



BAYERISCHE JULIUS-MAXIMILIANS
**UNIVERSITÄT
WÜRZBURG**
Psychologisches Institut

Die Bedeutung von Teilnehmereinschätzungen zu verschiedenen Zeitpunkten für die Vorhersage des Erfolgs von Personalentwicklungsmaßnahmen

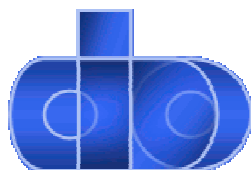
Vorgelegt von

Diana Beck

05. Oktober 2006

Betreuer

**Prof. Dr. Guido Hertel
Prof. Dr. K. Wolfgang Kallus**



**Arbeits-, Betriebs- und
Organisationspsychologie**

Diplomarbeit Nr. 12

Für meine Eltern.

Danksagung

Zunächst möchte ich mich bei Herrn Professor Dr. Guido Hertel, Bayerische Julius-Maximilians-Universität Würzburg, sowie bei Herrn Professor Dr. K. Wolfgang Kallus, Karl-Franzens-Universität Graz, für die Betreuung dieser Diplomarbeit bedanken. Ein weiterer Dank geht an meinen externen Betreuer vom Institut für Begleitforschung, Herrn Dipl.-Psych. Jens Brandt.

Durch den Einsatz und die freundliche Unterstützung von Frau Carin Jungmann (AchieveGlobal) sowie Herrn Peter Schulte (Ericsson) war es möglich, die Datenerhebung für die vorliegende Diplomarbeit durchzuführen. Beiden möchte ich für die angenehme Zusammenarbeit während des gesamten Projekts danken.

Mein Dank gilt ferner Frau Birgit Langenfels (Ericsson) und Frau Janet Grischkat (AchieveGlobal) als Ansprechpartnerinnen für organisatorische Fragen im Rahmen des Projekts. Vielen Dank auch an den Trainer, Herr Heiner Wiertz, für seine Unterstützung und Mithilfe bezüglich der Seminar-Feedbackbögen.

Zuletzt sei allen beteiligten Mitarbeitern der Firma Ericsson für die Unterstützung des Projekts und für ihre Geduld bei der Bearbeitung der Befragungen herzlich gedankt!

Zusammenfassung

Im Bereich der Evaluation von Personalentwicklungsmaßnahmen lassen sich vorrangig zwei bedeutsame Modelle finden: Das in der Praxis sehr bekannte Evaluationsmodell von Kirkpatrick (1959a, 1959b, 1960a, 1960b) mit den Ebenen *Reaktionen*, *Lernen*, *Verhalten* und *Ergebnisse* sowie das umfassende Rahmenmodell zur Trainingseffektivität von Tannenbaum (Cannon-Bowers, Salas, Tannenbaum & Mathieu, 1995). Obwohl gerade die erste Ebene des Modells von Kirkpatrick starke Kritik erfahren hat, bleibt das Maß der Teilnehmerreaktionen die am häufigsten erhobene Ebene. In der vorliegenden Arbeit wurde bei 47 Teilnehmern eines Vertriebsstrainings geprüft, welchen Beitrag die Messung der Teilnehmerreaktionen tatsächlich für die Vorhersage des Erfolgs von Personalentwicklungsmaßnahmen leisten kann. Hierfür wurde die Reaktionsebene in affektive (*affective reactions*) und nutzenbezogene Reaktionen (*utility reactions*) unterteilt. Diese Trennung offenbarte hohe Korrelationen der nutzenbezogenen Reaktionen mit der subjektiven Verhaltenseinschätzung, welche als Maß für die Verhaltensebene herangezogen wurde. Allgemein korrelierten die *nutzenbezogenen Reaktionen* höher mit den weiteren hier gemessenen Ebenen (*Lernen* und *Verhalten*) als die *affektiven Reaktionen*. Regressionsanalytisch konnte die starke Vorhersagekraft dieser Nutzeinschätzungen gezeigt werden. Über das Kirkpatrick-Modell hinaus wurden einige der im Tannenbaum-Modell identifizierten Einflussfaktoren erhoben. Die Unterstützung durch die Führungskraft bzw. durch das Arbeitsumfeld (Transferklima) stellte sich dabei als eine bedeutsame Größe heraus, die nicht nur auf das Lernen und das Verhalten, sondern ebenfalls auf die Teilnehmerreaktionen einwirkt. Ebenso hervorzuheben ist der Einfluss des subjektiv empfundenen Trainingsbedarfs der Teilnehmer für die unmittelbare Trainingsbewertung. Von Bedeutung sind letztlich auch die Erkenntnisse, die sich aus der Messung der Teilnehmerreaktionen zu drei Zeitpunkten ergeben. Ganz im Sinne von Clement (1982) ist durch die mehrfache Messung der Teilnehmerreaktionen das Erheben von Zusatz-Informationen möglich, die für ein nachhaltiges Qualitätsmanagement eine Rolle spielen können: So kann die Mehrfachmessung beispielsweise Informationen darüber liefern, wie der Transferprozess verläuft und schafft so die Möglichkeit, in diesen Verlauf unterstützend einzugreifen.

