmerischen Praxis gar nicht beachtet. So findet Gülpen (1996) in einer Befragung von 91 Personalverantwortlichen deutscher Unternehmen, dass 60% der befragten Unternehmen den Unternehmenserfolgsbeitrag von Trainingsmaßnahmen nie untersuchen und 25% selten. Damit wird die Tendenz einer früheren Studie von Pullig (1987) bestätigt. Aber auch außerhalb Deutschlands beschränken sich die Unternehmen meist auf die Evaluation der Zufriedenheit mit der Trainingsmaßnahme (ASTD 2003). Hinweise darauf, warum der für Unternehmen eigentlich zentrale Erfolgsbeitrag von Trainingsmaßnahmen so selten untersucht wird, finden sich ebenfalls in der Studie von Gülpen (1996). So wird den Kriterien „Aussagekraft der Ergebnisse“ sowie „Feedbackmöglichkeit durch Teilnehmer“ bei der Wahl einer Evaluationsform für Trainingsmaßnahmen die höchste Bedeutung beigemessen. Die vier unwichtigsten Kriterien finden sich in der Reihenfolge beginnend mit dem unwichtigsten Kriterium: Kontrolle über die Teilnehmer, wissenschaftliche Fundierung, quantitativer Charakter der Ergebnisse und Möglichkeit der Legitimation des Seminars. Aufschlussreich ist in Gülpen (1996) auch die Frage bezüglich der Gründe für die Nicht-Quantifizierung einer Erfolgsbewertung. Bei vier vorgegebenen Alternativen wird mit 59,6% fehlendes Know-How am häufigsten genannt, gefolgt von „zu teuer“ mit 58,30%. Der Grund „interne Widerstände“

¹ Eine Ausnahme bildet hier der Aufsatz von Gensler et al. (2005) und die genannten Referenzen hierin.

erreicht 50%, der Aspekt „technisch nicht möglich“ wird mit 40,8% am seltensten von den vier Alternativen genannt (Gülpen 1996: 91).

Die vorliegende Arbeit hat sich deshalb zum Ziel gesetzt, aufzuzeigen, wie sich die ökonomische Programmevaluation in das Spektrum existierender betriebswirtschaftlicher und psychologischer Ansätze zur Bewertung von Trainingsmaßnahmen einfügt. Auch wenn die vorhandenen und erhobenen Daten im Unternehmen im Einzelfall für eine direkte Anwendung ökonometrischer Methoden unzureichend sind, entsteht zumindest ein Bewusstsein dafür, ob aufgrund der durchgeführten Evaluation auf einen kausalen Effekt der Trainingsmaßnahme auf den Unternehmenserfolg geschlossen werden kann. Zudem kann mit Kenntnis der Methoden der ökonomischen Programmevaluation eine Trainingsmaßnahme so durchgeführt werden, dass die Identifikation kausaler Effekte zumindest grundsätzlich möglich wäre.

Zunächst werden die vorwiegend an das *four level* Modell von Kirkpatrick aus dem Jahr 1959 angelehnten betriebswirtschaftlichen Evaluationsansätze dargestellt. Die Beschreibung der ökonomischen Sicht des fundamentalen Evaluationsproblems leitet zum ökonomischen Evaluationsansatz über. Der ökonomische Evaluationsansatz folgt der Idee, dass ein kausaler Effekt einer Trainingsmaßnahme durch die Differenz der potentiellen Ergebnisse von Maßnahmenteilnehmern und Nicht-Teilnehmern identifiziert werden kann (Lechner 1998). In der betriebswirtschaftlichen Literatur wird die Notwendigkeit der Identifikation kausaler Effekte bei der Beurteilung von Personalentwicklungsmaßnahmen zwar thematisiert, die Möglichkeit hierzu jedoch aufgrund der Komplexität des betrieblichen Umfeldes von vorneherein verneint (Becker 2005). Dies sollte jedoch in gleichem Maße für Feld der aktiven Arbeitsmarktpolitik gelten, die dagegen nicht von der Möglichkeit der Identifikation wahrer Maßnahmeneffekte Abstand nimmt (Heckman et al. 1999). Ergänzend wird schließlich das Kosten-Nutzen-Kalkül der Psychologen Schmidt, Hunter und Pearlman zur Bewertung betrieblicher Trainingsmaßnahmen erläutert.

Die Darstellung der einzelnen ökonometrischen Evaluationsansätze muss notwendigermaßen exemplarisch bleiben. Bei der Darstellung der ökonomischen Programmevaluation wird auf weiterführende Literatur verwiesen, die insbesondere die hier nicht dargestellten Verfahren diskutiert (Blundell und Costa Dias 2000, Schmidt 2001, Caliendo und Hujer im Erscheinen). Der Beitrag stellt primär Methoden zur Bewertung von Personalentwicklungsmaßnahmen vor. Während die im Weiteren vorgestellten betriebswirtschaftlichen Konzepte und das psychologische Konzept von Schmidt et al. speziell auf die Bewertung von Trainingsmaßnahmen zugeschnitten sind, lassen sie die ökonomischen Konzepte der

Programmevaluation auf andere unternehmerische Maßnahmen anwenden, deren Ergebnis kausal auf die jeweilige Maßnahme zurückgeführt werden soll. Beispiele finden sich hier namentlich im Marketing, wie sie zum Beispiel in Gensler et al. (2005) bei der Darstellung der Matching Methode als einer Methode der ökonomischen Programmevaluation vorgestellt werden.

2. Betriebswirtschaftliche Konzepte der Programmevaluation

In der betriebswirtschaftlichen Literatur gibt es eine Vielzahl an Konzepten zur Programmevaluation. Der Begriff „Programm“ bezeichnet in diesem Zusammenhang ein komplexes Bündel an Interventionen, die zielgeführt und aufeinander abgestimmt sind. Die Adjektive „zielgeführt“ und „aufeinander abgestimmt“ drücken hierbei aus, dass jeder Intervention bzw. jedem Bestandteil einer Intervention ein konkretes Ziel zugeordnet wird und genau festgelegt wird, wie die verschiedenen Interventionen ineinander greifen. Programmevaluation bezeichnet folglich die Evaluation von „zielgeführten Lernarrangements in Organisationen“ (Beywl 1999: 20). Die betriebswirtschaftlichen Konzepte der Programmevaluation können großteils den zielorientierten und den prozessorientierten Evaluationsansätzen zugeordnet werden.

Die zielorientierten Ansätze konzentrieren sich darauf, zu überprüfen, inwieweit die Trainingsziele erreicht worden sind. Dies wird auch oft als Ergebnisevaluation bezeichnet. Hierfür wird eine genaue Definition der Trainingsziele im Vorfeld vorausgesetzt. Die prozessorientierten Ansätze analysieren im Rahmen der Evaluation den gesamten Trainingsprozess: Im Gegensatz zur zielorientierten Ansätzen setzt hierbei die Evaluation nicht erst nach Trainingsende ein, sondern begleitet den gesamten Ablauf von der Trainingsplanung bis zur Durchführung. Dieses Vorgehen schließt in der Regel eine Ergebnisevaluation mit ein, was deutlich macht, dass es sich hierbei um einen breiter aufgestellten Ansatz handelt.

Nur wenige Ansätze zur Programmevaluation im Trainingsbereich können nicht diesen zwei Kategorien zugeordnet werden. Hierzu gehören unter anderem der Ansatz von Holton (1996) und der Ansatz von Kraiger, Ford und Salas (1993).

Tab.1: Ausgewählte zielorientierte Evaluationsansätze

Ansatz	Evaluationskriterien
Kirkpatrick 1998	Vier Level: 1. Reaction: Erfassen der subjektiven Wahrnehmung des Trainingsteilnehmers hinsichtlich des Trainings 2. Learning: Feststellen, inwieweit die Lernziele des Trainings erreicht worden sind 3. Behaviour: Anwendung des Gelernten im Job 4. Results: Ermittlung der durch das Training erzielten Ergebnisse
Phillips 1996	Die ersten vier Level sind inhaltlich im wesentlichen identisch mit denen von Kirkpatrick (s.o.). Phillips bezeichnet sie wie folgt: 1. Reaction and planned action , 2. Learning , 3. Applied Learning on the Job , 4. Business Results . Phillips hat Kirkpatricks Ansatz durch das fünfte Level: Return on Investment (ROI) ergänzt.

Tabelle 1 liefert eine Übersicht über zwei ausgewählte zielorientierte Evaluationsansätze. Der in der Literatur vorherrschende Ansatz wurde von Kirkpatrick im Jahr 1959 entwickelt und gehört zu den zielorientierten Ansätzen. Bei seinem Ansatz handelt es sich um eine Ergebnisevaluation auf den vier Leveln: Reaction, Learning, Behaviour und Results, wobei Kirkpatrick für die ersten drei Level Standards² formuliert hat, deren Einhaltung die Qualität und den Erfolg der Evaluation sichern sollen (Kirkpatrick 1979: 78). Die Bedeutung des ersten Levels liegt laut Kirkpatrick darin, die Gefühle aller Teilnehmer hinsichtlich des Programms zu erfahren und so dessen Akzeptanz zu bestimmen. Er argumentiert, dass die Teilnehmer ein Training mögen müssen, um den maximalen Nutzen daraus erzielen zu können (Kirkpatrick 1979: 81). Das zweite Level Learning ist eine notwendige Ergänzung, weil selbst positive Reaktionen hinsichtlich des Programms nicht garantieren, dass die Teilnehmer auch die Trainingsinhalte im ausreichenden Maße gelernt haben. Unter Learning versteht Kirkpatrick das Verstehen und Aufnehmen von Prinzipien, Fakten und Techniken (Kirkpatrick 1979: 82). Dieses Level unterscheidet sich zum ersten Level Reaction besonders dadurch, dass jetzt objektive Maße im Vordergrund stehen sollen und nicht mehr die subjektive Wahrnehmung der Teilnehmer. Die Einführung des dritten Levels Behaviour berücksichtigt die Tatsache, dass ein Unterschied darin besteht, Prinzipien und Techniken zu wissen bzw. zu können und sie tatsächlich bei der Arbeit ein- und umzusetzen. Dieses Level evaluiert folglich die durch das Programm tatsächlich resultierten Verhaltensänderungen bei der Arbeit (Kirkpatrick 1979: 86). Das letzte Level Results in Kirkpatricks Ansatz unterliegt einer besonderen Schwierigkeit. Denn bei dem Unterfangen den Einfluss eines Programms

² So lauten zum Beispiel die Standards für das Level **Transfererfolg (Behaviour)** wie folgt: 1) Systematische Abschätzung der Leistung am Arbeitsplatz in Form eines Vorher-Nachher-Vergleichs; 2) Leistungsabschätzung durch Beobachtung bzw. Befragung von einer oder mehreren der folgenden Gruppen: Teilnehmer des Programms, deren Vorgesetzte, deren Untergebene, deren Freunde oder andere Personen, denen die Leistung bekannt ist; 3) Erfolg der Leistungsabschätzung nach Programmende erst nach ca. drei Monaten; 4) Statistische Analyse der Ergebnisse vor und nach Programm; 5) Einsatz einer Kontrollgruppe zum Vergleich der Ergebnisse (Kirkpatrick 1979: 86).