Inhaltsverzeichnis

1	Einleitung	8
2	Theoretischer Hintergrund	10
2.1	Betriebliche Weiterbildung und Personalentwicklung.....	10
2.1.1	Begriffsklärung.....	10
2.1.2	Funktionen, Ziele und Bedeutung	11
2.2	Evaluation.....	12
2.2.1	Begriffsklärung.....	12
2.2.2	Funktionen, Ziele und Bedeutung	13
2.3	Evaluationsmodelle in der PE – Stand der Forschung	14
2.3.1	Modell der vier Ebenen nach Kirkpatrick	14
2.3.2	Modell des Transfers nach Baldwin und Ford	19
2.3.3	Modell der Trainingseffektivität nach Tannenbaum	21
2.4	Evaluationsmodelle der PE und ihre Anwendung in der Praxis	25
2.4.1	Anwendung der einzelnen Ebenen nach Kirkpatrick	26
2.4.2	Hindernisse bei der Durchführung von Evaluation	27
2.4.3	Reaktionen als „Happiness index“? – Kritik an Kirkpatrick.....	28
2.4.4	Vorteile einer Reaktionsmessung zu mehreren Zeitpunkten.....	32
2.4.5	Bedeutung von Feedback im Evaluationsprozess	32
2.5	Schlussfolgerungen	34
3	Fragestellungen und Hypothesen	36
3.1	Korrelationen zwischen den Evaluationsebenen nach Kirkpatrick.....	36
3.2	Vorhersage des Seminarerfolgs anhand der Reaktionen.....	37
3.3	Auswirkung von Einflussgrößen auf die verschiedenen Ebenen	38
4	Methodische Umsetzung	41
4.1	Stichprobenbeschreibung	41
4.1.1	Seminarteilnehmer (Trainees)	41
4.1.2	Kontrollgruppe	42

4.1.3	Führungskräfte.....	43
4.1.4	Kontaktaufnahme	44
4.1.5	Datenschutz	44
4.2	Untersuchungsdesign	45
4.3	Untersuchungsablauf	47
4.4	Untersuchungsvariablen	49
4.4.1	Kriterien nach Kirkpatrick.....	49
4.4.1.1	Reaktionsebene.....	49
4.4.1.2	Lernebene - Wissen	53
4.4.1.3	Lernebene - Einstellungen.....	54
4.4.1.4	Verhaltensebene	55
4.4.1.5	Auswahl der Zielkriterien.....	55
4.4.2	Kriterien nach Tannenbaum	56
4.4.2.1	Motivation	57
4.4.2.2	Selbstwirksamkeit	58
4.4.2.3	Subjektiver Bedarf.....	58
4.4.2.4	Anwendungsmöglichkeit aus Sicht des Unternehmens.....	59
4.4.2.5	Vorerfahrung (Expertise)	59
4.4.2.6	Transferklima	59
4.4.2.7	Seminarbewertung.....	60
4.5	Datenaufbereitung	61
4.5.1	Rücklaufquote.....	61
4.5.2	Statistische Auswertung	62
4.5.2.1	Prüfung der Hypothesen.....	62
4.5.2.2	Effektstärke	64
4.5.2.3	Auswertungen im Vorfeld.....	65
5	Ergebnisse.....	69
5.1	Zusammenhänge zwischen den Ebenen nach Kirkpatrick.....	69
5.1.1	Lernebene und Verhaltensebene.....	71