auf Unternehmensergebnisse (wie z.B. Reduzierung der Fluktuation, Reduzierung der Anzahl an Beschwerden, erhöhte Kundenzufriedenheit etc.) zu messen, bleibt immer die Frage offen: Wieviel der Verbesserung ist durch das Training verursacht worden und wieviel durch andere Faktoren? Die Schwierigkeit liegt folglich darin, für Effekte und Faktoren (wie bspw. Zeit, Branchenentwicklung, Vorgeschichte der Teilnehmer etc.) zu kontrollieren. Kirkpatrick hat für diese Ebene keine Standards formuliert, sondern stellt exemplarisch einige Vorgehensweisen von Studien vor, die eine Evaluation von Unternehmensergebnissen durchgeführt haben (Kirkpatrick 1979: 89). Die Problematik bei den vorgestellten Fallstudien liegt darin, inwiefern selbst durch Vorher-Nachher Vergleiche und Kontrollgruppe wirklich nachgewiesen werden kann, ob die Verbesserung der Ergebnisse einzig und allein auf das Training zurückzuführen ist. Hier stellt sich die Frage, wie erstens alle weiteren Einflüsse identifiziert und wie in einem weiteren Schritt diese Einflussfaktoren kontrolliert werden können, da sicherlich ein Vorher-Nachher Vergleich nicht für alle Faktoren eine hinreichende Kontrolle bietet. Auf diese Problematik wird aber noch näher im nachfolgenden Abschnitt 3 eingegangen. Die Darstellung von Kirkpatricks Evaluationsansatz hat deutlich gemacht, dass der Durchführungsaufwand der Evaluation mit jedem Level zunimmt.

Phillips hat Kirkpatricks Ansatz durch ein fünftes Level ergänzt und Kirkpatricks Level Behaviour und Results umbenannt. Bei dem fünften Level handelt es sich um den Return on Investment (ROI)³. Der ROI setzt die erzielten Effekte eines Programms (gemessen in Geldeinheiten) in Verhältnis zu dem investierten Kapital (Kosten des Programms) (Werner und DeSimone 2005: 258). Für die Kalkulation des ROI müssen nach Programmende Daten erhoben, der Trainingseffekt isoliert, die erhobenen Daten in Geldeinheiten konvertiert werden und schließlich die Kosten des Programms ermittelt werden (Phillips 1996: 46). Diese Prozesskette veranschaulicht auch deutlich die Schwierigkeiten der Durchführung dieses Levels in der Praxis. Allein die Isolierung des Trainingseffekts und die Konvertierung des Trainingseffekts in Geldeinheiten sind sehr komplexe Probleme, die in der Regel keine einfache Lösung haben.⁴

Phillips argumentiert, dass es sinnvoll ist, den ROI nicht in dieselbe Ebene mit den Business Results aufzunehmen, sondern das 5. Level zu ergänzen. Denn es kann durchaus sein, dass die Business Results positiv sind (z.B. Anstieg der Verkaufszahlen nach einem

³ Der Return on Investment (zu Deutsch Kapitalrendite) ist eine Kennzahl, bei der das investierte Kapital im Verhältnis zum Gewinn gesetzt wird. $ROI = \text{Gewinn/Umsatz} * \text{Umsatz/investiertes Kapital}$ (in Worten: der ROI entspricht dem Umsatzerfolg multipliziert mit dem Umschlag des investierten Kapitals). Der ROI wird in der Regel bei der finanzwirtschaftlichen Bilanzanalyse eingesetzt (Spremann 1991: 200).

⁴ Es liegen Arbeitstabellen vor, welche die Kalkulation des ROI und die von Kosten-Nutzen-Analysen unterstützen (siehe hierzu Parry (2000)).

Verkaufstraining), der ROI desselben Trainings aber negativ ist, weil die Kosten für den erzielten Erfolg zu hoch sind (Stoel 2004: 47). Phillips hat die Level „Behaviour“ und „Results“ umbenannt in „Applied Learning on the Job“ und „Business Results“.⁵ Da die Level inhaltlich unverändert sind, gibt Phillips als Grund für die Umbenennung an, dass das Konzept durch die Umbenennung nicht nur zur Evaluation von Trainingsmaßnahmen einsetzbar ist, sondern auch zur Evaluation in den Bereichen Qualität, Technologie, Marketing und Strategieänderung (Stoel 2004: 48).

Tab.2: Ausgewählte prozessorientierte Ansätze

Ansatz	Evaluationskriterien
CIPP (Galvin 1983)	<ol style="list-style-type: none"> 1. Context: Ermittlung des relevanten Umfelds, Identifizierung eines bestimmten Bedarfs und nicht realisierten Möglichkeiten ⇒ Festlegung von Trainingszielen 2. Input: Identifikation von Strategien, die mit hoher Wahrscheinlichkeit das gewünschte Ergebnis erzielen 3. Process: Beurteilung der Durchführung des Programms 4. Product: Messung und Interpretation der erzielten Ergebnisse
Brinkerhoff 1987	<ol style="list-style-type: none"> 1. Goal Setting: Definition des Bedarfs und der Zielsetzung 2. Program Design: Auswahl einer Intervention für die definierte Zielsetzung und Analyse der Eignung dieser Intervention zur Zielerreichung 3. Program Implementation: Analyse der Durchführung hinsichtlich verschiedener Kriterien (reibungsloser und planmäßiger Ablauf, Zufriedenheit der Teilnehmer, Kosten etc.) 4. Immediate Outcomes: Erlernte Kenntnisse und Fähigkeiten 5. Intermediate or Usage Outcomes: Intensität der Anwendung der erlernten Kenntnisse und Fähigkeiten im Arbeitsalltag 6. Impacts and Worths: Nutzen auf Unternehmensebene, Identifizierung von relevanten Veränderungen
IPO (Bushnell 1990)	<ol style="list-style-type: none"> 1. Input: Ermittlung der System Performance Indikatoren (wie z.B. Qualifikation der Teilnehmer, Erfahrung des Trainers, Trainingsort und –ausstattung, Budget) hinsichtlich ihres Einflusses auf den Trainingserfolg 2. Process: Festlegung von Trainingszielen, Entwicklung von Kriterien für die Gestaltung, Auswahl von Trainingsmethoden Plan ⇒ Trainingskonzept ⇒ Umsetzung ⇒ Durchführung 3. Output: Reaktion der Teilnehmer, erlernte Kenntnisse und Fähigkeiten, erhöhte Leistung im Job ⇒ direkter Nutzen 4. Outcomes: Profit auf Unternehmensebene, Kundenzufriedenheit, Produktivität ⇒ i.d.R. nicht direkt feststellbar, erst ermittelbar durch Langzeitstudien
TVS (Fitz-Enz 1994)	<ol style="list-style-type: none"> 1. Situation: Ermittlung der Ausgangssituation (z.B. vorhandenes Leistungslevel) und Ableitung des Nutzens aus einer Veränderung dieser Ausgangssituation 2. Intervention: Formulierung und Konkretisierung der Problemstellung und Beschreibung der Lösung (bestehen u.U. andere Alternativen als Training zur Lösung) 3. Impact: Analyse der herbeigeführten Veränderungen 4. Value: Ermittlung des Nutzens (Anstieg in Qualität, Produktivität, Verkauf etc.) in Form von Geldeinheiten

Tabelle 2 liefert einen Überblick über ausgewählte prozessorientierte Ansätze. Die in der Tabelle dargestellten Ansätze unterscheiden sich zwar in der Benennung und teilweise auch in

⁵ Im abgedruckten Interview zwischen Kirkpatrick und Phillips bezeichnet Phillips „Applied Learning on the Job“ mit „Application“ und „Business Results“ mit „Business Impacts“ (Stoel 2004: 48).

der Anzahl der Evaluationsschritte, aber insgesamt ist ihre grundsätzliche Vorgehensweise analog: Im Gegensatz zu den zielorientierten Ansätzen integrieren alle dargestellten prozessorientierten Ansätze auch die Trainingsbedarfsanalyse, welche aber in jedem Ansatz anders bezeichnet wird (Context, Goal Setting, Process, Situation). Alle Ansätze befassen sich anschließend mit der Analyse der Trainingsgestaltung sowie der Trainingsdurchführung. Im Anschluss daran findet in allen Ansätzen eine Ergebnisevaluation statt. Der CIPP-Ansatz stellt diesen Grundgedanken der prozessorientierten Evaluation sehr anschaulich dar. Im ersten Schritt Context wird die Trainingsbedarfsanalyse durchgeführt. Anhand einer Analyse des Umfelds wird der Trainingsbedarf konkretisiert, anschließend werden auf dieser rationalen Grundlage Ziele für das Training festgelegt. Im zweiten Schritt Input werden Informationen ermittelt, die es ermöglichen festzustellen, welche Ressourcen benötigt werden, um die definierten Trainingsziele optimal umsetzen zu können. Die Ergebnisse dieser Phase können u.a. Budgetplanungen, Vorgehensweisen, Trainingsvorschläge etc. sein. Der dritte Schritt Process liefert ein detailliertes Feedback zur Durchführung des Programms. Hierfür sollen mögliche Faktoren für den Misserfolg überwacht werden und die Umsetzung an sich beobachtet und beschrieben werden. Im letzten Schritt Product wird gemessen, inwieweit die aufgestellten Trainingsziele erreicht wurden. Durch diese Gestaltung liefert der CIPP-Ansatz laut Galvin nicht nur Entscheidungshilfen für Überlegungen hinsichtlich der Trainingsfortführung durch die Ergebnisevaluation, sondern zusätzlich noch weitere für die Planung durch die Umfeldanalyse, für die Strukturierung durch die Inputanalyse sowie für die Durchführung durch die Prozessanalyse (Galvin 1983: 55). Die anderen aufgeführten prozessorientierten Ansätze unterscheiden sich hauptsächlich vom CIPP-Ansatz dadurch, dass sie die Ergebnisevaluation noch genauer differenzieren. Brinkerhoff unterscheidet beispielsweise drei Ergebnisarten: Erlernte Fähigkeiten und Kenntnisse (Intermediate Outcomes), Transfer der erlernten Kenntnisse und Fähigkeiten in den Arbeitsalltag (Intermediate or Usage Outcomes), Nutzen auf Unternehmensebene (Impacts and Worths). Der IPO-Ansatz und der TVS-Ansatz untergliedern den letzten Schritt Product des CIPP-Ansatzes jeweils in nur zwei Aspekte: direkter Nutzen (Output) und langfristiger Nutzen für das Unternehmen (Outcomes) beim IPO-Ansatz und herbeigeführte Veränderungen (Impact) und Nutzen (Value) beim TVS-Ansatz. Dies verdeutlicht, dass die prozessorientierten Ansätze in gewisser Weise auf den von Kirkpatrick eingeführten Grundgedanken: Ergebnisevaluation auf mehreren Leveln zurückgreifen und diesen durch Schritte im Vorfeld (Trainingsbedarfsanalyse, Trainingsgestaltung und –durchführung) ergänzen. Der Innovationsgehalt dieser Modelle erscheint allerdings begrenzt, wenn man sich vor Augen

führt, dass für die Umsetzung aller zielorientierten Ansätze eine genaue Formulierung der Trainingsziele im Vorfeld vorausgesetzt wird und durch das Level Reaction bei Kirkpatrick zumindest grobe Fehler bei der Durchführung aufgedeckt werden. Dies verdeutlicht, dass keine wirklich neuen Gesichtspunkte der Evaluation in den prozessorientierten Ansätzen enthalten sind. Folglich besteht der Beitrag der prozessorientierten Evaluationsansätze hauptsächlich darin, Aspekte, die bei zielorientierten Ansätzen vorausgesetzt worden sind, sinnvoll und vollständig in einen Ansatz zu integrieren. Für die Praxis sind diese Ansätze insofern von Bedeutung, da sie klarstellen, dass die Evaluation nicht erst nach Trainingsende beginnt, sondern den kompletten Prozess begleiten sollte.