5.2	Vorhersage des Seminarerfolgs anhand der Teilnehmerreaktionen.....	71
5.2.1	Lernebene	72
5.2.2	Verhaltensebene	73
5.3	Auswirkung möglicher Einflussgrößen auf die Evaluationsebenen.....	74
5.3.1	Motivation	76
5.3.2	Selbstwirksamkeit.....	76
5.3.3	Subjektiver Bedarf.....	77
5.3.4	Anwendungsmöglichkeit.....	77
5.3.5	Vorerfahrung	78
5.3.6	Transferklima	80
5.3.7	Allgemeine Seminarbewertung	83
5.3.8	Spezifische Seminarbewertung	84
5.4	Folgeanalysen.....	87
5.4.1	Konsequenzen der Trennung in <i>affective</i> und <i>utility reactions</i>	87
5.4.2	Nachgeschobene Regressionsanalysen.....	89
5.4.3	Zufriedenheitseinschätzungen zu t1 und t2	91
5.4.4	Zusammenhang der Anwendbarkeit t1 und Anwendung t2 und t3	92
6	Diskussion.....	94
6.1	Zusammenfassung der Ergebnisse	94
6.1.1	Zusammenhänge zwischen Reaktionen, Lernen und Verhalten.....	94
6.1.2	Die Rolle der Teilnehmerreaktionen zur Vorhersage des Seminarerfolgs.....	96
6.1.3	Faktoren mit Einfluss auf Reaktionen, Lernen und Verhalten	97
6.2	Verlauf der Einschätzungen zur Anwendung der Seminarinhalte	103
6.3	Bedeutung von Mehrfachmessungen	104
6.4	Methodische Limitationen.....	107
6.5	Fazit und Ausblick	110
7	Literaturverzeichnis	113
8	Eidesstattliche Erklärung.....	119

1 Einleitung

Zunehmende Globalisierung sowie immer stärkerer Wettbewerb führen im Rahmen der Personalentwicklung (PE) zu einem kontinuierlichen Bedarf an Weiterbildungsmaßnahmen und somit zu hohen Ausgaben (Arthur, Bennett, Edens & Bell, 2003). Hier stellt sich die Frage, wie der Erfolg solcher Maßnahmen im Sinne einer Qualitätssicherung sicherzustellen und abzubilden ist. Auf der Suche nach entsprechenden Evaluationsmöglichkeiten im PE-Bereich findet man in der Praxis am häufigsten das Evaluationsmodell von Kirkpatrick (1996), welches durch seine Einfachheit einen hohen Stellenwert innehat. Bei genauerem Hinsehen klafft jedoch eine Lücke zwischen der wahrgenommenen Bedeutung einer Evaluation auf mehreren Ebenen und ihrer tatsächlichen Realisierung in der Praxis: Das Vier-Ebenen-Modell Kirkpatricks mit den Ebenen *Reaktionen*, *Lernen*, *Verhalten* und *Ergebnissen* ist zwar das bekannteste Evaluationsmodell für PE-Maßnahmen (Salas & Cannon-Bowers, 2001), praktisch gesehen wird jedoch nur die erste Ebene am häufigsten angewandt (Borchert & Rutschke, 2005; Van Buren & Erskine, 2002). Dabei sind die erfassten Teilnehmerreaktionen nicht mit einer Verhaltensreaktion zu verwechseln – sie bilden die subjektive Zufriedenheit der Teilnehmer¹ mit verschiedenen Aspekten der Maßnahme ab.

Trotz ihres dominierenden Einsatzes in der Evaluationspraxis von Unternehmen und Weiterbildungsinstituten ist die erste Ebene nicht unumstritten (z.B. Holton, 1996; Bates, 2004). Den Hauptvorwurf spiegelt die gängige Bezeichnung „happiness index“ oder „smile sheets“ wider. Dahinter verbirgt sich die Kritik, eine Erhebung der Teilnehmerreaktionen sei lediglich ein Stimmungsbild darüber, wie „happy“ die Teilnehmer mit Trainer, Seminar, Ort, Verpflegung etc. sind. Von verschiedenen Seiten wird daher eine Differenzierung gefordert (Alliger, Tannenbaum, Bennett, Traver & Shotland, 1997; Morgan & Casper, 2000; Warr & Bunce, 1995), um diesen Feedbackbögen weitergehende Informationen zu entnehmen und Zusammenhänge zwischen den weiteren Erfolgskriterien einer Maßnahme aufzeigen zu können. Als Erfolgskriterien gelten dabei neben einem Wissenszuwachs Einstellungsänderungen sowie Verhaltensänderungen (im Sinne einer Transferleistung der neu