Tab.3: Weitere ausgewählte Evaluationsansätze

Ansatz	Evaluationskriterien
Kraiger, Ford & Salas 1993	Identifizierung von drei Kategorien von Lernergebnissen: Kognitive Lernergebnisse, fähigkeits-basierte Lernergebnisse, affektive Lernergebnisse
Holton 1996	Ansatz mit drei Größen zur Ermittlung des Outcomes: Learning, Individual Performance und Organizational Results und Identifizierung von exogenen Faktoren, die in drei Gruppen: Motivation, Umwelt, Fähigkeit/Befähigung aufgeteilt sind. Schwerpunkt des Modells ist die Ermittlung eines Beziehungsgeflechts zwischen den Outcomegrößen und den exogenen Faktoren.

Tabelle 3 stellt zwei Evaluationsansätze dar, die weder den zielorientierten noch der prozessorientierten Ansätzen direkt zugeordnet werden können. Anders als die bisher vorgestellten Evaluationsansätze ist der Ansatz von Kraiger, Ford und Salas stark psychologisch fundiert. Die Autoren sehen eine Schwäche von Kirkpatricks Ansatz besonders darin, dass dieser nicht genau differenziert, welche Änderungen durch das Lernen erzielt werden sollen und welche Evaluationstechniken für diese Art des Lernens und der daraus resultierenden Verhaltensänderungen geeignet sind (Kraiger, Ford und Salas 1993: 311). Ausgehend von früheren Arbeiten haben Kraiger, Ford und Salas drei Kategorien von Lernen (kognitiv, fähigkeitsbasiert und affektiv) aufgestellt und diese jeweils weiter untergliedert. Die kognitiven Lernergebnisse werden weiter systematisiert in verbales Wissen, Wissensorganisation und kognitive Strategien (Kraiger, Ford und Salas 1993: 313-316). Die fähigkeitsbasierten Lernergebnisse werden in Ansammlung und Automatisierung unterteilt, wobei Ansammlung selbst das Resultat von zwei zusammenhängenden Prozessen ist: Verinnerlichung eines Verfahrens und Zusammensetzung von früher und neu Gelerntem zu einer gedanklichen Einheit (Kraiger, Ford und Salas 1993: 316-318). Die affektiven Lernergebnisse werden aufgefächert in Veränderung der Einstellungen und Veränderung der Motivation. Bei Veränderungen der Motivation spielen folgende drei Faktoren eine Rolle: Haltung hinsichtlich Motivation, Selbst-Wirksamkeit und Zielsetzung und –bindung (Kraiger, Ford und Salas 1993: 318-321). Anhand dieser Kategorisierung liefern Kraiger, Ford und

Salas Evaluationsmethoden für jeden Aspekt in den drei dargestellten Kategorien.⁶ Der Beitrag, den Kraiger, Ford und Salas hiermit für die Programmevaluation liefern, liegt hauptsächlich darin, dass sie verdeutlichen, dass Programme zu unterschiedlichen Arten von Lernergebnissen führen können und dementsprechend die jeweiligen Evaluationsinstrumente angepasst werden müssen. Sie konzentrieren sich auf die detaillierte Darstellung der Learning-Ebene, klammern aber weitere Ebenen der Programmevaluation (Transferebene und Ergebnis auf Unternehmensebene) völlig aus. Gerade für Unternehmen ist es nicht nur wichtig zu evaluieren, wieviel gelernt wurde, sondern wie viel des Gelernten tatsächlich bei der Arbeit umgesetzt wird. Hieraus lässt sich schlussfolgern, dass es sich eher um eine punktuelle Erweiterung des Modells von Kirkpatrick hinsichtlich seines zweiten Levels Learning handelt.

Holton kritisiert Kirkpatrick mit seinem Ansatz sehr stark. Er stützt sich bei seiner Kritik auf die Ergebnisse von Alliger und Janak (1989). Alliger und Janak haben dargelegt, dass Kirkpatricks Ansatz folgende drei Vermutungen nahelegt: 1) Jedes folgende Level ist informativer als das vorangegangene, 2) jedes Level wird durch das vorangegangene verursacht und 3) alle Korrelationen zwischen den Level sind positiv. Anhand einer empirischen Analyse stellen sie fest, dass alle drei Annahmen problematisch sind (Alliger und Janak 1989: 332 ff.). Holton nimmt genau diese Kritikpunkte wieder auf und kommt damit seinerseits zu der Aussage, dass es sich bei Kirkpatricks Ansatz nicht um ein Modell im wissenschaftlichen Sinn handelt, weil ihm hierfür notwendige Eigenschaften (wie u.a. kausale Beziehungen) fehlen (Holton 1996: 6f.). Nach Holton ist ein Evaluationsmodell notwendig, welches durch die Forschung fundiert ist. Mit Hilfe des Rückgriffs auf zahlreiche Forschungsergebnisse hat Holton ein Modell entwickelt, bei dem im Zentrum die drei Ergebnisgrößen: Lernerfolg (Learning), Individuelle Leistung (Individual Performance) und Unternehmenserfolg (Organizational Results) stehen.⁷ Weiterhin identifiziert Holton drei Kategorien von Einflussgrößen: Motivation, Umwelt und Fähigkeit/Befähigung, welche er noch weiter aufgliedert. Anhand von früheren Forschungsergebnissen stellt er kausale Zusammenhänge zwischen den einzelnen Faktoren und den Ergebnisgrößen sowie zwischen den Faktoren untereinander her. Holtons Beitrag für die Praxis der Programmevaluation liegt besonders darin, dass er die Komplexität der Evaluation noch einmal durch seine Arbeit verdeutlicht und so auch auf die Problematik hinweist, den Effekt des Trainings zu isolieren.

⁶ Für eine Übersicht über die von Kraiger et al. aufgestellten Evaluationsmethoden vgl. Kraiger et al. (1993: 323).

⁷ Die genannten Ergebnisgrößen ähneln denen von Kirkpatrick sehr stark. Trotzdem ist Holtons Ansatz nicht unter den zielorientierten Ansätzen aufgeführt worden, da im Zentrum des Modells nicht nur die Ergebnisgrößen, sondern auch die beeinflussenden Faktoren stehen.

Holton identifiziert eine Vielzahl an Faktoren (z.B. Transferklima, externe Ereignisse, Motivation zu Lernen, Motivation etw. anzuwenden, Fähigkeit des Einzelnen etc.), welche auf die Trainingsergebnisse wirken. Gerade aus dieser Perspektive steigt die Bedeutung von Untersuchungsdesigns, welche wirklich kausale Schlussfolgerungen zulassen. Dies wird im nächsten Abschnitt näher beleuchtet. Beim Vergleich der vorgestellten Ansätze lässt sich feststellen, dass die dargestellten zielorientierten und prozessorientierten Ansätze im Vergleich zu den dargestellten Ansätzen von Kraiger et al. und Holton eher pragmatisch als wissenschaftlich geprägt sind. Kirkpatrick selbst weist in seiner Antwort auf Holton (1996) darauf hin, dass sein Ansatz dazu beitragen soll, Praktikern einen Weg zu weisen, Evaluation systematisch anzugehen und dass dieser keine wissenschaftliche Ausarbeitung der Thematik Evaluation darstellt (Kirkpatrick 1996: 24). Durch die Beiträge von Holton und Kraiger et al. werden sicherlich die Evaluationsansätze aus wissenschaftlicher Sicht fundiert und erweitert. Hierdurch gewinnen sie aber auch an Komplexität, die der Praxis bei der Umsetzung immer mehr Expertise abverlangt.⁸

3. Die Ergebnisevaluation als gemeinsamer Bestandteil aller Konzepte

Im vorangegangenen Abschnitt wurde bereits angesprochen, dass alle vorgestellten Ansätze die Ergebnisevaluation integrieren. Aber gerade die durchdachte Durchführung einer Ergebnisevaluation ist keineswegs trivial: Neben der inhaltlichen Schwierigkeit dieser Evaluationsart (Operationalisierung der Ziele etc.) verlangt ein Untersuchungsdesign, das kausale Aussagen erlaubt, bereits vor Konzeption ein klares Wissen über die benötigten Daten und die jeweils unterstellten Annahmen. Bei der Ergebnisevaluation liegt die besondere Herausforderung darin nachweisen zu können, dass die beobachteten Effekte tatsächlich auf das Training zurückzuführen sind. Die entscheidende Frage hierbei lautet: Wie groß ist der mittlere Effekt für die Teilnehmer durch die Programmteilnahme? (Lechner 1998: 15). Es werden im folgenden Untersuchungsdesigns vorgestellt, die versuchen sicherzustellen, dass die beobachteten Auswirkungen bzw. Ergebnisse auch tatsächlich das Resultat des Programms sind und sich nicht zufällig ergeben haben bzw. durch andere Umstände verursacht wurden.

⁸ Ein tieferes Verständnis von Kraiger et al. Ansatz setzt fundierte Kenntnisse im psychologischen Bereich voraus. Holtons Modell zeigt so viele Einflussfaktoren auf, dass eine saubere Evaluation des Trainingseffekts in der Praxis fast unmöglich erscheint (vgl. hierzu Abbildung 2 in Holton (1996)).

Der ökonomische Evaluationsansatz folgt dabei dem *Modell potentieller Ergebnisse* (Roy 1951 und Rubin 1974), welches die Ergebnisse von Maßnahmenteilnehmern mit denen von Nicht-Teilnehmern vergleicht. Ausgangslage ist die Feststellung, dass der wahre Effekt eines Programms nur festgestellt werden kann, wenn der Effekt nach Teilnahme am Programm zum Zeitpunkt t für eine Person ermittelt und gleichzeitig der Effekt für dieselbe Person ohne Teilnahme am Programm zum selben Zeitpunkt t beobachtet werden könnte. Da eine Person entweder an einem Programm teilnehmen kann oder nicht, ist aber zumindest einer dieser Zustände nicht beobachtbar. Das jeweils unbeobachtbare Ergebnis wird in der Literatur auch als *Counterfactual Outcome* bezeichnet (Caliendo und Hujer, im Erscheinen). Dies stellt das zentrale Evaluationsproblem in der Programmevaluation dar. Als möglicher Lösungsansatz wird in der Literatur die Wahl einer geeigneten Kontrollgruppe diskutiert. Die Kontrollgruppe liefert die Effekte, die in Abwesenheit des Programms bzw. bei der Durchführung eines alternativen Programms eintreten. Die verschiedenen Ansätze in der ökonomischen Programmevaluation lassen sich zuvorderst danach unterscheiden, ob die Daten aus einem Experiment stammen, bei dem die Teilnehmer zufällig in eine Trainingsmaßnahme gelost wurden oder ob es sich um Beobachtungsdaten handelt, bei denen die Kontrollgruppe allein durch die Tatsache der Nicht-Teilnahme am Trainingsprogramm bestimmt wird, jedoch Selektionseffekte, also Effekte, die auf eine systematische Auswahl der Trainingsteilnehmer zurückgehen können, eine Rolle spielen (Blundell und Costa Dias 2000 und Caliendo und Hujer im Erscheinen). Nach der formalen Darstellung des Evaluationsproblems werden die verschiedenen Verfahren zur Analyse von Beobachtungsdaten exemplarisch anhand von zwei prominenten Schätzern dargestellt. Hierbei stellen die Annahmen, die es ermöglichen, die unbeobachtbaren Ergebnisse durch die Ergebnisse der jeweiligen Kontrollgruppe zu ersetzen, den entscheidenden Unterschied zwischen den einzelnen Schätzern dar. Die Ergebnisse der Evaluation basieren dementsprechend auf der vorgestellten Annahme und sind somit auch nur vertretbar, sofern die gewählte Annahme für das Untersuchungsziel plausibel erscheint. Der Idealfall der experimentellen Untersuchung, die das Evaluationsproblem faktisch beseitigt, wird im Anschluss kurz dargestellt und die Probleme bei der Anwendung in der betrieblichen Praxis diskutiert. Zuletzt wird das vielleicht prominenteste ökonometrische Schätzverfahren, die lineare Kleinst-Quadrate Regression auf den vorgestellten Modellrahmen zurückgeführt.

3.1. Formale Darstellung des Evaluationsproblems

Die im vorausgegangenen Abschnitt beschriebene gesuchte Zielgröße der Evaluation, der mittlere Effekt für die Teilnehmer durch die Programmteilnahme wird formal mit einem üblichen Evaluationsparameter, dem sogenannten *Mean Effect of Treatment on the Treated (METT)* beschrieben:

$$\begin{aligned} M_{X=k} &= E(Y_{1t} - Y_{0t} | X = k, D = 1) \\ &= E(Y_{1t} | X = k, D = 1) - E(Y_{0t} | X = k, D = 1) \end{aligned} \quad (1)$$

mit

Y_{1t} : Ergebnis (nach Programmende: t) bei Programmteilnahme

Y_{0t} : Ergebnis (nach Programmende: t) ohne Programmteilnahme

X : Merkmale, welche die Individuen charakterisieren und eine bestimmte Gruppenzugehörigkeit festlegen

D : wobei $D=1$: Teilnahme am Programm und $D = 0$: keine Teilnahme am Programm (Schmidt 2001: 15)

In diesem Modell wird der kausale Effekt als die Differenz zwischen den Ergebnissen ($Y_{1t} - Y_{0t}$) aufgefasst. Damit in diesem Zusammenhang rein formal von einem kausalen Effekt gesprochen werden kann, muss die Annahme getroffen werden, dass die individuellen Ergebnisse unabhängig davon sind, wie sich die anderen Individuen verhalten (*Stable unit treatment value-Annahme, SUTVA*) (Schmidt 2001: 14). Diese Annahme ist für Trainingsprogramme in Unternehmen relativ unproblematisch, dagegen kann sie bei groß angelegten Arbeitsmarktprogrammen durchaus kritisch werden. Hier kann über den Marktmechanismus das Programmresultat des Einzelnen (z.B. Stundenlohn in einem umgeschulten Beruf) durch die Teilnahmeentscheidungen anderer beeinflusst werden. Denn je mehr Leute umgeschult werden, desto niedriger wird ab einer bestimmten Anzahl der Umgeschulten der Stundenlohn des Einzelnen ausfallen.

Die Formel für den METT verdeutlicht noch einmal das bereits angesprochene Evaluationsproblem. Der Erwartungswert für das Ergebnis ohne Programmteilnahme unter der Bedingung, dass das Individuum am Programm teilgenommen hat und das Merkmal $X=k$ aufweist ($E(Y_{0t} | X = k, D = 1)$) ist nicht beobachtbar und kann somit auch nicht mit einfachen Methoden geschätzt werden. Das Ergebnis eines Teilnehmers ohne Programmteilnahme nach Programmende ist das bereits angesprochene Counterfactual Outcome für den Teilnehmer. Es muss dementsprechend eine Annahme getroffen werden, die es ermöglicht, diesen Erwartungswert $E(Y_{0t} | X = k, D = 1)$ zu schätzen. Die folgenden Abschnitte stellen Lösungen für das Evaluationsproblem in Form von zwei prominenten Schätzern für Beobachtungsdaten

und der Durchführung von experimentellen Untersuchungen vor. Für die Darstellung weiterer Verfahren wird auf einschlägige Übersichtsartikel von Lechner (1998), Schmidt (2002), Caliendo und Hujer (im Erscheinen) verwiesen. Eine Darstellung der Matching Methode, des Difference-in-Difference und des Conditional Difference-in-Difference Schätzers findet sich in Caliendo und Kopeing (2005) und mit speziellem Bezug auf die betriebswirtschaftliche Anwendung insbesondere in Gensler et al. (2005).

3.1.1. Cross-section Schätzer

Der Cross-Section (CS) Schätzer ist ein häufig angewandter Schätzer, weil er keine Längsstudie voraussetzt. Die Annahme dieses Schätzers lautet, dass das Counterfactual Outcome für Teilnehmer zum Zeitpunkt t im Mittel identisch ist mit dem Ergebnis der Nicht-Teilnehmer zum Zeitpunkt t. Formal bedeutet dies:

$$E(Y_{0t}|X = k, D = 1) = E(Y_{0t}|X = k, D = 0). \quad (\text{Ann.1})$$

Dies kann über die Abweichung zum METT hergeleitet werden:

$$\begin{aligned} B^{CS} &= M^{CS} - M \\ &= E(Y_{1t}|X = k, D = 1) - E(Y_{0t}|X = k, D = 0) - \left[E(Y_{1t}|X = k, D = 1) - E(Y_{0t}|X = k, D = 1) \right] \\ &= E(Y_{0t}|X = k, D = 1) - E(Y_{0t}|X = k, D = 0). \end{aligned} \quad (2)$$

Setzt man die letzte Zeile gleich Null, erhält man die oben aufgestellte Annahme (Ann.1).

Das resultierende Problem ($E(Y_{1t}|X = k, D = 1) - E(Y_{0t}|X = k, D = 0)$) kann leicht geschätzt werden, weil beide Erwartungswerte beobachtbar sind und folglich durch das jeweilige arithmetischen Mittel geschätzt werden können.

Der Schätzer für dieses Problem lautet dementsprechend:

$$\hat{M}_{X=k}^{CS} = \frac{1}{N_{1,X=k}} \sum_{i \in I_1, X=k} Y_{1t_i} - \frac{1}{N_{0,X=k}} \sum_{j \in I_0, X=k} Y_{0t_j}. \quad (3)$$

Dieser Schätzer ist anfällig für systematische Verzerrungen, wenn der Selektionsprozess für das Training nicht exogen läuft. Formal ausgedrückt bedeutet dies, dass das Ergebnis ohne Programmteilnahme, aber nach Programmende und die Teilnahmeentscheidung für ein gegebenes Merkmal X statistisch unabhängig sein müssen. Das Problem liegt hierbei insbesondere bei den nicht beobachtbaren Faktoren (d.h. für Faktoren, für die in der Analyse nicht kontrolliert werden kann). Ein typisches Beispiel für einen solchen nicht beobachtbaren Faktor ist die Motivation der Trainingsteilnehmer. Die Motivation hat in der Regel sowohl

einen Einfluss auf die Teilnahmeentscheidung als auch auf das resultierende Ergebnis. In diesem Szenario würde die Gruppe der Teilnehmer (mit $D=1$) aus motivierteren Individuen bestehen, welche in der Regel auch in Abwesenheit des Programms ein höheres Ergebnis erzielen als die Kontrollgruppe, die aus weniger motivierten Individuen besteht. Dies würde zu einer positiven Korrelation zwischen der Teilnahmeentscheidung $D=1$ und dem Ergebnis nach Programmende ohne Programmteilnahme (Y_{0t}) führen. Hierdurch wird die oben aufgestellte Annahme (Ann.1) verletzt (Schmidt 2001: 30).

3.1.2 Before-After Schätzer

Der Before-After (BA) Schätzer impliziert, dass es sich um eine Langzeitstudie handelt, bei der an zwei Messzeitpunkten Daten von einem anfangs festgelegten Teilnehmerkreis des Programms erhoben werden. Der Pretest findet vor Trainingsbeginn statt, der Posttest wird nach Trainingsende durchgeführt. Gerade in der Praxis kann es in manchen Fällen schwierig werden, eine Gruppe vergleichbarer Personen zu finden, welche die Effekte für Nicht-Teilnehmer wiedergibt. Die Idee des BA Schätzers ist nun, dass jeder Programmteilnehmer vor Programmbeginn ein Nicht-Teilnehmer ist und so der kausale Effekt eines Programms durch den Vergleich von den Ergebnissen eines Teilnehmers nach Programmende mit den Ergebnissen des Teilnehmers vor Programmbeginn ermittelt werden kann.

Diesem Vorgehen liegt die Annahme zugrunde, dass über die gesamte Population von Programmteilnehmern im Mittel die Ergebnisse zum Zeitpunkt vor Programmbeginn, t' , identisch sind mit den Ergebnissen im Mittel, die diese Gruppe zum Zeitpunkt t (nach Programmende) erfahren hätte, wenn sie nicht am Programm teilgenommen hätte.

Formal bedeutet dies:

$$E(Y_{0t} | X = k, D = 1) = E(Y_{0t'} | X = k, D = 1) \quad (\text{Ann.2})$$

(Schmidt 2001:28).