¹ In der vorliegenden Arbeit werden aufgrund der besseren Lesbarkeit geschlechtsspezifische Wörter stets in der männlichen Form verwendet. Alle Teilnehmerinnen, Mitarbeiterinnen etc. werden jedoch gebeten, sich gleichermaßen angesprochen zu fühlen.

erworbenen Inhalte in die Praxis). Verschiedenen Seiten (z.B. Alliger et al., 1997; Warr & Bunce, 1995) fordern daher, von den Teilnehmern nicht nur affektive Zufriedenheitseinschätzungen (*affective reactions*) einzuholen, sondern auch solche, die den Nutzen der Maßnahme (*utility reactions*) erfassen. Ein weiterer Vorwurf an das Modell Kirkpatrick's liegt in seiner Einfachheit, d.h. in der fehlenden Berücksichtigung der Einflussfaktoren, die auf den gesamten Prozess einwirken können (Bates, 2004). Diesem Vorwurf wird im Modell zur Effektivität von PE-Maßnahmen von Tannenbaum (Cannon-Bowers et al., 1995) begegnet: Darin werden die Kirkpatrick-Ebenen mit individuellen, organisationalen und trainingsbezogenen Merkmalen in Verbindung gebracht, die die Effektivität einer Maßnahme vor, während und nach dem Training beeinflussen können (vgl. Salas & Cannon-Bowers, 2001).

Die vorliegende Arbeit soll einen Beitrag leisten, die Zusammenhänge der Effektivitätsmaße des Modells von Kirkpatrick (1996) in der Evaluationspraxis von PE-Maßnahmen näher zu spezifizieren. Es wird der Frage nachgegangen, welchen Beitrag die Teilnehmerreaktionen zur Ermittlung bzw. Vorhersage des Seminarerfolgs leisten können. Dafür wird entsprechend der Forderung von Alliger et al. (1997) eine Trennung in affektive sowie nutzenbezogene Teilnehmerreaktionen (*affective* und *utility reactions*) vorgenommen. Auch wird die Empfehlung von Clement (1982) übernommen, über ein zeitlich gestaffeltes Vorgehen, d.h. über mehrfache und inhaltlich verschiedene Reaktionsmessungen Anhaltspunkte über hinderliche und förderliche Faktoren zu erhalten. Zusätzlich zum unmittelbaren Feedback am Ende des Seminars werden die Teilnehmer mittels zweier Online-Befragungen (ca. zwei Wochen sowie ca. drei Monate später) zu ihrer Zufriedenheit mit dem Seminar und zur Umsetzung der Inhalte befragt. Darüber hinaus wird im Lichte der Forschungsergebnisse bezüglich des Trainingseffektivitäts-Modells von Tannenbaum (Cannon-Bowers et al., 1995) untersucht, welche Einflussgrößen sich auf die hier erhobenen Effektivitätsmaße *Reaktionen*, *Lernen* und *Verhalten* auswirken.

Im nachfolgenden Kapitel 2 wird zuerst näher auf die theoretische Basis dieser Arbeit eingegangen, bevor in Kapitel 3 die zu untersuchenden Fragestellungen dargestellt und die dazugehörigen Hypothesen formuliert werden. Der methodische Aufbau und die verwendeten Instrumente werden in Kapitel 4 ausführlich vorgestellt. In Kapitel 5 werden die gefundenen Ergebnisse beschrieben. Eine Interpretation dieser Ergebnisse und ihre Implikationen bzw. die daraus abzuleitenden Konsequenzen werden abschließend im Kapitel 6 dargestellt.