Diese Annahme (Ann.2) kann hergeleitet werden, indem die Abweichung (Bias) zu dem METT analysiert wird:

$$\begin{aligned} B^{BA} &= E(Y_{1t} | X = k, D = 1) - E(Y_{0t} | X = k, D = 1) - \left[E(Y_{1t} | X = k, D = 1) - E(Y_{0t'} | X = k, D = 1) \right] \quad (4) \\ &= E(Y_{0t'} | X = k, D = 1) - E(Y_{0t} | X = k, D = 1) \end{aligned}$$

⁹ Der Index t' steht für den Zeitpunkt vor Programmteilnahme.

Damit keine systematische Abweichung zum METT vorliegt, muss die Abweichung gleich Null sein. Wird die Gleichung (4) gleich Null gesetzt, erhält man die oben definierte Annahme (Ann.2).

Durch die Annahme kann der BA Schätzer leicht ermittelt werden. Es muss nun folgendes Problem geschätzt werden:

$$M^{BA}_{X=k} = E(Y_{1t}|X = k, D = 1) - E(Y_{0t}|X = k, D = 1). \quad (5)$$

Diese Formulierung des Problems basiert auf der oben getroffenen Annahme.

Der Schätzer lautet folglich:

$$\hat{M}^{BA}_{X=k} = \frac{1}{N_{1,X=k}} \sum_{i \in I, X=k} Y_{1ti} - \frac{1}{N_{0,X=k}} \sum_{i \in I, X=k} Y_{0ti} = \frac{1}{N_{1,X=k}} \sum_{i \in I, X=k} (Y_{1ti} - Y_{0ti}). \quad (6)$$

Die oben definierte Annahme (Ann.2) setzt aber voraus, dass ein Individuum sich vom Zeitpunkt vor Programmbeginn (t') zum Zeitpunkt nach Programmende (t) nicht weiterentwickelt. Dies ist besonders problematisch bei langen Programmen, bei denen die zwei Zeitpunkte weit auseinander liegen. Hier kann es vorkommen, dass der BA Schätzer die Effekte des Programms überschätzt, weil Programmteilnehmer und Nicht-Teilnehmer eine systematisch unterschiedliche Entwicklung bezüglich der gemessenen Programmergebnisse durchlaufen. So bleibt die Frage, ob das Individuum nicht auch ohne Programmteilnahme, sondern einfach durch die Ereignisse in der Zwischenzeit zu denselben Ergebnissen gekommen wäre, unbeantwortet. Zum Beispiel kann bei einer Excel-Schulung, die aus mehreren, über einen längeren Zeitraum verteilten Trainingseinheiten besteht, hinterfragt werden, ob die Person sich nicht auch durch den Gebrauch des Programms bei der täglichen Arbeit zum Zeitpunkt t systematisch gegenüber Zeitpunkt t' verbessert hätte. Der BA Schätzer würde in diesem Fall den tatsächlichen Effekt des Trainingprogramms überschätzen (Schmidt 2001: 29). Dem zentralen Problem des BA Schätzers kann begegnet werden, wenn das Vorher-Nachher-Untersuchungsdesign noch zusätzlich durch eine Kontrollgruppe ergänzt wird. Bei diesem Untersuchungsdesign wird folglich verglichen, wie sich die Ergebnisvariable der Kontrollgruppe (Nicht-Programmteilnehmer) im Vergleich zur Teilnehmergruppe (Programmteilnehmer) verändert. Dieses Design erfordert aber sowohl die Festlegung einer Kontrollgruppe als auch Langzeitstudien für beide Gruppen. Der sogenannte Difference-in-Difference Schätzer berücksichtigt im Gegensatz zum BA Schätzer also einen möglichen Trendeffekt, ist jedoch auch mit der starken Annahme verbunden, dass dieser für Teilnehmer und Nicht-Teilnehmer im gleichen Maße vorliegt (Schmidt, 2001 und Gensler et al. 2005).

3.1.3 Experimentelle Untersuchungen

Ein weiteres Konzept für die Lösung des Evaluationsproblems ist die zufällige Zuordnung von Individuen in Teilnehmergruppe und Kontrollgruppe. Ein Zufallsgenerator legt hierbei fest, ob Teilnehmer, die an einem Programm freiwillig teilnehmen wollen, tatsächlich auch teilnehmen können. Es gibt demnach eine zusätzliche Variable R_i für jedes Individuum, welche den durch den Zufallsgenerator generierten Status angibt ($R=1$: Individuum zufällig in Teilnehmergruppe gewählt und $R=0$: Individuum zufällig in Kontrollgruppe gewählt). Das Counterfactual Outcome kann durch folgende die Annahme 1 identifiziert werden:

$$E(Y_{0t} | X = k, D = 1) = E(Y_{0t} | X = k, D = 1, R = 0) \quad (\text{Ann.3})$$

(Schmidt 2001: 21).

Hinter diesem Ergebnis für das Counterfactual Outcome verbergen sich noch weitere Annahmen, welche besagen, dass die Einführung des Zufallsgenerators nicht den gewöhnlichen Ablauf (course of affairs) verändert. Formal bedeutet dies, dass sowohl

$$E(Y_{1t} | X = k, D = 1) = E(Y_{1t} | X = k, D = 1, R = 1) \quad (\text{Ann.3a})$$

als auch

$$E(Y_{0t} | X = k, D = 0) = E(Y_{0t} | X = k, D = 0, R = 0) \quad (\text{Ann.3b})$$

gelten muss (Schmidt 2001: 21 und Heckman, LaLonde und Smith 1999: 1899). Die für das Identifikationsproblem entscheidende Annahme (Ann.3) kann hergeleitet werden, indem man die Abweichung vom METT analysiert und hierbei berücksichtigt, dass Annahme (Ann.3a) gelten muss:

$$B^R = E(Y_{1t} | X = k, D = 1, R = 1) - E(Y_{0t} | X = k, D = 1, R = 0) - [E(Y_{1t} | X = k, D = 1) - E(Y_{0t} | X = k, D = 1)] \quad (7)$$

$$B^R = E(Y_{0t} | X = k, D = 1) - E(Y_{0t} | X = k, D = 1, R = 0).$$

Setzt man die Abweichung gleich Null, d.h. versucht man den systematischen Fehler zu beheben, dann erhält man die oben beschriebene Annahme (Ann. 4), die es ermöglicht, das Counterfactual Outcome durch ein beobachtbares Ergebnis zu ersetzen und so mit Hilfe des Mittelwerts einfach zu schätzen.

Ausgehend von der oben hergeleiteten Identifikationsannahme lautet der Schätzer für den METT folglich:

$$\hat{M}_{X=k}^R = \frac{1}{N_{R=1, X=k}} \sum_{i \in I_1, X=k, R=1} Y_{1ti} - \frac{1}{N_{R=0, X=k}} \sum_{j \in I_0, X=k, R=0} Y_{0tj} \quad (8)$$

(Schmidt 2001: 22).

Obwohl der Schätzer bezogen auf das Evaluationsproblem geradezu ideal wirkt, dürfte die praktische Durchführung bei der Bewertung von Trainingsmaßnahmen in den meisten Unternehmen kaum möglich sein. Das Adressieren der Trainingsmaßnahme an bestimmte Mitarbeitergruppen ist Teil eines professionellen Personalmanagements und wird schwerlich hinter eine klare Erfolgsmessung zurückgestellt werden. Neben den hohen Kosten, die mit solchen Experimenten verbunden sind, existieren weitere Nachteile die im Kontext der Literatur zu sogenannten „sozialen Experimenten“ diskutiert werden (Burtless 1995) und auf Untersuchungsdesigns im Unternehmenskontext übertragen werden können.

3.1.4. Bedeutung der Konditionierung auf ein gemeinsames Merkmal

Alle drei Annahmen sind bereits für eine Konditionierung auf ein gemeinsames Merkmal X in den drei vorangegangenen Abschnitten vorgestellt worden. Heckmann et al. stellen in ihrer Darstellung die drei Schätzer ohne Konditionierung auf ein gemeinsames Merkmal X dar (Heckman, LaLonde und Smith 1999: 1891-1897). Der Ausdruck „Konditionierung auf ein Merkmal X“ bedeutet in diesem Zusammenhang, dass der Schätzer für eine Untergruppe innerhalb der Stichprobe ermittelt wird. Innerhalb dieser Gruppe weisen alle Individuen das Merkmal X auf. Heckman stellt klar, dass eine Konditionierung die Einhaltung der dargestellten Annahmen (Ann.1, Ann.2, Ann.3) wahrscheinlicher macht, aber nicht garantiert. Wenn nämlich die Verteilung des Merkmals X für Teilnehmergruppe und Kontrollgruppe verschieden ist, kann eine Konditionierung auf das Merkmal X dazu führen, dass der systematische Unterschied behoben wird. Wird dieser aber nicht durch das Merkmal X, sondern durch unbeobachtete Faktoren verursacht, kann eine Konditionierung unter bestimmten Umständen dazu führen, dass der Effekt des systematischen Unterschieds noch verstärkt und nicht eliminiert wird¹⁰ (Heckman, LaLonde und Smith 1999: 1898).

3.2. Lineare Regression

¹⁰ Nachweis: Falls $|E(Y_{0t}|D=1) - E(Y_{0t}|D=0)| = M$, dann muss nicht unter allen Umständen auch gelten $|E(Y_{0t}|D=1, X=k) - E(Y_{0t}|D=0, X=k)| < M$.

Die Regressionsanalyse ist ein prominentes ökonometrisches Verfahren, das bei vielen wissenschaftlichen Untersuchungen eingesetzt wird, wobei die Methode der kleinsten Quadrate die am häufigsten eingesetzte Technik zur Schätzung der Regressionskoeffizienten ist. Aufgrund der großen Bedeutung dieser Methode wird im folgendem dargelegt, dass die lineare Regression (mit einer unabhängigen Variablen) auf eine leicht abgewandelte Form einer bereits vorgestellten Annahme zurückzuführen ist.

Ausgangspunkt für die Anwendung der linearen Regression ist ein Datensatz der mindestens die Ergebnisse für Teilnehmer sowie für Nicht-Teilnehmer liefert. Das Ergebnis für jeden Teilnehmer i kann folgendermaßen dargestellt werden:

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i} = Y_{0i} + (Y_{1i} - Y_{0i}) D_i = Y_{0i} + \Delta_i D_i . \quad (9)$$

Um dieses Problem mit dem Standardmodell der linearen Regression lösen zu können, müssen einige Annahmen getroffen werden: 1) Der Programmeffekt (Δ) ist für alle identisch und 2) das Ergebnis ohne Programmteilnahme (Y_0) stimmt für alle überein. Das mit der Regression zu schätzende Problem lautet dann:

$$Y_i = Y_0 + \Delta D_i . \quad (10)$$

Dies kann durch folgende Regressionsgleichung geschätzt werden:

$$Y_i = \alpha + \beta D_i + \varepsilon_i \quad (11)$$

(Cobb-Clark, Crossley 2003: 498).

Das Ergebnis ohne Programmteilnahme (Y_0 bzw. im Regressionsmodell: α) wird in diesem Fall als konstant vorausgesetzt. Es ist also unabhängig davon, ob es sich bei der Person um einen Teilnehmer handelt oder nicht. Dies ist äquivalent zu folgender Formulierung:

$$E(Y_{0i} | D_i = 1) = E(Y_{0i} | D_i = 0) .^{11} \quad (\text{Ann.4})$$

In leicht abgewandelter Form ist diese Annahme bereits beim CS Schätzer vorgestellt worden. Der Unterschied liegt nur in der weiteren Konditionierung auf ein gemeinsames Merkmal X (vgl. Ann. 1). Bei der Annahme (Ann. 4) handelt es sich folglich um die Annahme des CS Schätzers ohne Konditionierung auf ein gemeinsames Merkmal $X=k$. Folglich geht dieses Basisregressionsmodell von der Annahme aus, dass für alle Beobachteten (egal ob Teilnehmer oder Nicht-Teilnehmer) das Ergebnis vor Programmbeginn identisch ist. Weiterhin geht das Modell davon aus, dass der Programmeffekt Δ für alle konstant ist. Beide Annahmen sind bei einer Anwendung kritisch zu hinterfragen.

¹¹ Exakt formuliert ist die Annahme für einen unverzerrten Schätzer für β bei der linearen Regression:

$E(\varepsilon_i | D_i = 0) = E(\varepsilon_i | D_i = 1) = E(\varepsilon_i) = 0$. Dies ist aber äquivalent zur der Formulierung der Annahme (Ann.4)

(Cobb-Clark, Crossley 2003: 498).

Da es sich bei dem oben aufgestellten Regressionsmodell um das Basismodell der linearen Regression handelt, ist der Schätzer für β noch einfach zu ermitteln. Der Kleinste-Quadrate-Schätzer für β lautet folglich:

$$\beta = \frac{\sum (D_i - \bar{D}) Y_i}{\sum (D_i - \bar{D})^2} \quad (12)$$

Mit einigen zusätzlichen Umformungen kann gezeigt werden, das gilt:

$$\beta = \frac{1}{n\bar{D}} \sum_{\text{treated}} Y_i - \frac{1}{n(1-\bar{D})} \sum_{\text{untreated}} Y_j \quad (13)$$

Hierbei steht n für die Anzahl aller Personen in der Stichprobe und $n\bar{D}$ für die Anzahl der Teilnehmer¹² bzw. $n(1-\bar{D})$ für die Anzahl der Nicht-Teilnehmer¹³. Insgesamt erhält man also die Differenz vom Mittelwert der Teilnehmer und dem Mittelwert der Nicht-Teilnehmer (Cobb-Clark, Crossley 2003: 499). Da der Kleinste-Quadrate-Schätzer bei Gültigkeit seiner Annahmen der beste lineare, unverzerrte Schätzer ist, ist somit diese Gewichtung optimal (Cobb-Clark, Crossley 2003: 500).

4. Kosten-Nutzen-Analyse als Möglichkeit der Programmevaluation

Die Kosten-Nutzen-Analyse der Psychologen Schmidt, Hunter und Pearlman (1982: 333ff) zur Bewertung des Erfolgs von Trainingsmaßnahmen hat sich im psychologischen Bereich durchgesetzt und wurde bereits in einigen amerikanischen Studien angewandt, im deutschsprachigen Raum liegen dagegen nur wenige Studien, die diese Methode nutzen, vor (Gülpen 1996: 39). Die Kosten-Nutzen-Analyse von Schmidt et al. stellt ein mögliches Instrument zur Evaluation auf der Unternehmensebene dar. Im Rahmen dieser Analyse wird der Nettonutzen eines Programms ermittelt, indem von dem Gesamtnutzen des Programms ausgedrückt in Geldeinheiten (auch als Bruttonutzen bezeichnet) die aufgetragenen Kosten abgezogen werden. Der Schwerpunkt bei der Analyse von Schmidt et al. liegt auf der Ermittlung der durch das Programm erzielten Leistungsveränderung bei den Teilnehmern, also auf dem durch das Programm erzielte Output (Schmidt et al. 1982: 341). Es

¹² Hierfür muss man sich klar machen, dass gilt: $n\bar{D} = n \frac{1}{n} \sum D_i = \sum D_i$. Da es sich bei D_i um eine

Binärvariable mit der folgenden Kodierung 1 Teilnehmer und 0 Nicht-Teilnehmer handelt, ergibt die Summe der D_i über alle i genau die Anzahl an Teilnehmern.

¹³ Es gilt: $n(1-\bar{D}) = n - n\bar{D}$. Folglich wird von der Summe aller die Anzahl der Teilnehmer abgezogen und man erhält die Anzahl der Nicht-Teilnehmer.

handelt sich folglich um eine outputbezogene Analyse­methode. Die Formel zur Berechnung des Nettonutzens lautet (Schmidt et al. 1982: 335):

$$\Delta U = T N d_t SD_y - N C, \quad (14)$$

wobei:

ΔU = Nettonutzen des Trainingsprogramms in Geldeinheiten

T = Dauer des Trainingseffekts auf die Arbeitsleistung in Jahren

N = Anzahl der Teilnehmer

d_t = wahrer Unterschied in der Berufsleistung zwischen Trainierten und Untrainierten ausgedrückt in Standardabweichung

SD_y = Standardabweichung der Berufsleistung der untrainierten Gruppe in Geldeinheiten

C = Kosten des Trainings pro Teilnehmer

Der zweite Teil der Formel ($N \cdot C$) ist einfach zu ermitteln, sowohl die Kosten als auch die Anzahl der Teilnehmer sind leicht zu quantifizieren. Um die Kosten eines Programms systematisch zu erfassen, gibt es Übersichten, welche dieses Vorhaben wesentlich erleichtern.¹⁴

Wesentlich problematischer ist der erste Teil der Formel ($T \cdot N \cdot d_t \cdot SD_y$). Die Ermittlung des wahren Unterschieds in der Berufsleistung zwischen Trainierten und Untrainierten ausgedrückt in Standardabweichung (d_t) und Standardabweichung der Berufsleistung der untrainierten Gruppe in Geldeinheiten (SD_y) sind komplexer und nicht selbsterklärend. Um diesen Teil der Formel nachzuvollziehen, muss man sich als erstes vor Augen führen, dass der Bruttonutzen des Programms in Geldeinheiten ermittelt wird. Würden folglich für alle Trainingsprogramme in Unternehmen Daten vorliegen, welche den Anstieg der Berufsleistung der Trainierten bezogen auf die Untrainierten direkt in Geldeinheiten messen würden, würde sich die Formel vereinfachen: Der durchschnittliche Unterschied in der Berufsleistung in Geldeinheiten pro Teilnehmer ($\bar{Y}_1 - \bar{Y}_0$ jeweils in Geldeinheiten) könnte direkt mit dem Zeiteffekt (T) und der Anzahl an Teilnehmer (N) multipliziert werden.¹⁵ Da dieses Szenario aber eher eine seltene Ausnahme als die Regel bei betrieblichen Trainingsprogrammen ist, haben Schmidt et al. die Formel so angepasst, dass sie auch für Trainingsprogramme, bei denen die durchschnittliche Veränderung der Berufsleistung nicht direkt in Geldeinheiten messbar ist, anwendbar ist.

¹⁴ Siehe hierzu beispielsweise Gülpen (1996: 45).

¹⁵ Ein Beispiel für ein solches Szenario ist ein Verkaufstraining für ein bestimmtes Produkt. Nach Abschluß des Trainings könnte gemessen werden, wieviel mehr dieser Produkte die Teilnehmer durchschnittlich verkaufen als die Nicht-Teilnehmer. Das ermittelte Ergebnis könnte dann direkt in Geldeinheiten umgerechnet werden.

Die Grundidee der Nutzenformel von Schmidt et al. liegt darin, den Effekt eines Trainingsprogramms in Einheiten Standardabweichungen auszudrücken und in einem weiteren Schritt zu ermitteln, wieviel eine Steigerung der Arbeitsleistung um genau eine Standardabweichung in Geldeinheiten wert ist. Der Ausgangspunkt ist somit, die erzielte Leistungssteigerung im Beruf durch das Training in Einheiten Standardabweichungen auszudrücken. Aufgrund dessen wird die mit Hilfe eines Instruments gemessene durchschnittliche Veränderung der Berufsleistung durch die Standardabweichung der beobachteten Ergebnisse der Nicht-Teilnehmer¹⁶ und den Reliabilitätskoeffizienten¹⁷ des jeweiligen Instruments geteilt. Hierdurch erhält man die wahre Leistungsveränderung in Einheiten Standardabweichungen (d_t). In einem weiteren Schritt wird ermittelt, wie stark die Arbeitsleistung in Geldeinheiten für die zu untersuchende Gruppe an Arbeitnehmern variiert, einfacher ausgedrückt, wie groß ist der Unterschied der Arbeitsleistung in Geldeinheiten zwischen einem durchschnittlichen Arbeitnehmer der zu untersuchenden Gruppe und einem sehr guten Arbeitnehmer dieser Gruppe. Da man genau eine Einheit Standardabweichung ermitteln möchte, ist die exakte Definition eines „sehr guten Arbeitnehmers“ ein Arbeitnehmer auf dem 85ten Prozentpunkt (d.h. nur ca. 15% der Arbeitnehmer erbringen eine bessere Leistung, dagegen aber ca. 85% eine schlechtere Leistung). Dieses Vorgehen setzt voraus, dass die Arbeitsleistung in Geldeinheiten normalverteilt ist. Die Schlussfolgerung aus dieser Überlegung ist folglich: Wenn die Berufsleistung des Arbeitnehmers um eine Einheit Standardabweichung gesteigert werden kann, dann steigt die Berufsleistung in Geldeinheiten um den ermittelten Unterschied zwischen dem durchschnittlichen Arbeitnehmer (auf 50ten Prozentpunkt) und dem sehr guten Arbeitnehmer (auf 85ten Prozentpunkt). Der ermittelte wahre Trainingseffekt (d_t), ausgedrückt in Standardeinheiten, wird anschließend mit der Standardabweichung der Berufsleistung in Geldeinheiten (SD_y) multipliziert. Als Ergebnis erhält man den gewünschten Bruttonutzen (pro Jahr und Teilnehmer) in Geldeinheiten. Durch die Multiplikation des Bruttonutzens pro Jahr und Teilnehmer mit der Teilnehmeranzahl (N) und der Dauer des Trainingseffekts (T) wird der gesamte Bruttonutzen des Trainings ermittelt.

¹⁶ Schmidt et al. argumentieren, dass die SD der Nicht-Teilnehmer und nicht die Standardabweichung der Teilnehmer gewählt werden sollte, weil sich unter Umständen die Standardabweichung der Teilnehmer durch die Trainingsteilnahme verändert hat (Schmidt et al. 1982: 336).

¹⁷ Die Reliabilität ist ein Maß, dass die Verlässlichkeit bzw. Konsistenz eines Instruments beurteilt. Der Reliabilitätskoeffizient ist der quadrierte Korrelationskoeffizient zwischen beobachteten und wahren Werten eines Test. Er gibt den Anteil der Varianz der wahren Werte an der Varianz der beobachteten Werte an: $\rho_{XT}^2 = \frac{\sigma_T^2}{\sigma_X^2}$.

Da die wahren Werte nicht beobachtbar sind, gibt es verschiedene Verfahren, die Reliabilität zu ermitteln (Paralleltest Verfahren, Test-Retest Verfahren).

Im vorangegangenen Absatz ist deutlich geworden, dass die Ermittlung des wahren Unterschieds in der Berufsleistung zwischen Trainierten und Untrainierten ausgedrückt in Standardabweichung (d_i) und Standardabweichung der Berufsleistung der untrainierten Gruppe in Geldeinheiten (SD_y) nicht trivial, aber gleichzeitig auch zentral für die Ermittlung des Nutzens sind. Bei der Darstellung der Intuition für SD_y ist bereits ein möglicher Ansatz für die Schätzung von SD_y vorgestellt worden. Wie bereits dargelegt, entspricht eine Einheit Standardabweichung der Differenz der Berufsleistung in Geldeinheiten von einem durchschnittlichen Angestellten (50ter Prozentpunkt) und dem Angestellten auf dem 85ten Prozentpunkt. Aufgrund der Symmetrie der Normalverteilung muss die Differenz zwischen dem 15ten Prozentpunkt (also einem schlechten Angestellten) und dem durchschnittlichen Angestellten identisch sein. Die Berufsleistung in Geldeinheiten wird hierbei durch die hergestellten Produkte und geleisteten Dienstleistungen operationalisiert. Eine erfahrene Führungskraft soll nun die Berufsleistung in Geldeinheiten für Angestellte an diesen drei Punkten schätzen. Als Orientierungshilfe für die Schätzung können die Kosten erwogen werden, die entstehen würden, wenn man die vom Angestellten erbrachten Leistungen outsourcen würde. Es ist hierbei zentral, dass den Befragten der Unterschied zwischen Berufsleistung in Geldeinheiten und Gehalt klar ist. Im Normalfall sollte die Berufsleistung in Geldeinheiten höher als das Gehalt sein, damit ein Unternehmen langfristig am Markt bestehen kann. Diese Schätzmethode verlangt einen sorgfältig entwickelten Fragebogen, der Experten im Unternehmen vorgelegt werden kann.

Das gerade beschriebene Verfahren ist in seiner Durchführung sehr aufwendig und es ist fraglich, ob in jedem Unternehmen Experten für eine derart schwierige Schätzung zur Verfügung stehen. Schmidt et al. haben aber noch eine weitere, einfachere Schätzung durch empirische Forschung entwickelt, die besagt, dass die Standardabweichung der Berufsleistung in Geldeinheiten (SD_y) ca. 40-70% des Gehalts beträgt. Eine konservative Schätzung von SD_y liegt also bei 40% des Gehalts. Im Folgenden wird auf diese Faustregel als die 40%-Regel verwiesen. Die Autoren haben bereits in ihrer Publikation 1982 dargelegt, dass sie diese These empirisch stützen können. Der Grundgedanke ist, dass SD_y von der Tendenz her proportional zur Höhe des Outputs (d.h. der Arbeitsleistung) verläuft und somit auch proportional zum Gehalt. Es ergab sich, dass das Output eines durchschnittlichen Arbeitnehmers ungefähr doppelt so hoch wie sein Gehalt ist. Dementsprechend beträgt SD_y vom Output ca. 20% bis 35% (Schmidt et al. 1982).¹⁸

¹⁸ Die Standardabweichung der Berufsleistung in GE (SD_y) beträgt ca. 40-70% des Gehalts. Da der Output eines durchschnittlichen AN das Doppelte seines Gehalts beträgt, verdoppelt sich die Basis, auf die sich die SD_y

Für die Ermittlung von d_t gehen Schmidt et al. in ihrer Darstellung davon aus, dass die Veränderung der Arbeitsleistung der Teilnehmer und Nicht-Teilnehmer durch deren Vorgesetzten über einen längeren Zeitraum nach der Trainingsteilnahme beurteilt wird (Schmidt et al. 1982: 335). Es lässt sich hieraus schließen, dass es sich bei der Leistungsveränderung um eine beobachtete Veränderung auf der Transferebene handelt und nicht auf der Lernerfolgsebene. Schmidt et al. gehen nicht weiter auf diesen Aspekt ein. Es ist hierbei aber anzumerken, dass es sich bei der gemessenen Leistungsveränderung um jeden Fall um eine Größe handeln muss, die in direktem Bezug zur Berufsleistung steht. In ihrem Beispiel werden diese Vorgesetztenbeurteilungen an sechs Zeitpunkten im Abstand von einem Monat nach dem Training erhoben (Schmidt et al. 1982: 335).

Der Index t bei dem Parameter d_t drückt aus, dass es sich um die wahre (auf englisch „true“) Veränderung handelt, aufgrund dessen wird die in Standardeinheiten gemessene Veränderung ($d = \frac{Y_1 - Y_0}{SD}$) noch um die Reliabilität des zugrunde liegenden Messinstruments korrigiert.

Hierfür wird der Parameter d noch mit Hilfe des geschätzten Reliabilitätskoeffizienten r_{yy} berichtigt. Durch diese Operation erhält man $d_t = \frac{Y_1 - Y_0}{SD * \sqrt{r_{yy}}}$. Wie oben bereits thematisiert,

gehen Schmidt et al. in ihren Ausführungen davon aus, dass die Leistungsveränderung durch Bewertungen von Führungskräften ermittelt werden. In diesem Fall muss für die Tatsache kontrolliert werden, dass die Bewertungen unter Umständen vom Bewerter abhängen. Um dies zu kontrollieren, wird die Interrater-Reliabilität¹⁹ betrachtet (Schmidt et al. 1982: 336). Sofern aber andere individuelle Instrumente für die Leistungsbewertung benutzt werden, ist die Reliabilität des jeweiligen Instruments zu ermitteln. Ein weiterer kritischer Parameter bei der Schätzung ist die Dauer des Trainingseffekts auf die Berufsleistung in Jahren (T), der im Folgenden als Zeitfaktor bezeichnet wird. Schmidt et al. schlagen vor, als Schätzer für den Zeiteffekt die geschätzte Amortisationsdauer des Trainings zu halbieren (Schmidt et al. 1982: 339). Leider bleibt ungeklärt, ob sich diese Schätzmethode auf empirische Ergebnissen stützt. In Anwendungen der Schmidt, Hunter und Pearlman Formel wird der Zeitfaktor mit Hilfe von Schätzungen von Betroffenen ermittelt (Gülpen 1996: 52).

bezieht, um das Doppelte und so halbiert sich die SD_y um den Faktor 0,5, d.h. SD_y beträgt 20-35% bezogen auf das totale Output.

¹⁹ Die Interrater-Reliabilität wird ermittelt, indem der Korrelationskoeffizient zwischen den Beurteilungen der zwei Beurteiler berechnet wird.

5. Fazit

Zunächst wurde ein Überblick über bestehende Evaluationsansätze in der betriebswirtschaftlichen Literatur gegeben. Auffallend hierbei ist, dass sich alle Ansätze in irgendeiner Form an Kirkpatrick's Ansatz von 1959 orientieren oder zumindest Teilaspekte wieder aufgreifen. Zusammenfassend lässt sich hier feststellen, dass Kirkpatrick's Grundidee von der Evaluation auf verschiedenen Levels immer noch Bestand hat. Die Ansätze, die zeitlich folgen, haben diese Grundidee teilweise punktuell weiterentwickelt (vgl. Ansatz von Kraiger, Ford und Salas) oder durch Einbezug von vorgeschalteten Bereichen erweitert (vgl. prozessorientierte Ansätze) bzw. um weitere Levels ergänzt (vgl. Phillips). Gerade Phillips Erweiterung von Kirkpatrick's Ansatz verdeutlicht die Tendenz, die in der Theorie besteht, den Nutzen bzw. Erfolg von Trainingsmaßnahmen verstärkt in quantitativer Form zu erfassen. Als eine zentrale Gemeinsamkeit aller Ansätze lässt sich feststellen, dass jeder Ansatz die Evaluation des „Outcomes“ integriert. Keiner der genannten Ansätze thematisiert aber wirklich die Schwierigkeit, die bei einer derartigen Evaluation besteht. Kirkpatrick erwähnt in seinen Standards für die zwei Level *Learning* und *Behaviour* den Einsatz von Vorher-Nachher Designs und die Bildung einer Kontrollgruppe, geht aber nicht weiter darauf ein. In den Publikationen zu den anderen Ansätzen werden mögliche Evaluationsdesigns bei der Ergebnisevaluation in der Regel gar nicht thematisiert. Gerade dieser Bereich ist Gegenstand der volkswirtschaftlichen Literatur zur Evaluation von Arbeitsmarktprogrammen. Die Entwicklung neuer Methoden in diesem Bereich und die Bereitstellung der Methoden für den angewandten Forscher machen diese Literatur zu einem regen Forschungsfeld. Schließlich führen diese Erkenntnisse dazu, dass eine Entscheidung über ein Evaluationsdesign unter einem genauen Bewusstsein der unterstellten Annahmen getroffen werden kann. Aus dieser Perspektive sollte die Kenntnis dieser Literatur, zumindest in ihren Grundzügen, für jeden Evaluationsverantwortlichen selbstverständlich sein. Innerhalb der quantitativ orientierten Forschung wird die Programmevaluation oft auch mit Hilfe einer linearen Regression betrieben. Da diese Methode in vielen wissenschaftlichen Studien zur Anwendung kommt wurde in Abschnitt 3.2 dargestellt, dass das Ergebnis einer linearen Regression (mit einer unabhängigen Variablen) auf der Annahme des CS Schätzers beruht, welcher aufgrund seiner geringen Ansprüche an die Daten sehr beliebt ist.

Als eine Möglichkeit der Evaluation auf der Unternehmenserfolgsebene ist die Kosten-Nutzen-Analyse von Schmid, Hunter und Pearlman vorgestellt worden. Diese Formel ermöglicht die Ermittlung des Nettonutzens von Trainingsprogrammen in Geldeinheiten.

Zusammenfassend lässt sich zum Stand der Evaluationsforschung sagen, dass zumindest in der betriebswirtschaftlichen Literatur Kirkpatrick's Ansatz immer noch große Bedeutung zukommt. Kirkpatrick's Evaluationsansatz wird durch die Forschung in anderen Bereichen ergänzt. So finden sich beispielsweise in der psychologischen Literatur Forschungsergebnisse für Instrumente zur Messung des Lernerfolgs und des Transfererfolgs; die volkswirtschaftliche Literatur liefert Erkenntnisse für die Wahl von Evaluationsdesigns. Schmidt, Hunter und Pearlman (1982) stellen, gemessen an den aktuellen Entwicklungen der volkswirtschaftlichen Forschung, schon früh ein Instrument zur Quantifizierung des Nutzens von Trainingsprogrammen bereit. Die offensichtlichen Schwächen, die mit der Formel von Schmidt, Hunter und Pearlman (1982) verbunden sind, begründen jedoch den weiteren Handlungsbedarf in diesem Forschungsfeld. So ist der entscheidende Nachteil der Formel, dass eine Vielzahl von Schätzern multiplikativ miteinander verknüpft sind. Ist auch nur einer der Schätzer nicht vertrauenswürdig oder sogar falsch ermittelt worden, verliert die ganze Berechnung an Substanz. Sturman (2000) identifiziert die Komplexität der Nutzenanalyse und das Fehlen an Schulungen für HR Managern in diesem Bereich als zentrale Probleme der Nutzenanalyse von HR-Maßnahmen (Sturman 2000: 296).

Der Blick in die Unternehmenspraxis zeigt, dass unsystematische und subjektive Evaluationen die betriebliche Evaluationspraxis dominieren. Viele Unternehmen führen nur eine Evaluation der Zufriedenheit durch, was einer Evaluation auf dem ersten Level bei Kirkpatrick's Evaluationsansatz entspricht. Dies zeigt, dass zwischen Theorie und Praxis eine starke Diskrepanz besteht. Der in der Einleitung ausgeführte Blick auf die Kriterien bei der Wahl einer Evaluationsform anhand der Studie von Gülpen (1996) war entlarvend. Hierbei zeigt sich, dass die Befragten die Kriterien „Quantitativer Charakter der Ergebnisse“, „Legitimation des Seminars“ und „Wissenschaftliche Fundierung“, welche sowohl die Art und die Zielsetzung einer quantitativen Evaluation beschreiben, zu den unwichtigsten aller Kriterien zählen. Die Ergebnisse der Untersuchung der Gründe für die Nicht-Quantifizierung zeigen, dass viele Unternehmen der Meinung sind, dass ihnen sowohl das Know-How als auch die technischen Möglichkeiten für eine quantitative Evaluation fehlen. Verbunden mit den sehr hoch eingeschätzten Kosten für eine quantitative Evaluation erscheint es nahe liegend, dass Unternehmen eher qualitative und informelle Evaluationen als quantitative Evaluation betreiben. Die Unternehmen fühlen sich kompetent genug, diese durchzuführen und erhalten gewünschte Ergebnisse und Entscheidungshilfen für die Planung und Durchführung, die eine quantitative Evaluation nicht immer liefern kann. Die Frage bleibt jedoch, ob die Kosten quantitativer Evaluationsformen tatsächlich höher als der mögliche Nutzen sind, wie von den

Unternehmen vermutet. Ein weiteres Problem liegt sicherlich darin, dass Unternehmen (wie oben bereits angesprochen) kein Vertrauen in die Ergebnisse einer quantitativ geprägten Evaluation haben. Berücksichtigt man, dass geschulte Kenntnisse und Erfahrungen aus verschiedenen wissenschaftlichen Bereichen für eine fundierte quantitative Evaluation verlangt werden, so ist es nicht erstaunlich, dass die Praxis den Aufwand für eine quantitative Evaluation als sehr hoch beurteilt und den Nutzen kaum einschätzen kann. Dies führt in der betrieblichen Praxis zu einer starken Voreingenommenheit gegenüber einer quantitativen Evaluation, welche durch ein einheitliches und interdisziplinäres Konzept mit praxisnahen Lösungen abgebaut werden könnte. Neben dem Problem der Komplexität weist Dionne (1996) auf eine weitere Schwierigkeit hin: Es ist problematisch, dass die drei im Bereich der Evaluationsforschung betroffenen Parteien: Wissenschaftler, Trainer und HR-Manager andere Standards und Verwendungszwecke für die erzielten Ergebnisse haben (Dionne 1996: 282). Hinsichtlich der aufgezeigten Diskrepanz zwischen Theorie und Praxis ist es deshalb von Bedeutung, dass die Wissenschaftler bei ihrer weiteren Forschung den Blickwinkel der Praxis verstärkt berücksichtigen. Denn sobald HR-Manager eine quantitative Evaluation für umsetzbar und die Ergebnisse als nützlich beurteilen, wird diese Evaluationsform auch stärker in der betrieblichen Praxis eingesetzt. Dies würde indirekt auch dazu führen, dass ein zentrales von Sturman angesprochenes Problemfeld, die mangelnde Schulung der HR-Manager im Bereich der Nutzenanalyse bzw. quantitativer Evaluation im Allgemeinen gelöst wird.

Verzeichnis der zitierten Literatur

- Alliger, George; Janak, Elizabeth (1989): Kirkpatrick's Levels of Training Criteria: Thirty years later. In: *Personnel Psychology*, Vol. 42, S. 331-342.
- ASTD (2003): *State of Industry Report*. Alexandria (American Society for Training and Development).
- Becker, Fred G. (2005): Den Return on Development messen: Möglichkeiten und Grenzen der Evaluation. In: *Personalführung*, Heft 4/2005, S. 48-53.
- Beywl, Wolfgang (1999): Evaluatives Denken im betrieblichen Bildungsmanagement. In: *Evaluationsbedarf in der betrieblichen Bildung. Rostocker Arbeitspapiere zur Wirtschaftsentwicklung und Human Resource Development*, Rostock, S. 17-34.
- Blundell Richard; Costa Dias, Monica (2000): Evaluation Methods for Non-Experimental Data. In: *Fiscal Studies*, Vol. 21, Nr.4, S. 427-468.
- Brinkerhoff, Robert (1987): *Achieving Results from Training. How to Evaluate Human Resource Development to Strengthen Programs and Increase Impact*. San Francisco.
- Burtless, Gary (1995): The Case for Randomized Field Trials in Economic and Policy Research. In: *Journal of Economic Perspectives*, Vol. 9, Nr. 2, S. 63-84.
- Bushnell, David (1990): Input, Process, Output: A Model for Evaluating Training. In: *Training and Development*, Vol. 44, S. 41-43.
- Caliendo, M.; Hujer, R. (im Erscheinen): The Microeconomic Estimation of Treatment Effects – An Overview. In: *Allgemeines Statistisches Archiv*.
- Caliendo, M.; Kopeing, S. (2005): Some Practical Guidance for the Implementation of Propensity Score Matching. Discussion Paper Nr. 1588, IZA, Bonn
- Cobb-Clark, Deborah; Crossley, Thomas. (2003): *Econometrics for Evaluations: An Introduction to Recent Developments*. In: *Economic Record*, Vol. 79, S. 491-511.
- Dionne, Pierre (1996): *The Evaluation of Training Activities: A Complex Issue Involving Different Stakes*. *Human Resource Development Quarterly*, Vol. 7, S. 279-286
- Fitz-Ens, Jac (1994): Yes...You can weigh training's value. In: *Training and Development*, Vol. 31, S. 54-58.
- Galvin, James C. (1983): What can trainers learn from educators about evaluating management training? In: *Training and Development*, Vol. 37, S.52-57.
- Gensler, Sonja; Skiera, Bernd; Böhm, Martin (2005): Einsatzmöglichkeiten der Matching Methode zur Berücksichtigung von Selbstselektion. In: *Journal für Betriebswirtschaft*, Vol. 55, S. 37-62.
- Gülpen, Barbara (1996): *Evaluation betrieblicher Verhaltenstrainings unter der besonderen Berücksichtigung des Nutzens*, München (Zugl.: Erlangen, Nürnberg, Univ. Diss., 1995).
- Heckman, J.; LaLonde, R.; Smith, J. (1999): *The Economics and Econometrics of Active Labor Market Programs*. In: *Handbook of Labor Economics*. Vol.3, Ashenfelter, O.; Card, D. (Hrsg). Amsterdam.
- Holton, Elwood (1996): The Flawed Four-Level Evaluation Model. In: *Human Resources Development Quarterly*, Vol. 7, S. 5-21.
- Kirkpatrick, Donald (1979): Techniques for evaluating training programs. In: *Training and Development*, Vol. 33, S. 78-92.
- Kirkpatrick, Donald (1996): Invited Reaction to Holton Article. In: *Human Resources Development Quarterly*, Vol. 7, S. 23-25.
- Kirkpatrick, Donald (1998): *Evaluating training programs – the four levels*. 2. Aufl., San Francisco.
- Kraiger, K.; Ford, K.; Salas, E. (1993): Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation. In: *Journal of Applied Psychology*, Vol. 78, S. 311-328.

- Lechner, Michael (1998): Mikroökonomische Evaluationsstudien: Anmerkungen zu Theorie und Praxis. In: Qualifikation, Weiterbildung und Arbeitsmarkterfolg, Pfeiffer, F., Pohlmeier, W. (Hrsg.), 1. Auflage, Baden-Baden.
- Parry, S.B. (2000): Training for Results. ASTD, Alexandria.
- Phillips, Jack J. (1996): ROI: The Search for Best Practices. In: Training and Development, Vol. 50, S. 42-47.
- Pullig, Karl-Klaus (1987): Vorwort: Weiterbildung im Wandel – Ergebnisse einer Befragung. In: Weiterbildung im Wandel, Pullig, K.-K., Schäkel, U. (Hrsg.), Hamburg.
- Roy, A. D. (1951): Some Thoughts on the Distribution of Earnings. In: Oxford Economic Papers, Vol. 3, S. 135-146.
- Rubin, Donald B. (1974): Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. In: Journal of Educational Psychology, Vol. 66, S. 688-701.
- Schmidt, Christoph (2001): Knowing what works. The case for rigorous programme evaluation. Universität zu Heidelberg, CEPR Diskussionspapier Nr. 2826.
- Schmidt, F; Hunter, J.; Pearlman, K. (1982): Assessing the economic impact of personnel programs on workforce productivity. In: Personnel Psychology, Vol. 35, S. 333-347.
- Spremann, Klaus (1991): Investition und Finanzierung. 4. verb. Aufl. München.
- Stoel, Diedrick (2004): The Evaluation – Heavy weight Match. In: Training and Development, Vol. 58, S. 46-48.
- Sturman, Michael (2000): Implication of Utility Analysis Adjustments for Estimates of Human Resource Intervention Value. In: Journal of Management, Vol. 26, S. 281-299.
- Werner, J.; De Simone, R. (2005): Human Resources Development. 4. Auflage., London.