2 Theoretischer Hintergrund

Einleitend wird in Abschnitt 2.1 auf die Begriffe Weiterbildung und PE-Maßnahmen eingegangen, bevor ihre Bedeutung herausgearbeitet wird. Abschnitt 2.2 geht zu Beginn auf den Begriff der Evaluation ein und erläutert die unterschiedlichen Arten von Evaluation sowie ihre Ziele und Bedeutung. Die einzelnen Evaluationsmodelle für PE-Maßnahmen werden in Abschnitt 2.3 vorgestellt, allen voran das Vier-Ebenen-Modell von Kirkpatrick (1996), das Transfermodell von Baldwin und Ford (1988) sowie das Modell zur Trainingseffektivität von Tannenbaum (Cannon-Bowers et al., 1995; Höft, 2001). Nach der Beschreibung der einzelnen Modelle wird in Abschnitt 2.4 darauf eingegangen, wie es um die Anwendung dieser Modelle in der Praxis steht, welche Kritik am Modell von Kirkpatrick angebracht wird und wie dieser zu begegnen ist. Abschließend wird mit Abschnitt 2.5 ein Resümee gezogen und die Bedeutung der vorliegenden Arbeit herausgearbeitet.

2.1 Betriebliche Weiterbildung und Personalentwicklung

2.1.1 Begriffsklärung

Weiterbildung wird als wichtiges Instrument der betrieblichen Personal- und Organisationsentwicklung angesehen (Sauter, 1995). Die Personalentwicklung ihrerseits stellt wiederum einen wichtigen Bereich des Human Resource Management dar und ist ein zentrales Aufgabengebiet innerhalb der Arbeits- und Organisationspsychologie.

Nach Holling und Liepmann (1995) wird der Begriff „Personalentwicklung“ erst seit Mitte der 1970er Jahre systematisch verwendet. Inzwischen ist der Begriff jedoch etabliert und umfasst ein weites Gebiet. Unter Personalentwicklung (PE) sind laut Holling und Liepmann (1995) alle planmäßigen personen-, stellen- und arbeitsplatzbezogenen Maßnahmen zur Ausbildung, Erhaltung oder Wiedererlangung der beruflichen Qualifikation zu verstehen. Diese systematische Förderung der beruflichen Qualifikation von Mitarbeitern beschränkt sich nicht nur auf Fachkenntnisse und Fertigkeiten, sondern umfasst zusätzlich auch z.B. die Lernfähigkeit, Motivation, soziale Kompetenzen und den Umgang mit Belastungen (Holling & Liepmann, 1995).

Unter beruflicher Weiterbildung werden laut dem Arbeitsförderungsgesetz (§41 AFG) alle Maßnahmen verstanden, „ ... *die das Ziel haben, berufliche Kenntnisse und Fertigkeiten festzustellen, zu erhalten, zu erweitern oder der technischen Entwicklung anzupassen ...* “. Ob eine Weiterbildungsmaßnahme nun Training, Seminar, Kurs, Fortbildung, Schulung, PE-Maßnahme oder einfach nur Maßnahme genannt wird, soll nicht weiter verfolgt werden, da es den Rahmen dieser Diplomarbeit überschreiten würde: In der vorliegenden Arbeit werden die Begriffe synonym verwendet, da alle diese Begriffe das gemeinsame Ziel haben, das Wissen und/ oder das Verhalten von Mitarbeitern derart zu modifizieren und effizient(er) zu gestalten, dass die Ausführung der Tätigkeit nach der Maßnahme besser zur Erreichung von Unternehmenszielen beiträgt.

2.1.2 Funktionen, Ziele und Bedeutung

In Zeiten von Globalisierung und starkem Wettbewerb ist es für jedes Unternehmen wichtig, seine Mitarbeiter auf die neuen Herausforderungen und sich verändernde Anforderungen vorzubereiten. Eine Investition in die Qualifikation der Mitarbeiter soll eine höhere Produktivität des Unternehmens nach sich ziehen und hat demnach Auswirkungen auf den zukünftigen Erfolg des Unternehmens. Um wettbewerbsfähig zu bleiben, setzt man am häufigsten auf Trainingsmaßnahmen (Arthur et al., 2003). Angesichts von jährlichen Ausgaben über 54 Milliarden US-Dollar (Industry Report, 2000) sind Trainings als teure Investitionen zu betrachten. Betrugen die Ausgaben für Weiterbildung in Deutschland Anfang der 80er Jahre noch ca. 8 Milliarden DM², waren es 1992 bereits über 36 Milliarden DM³ (Kühnlein, 1997). Laut einer Studie des Instituts der deutschen Wirtschaft Köln lagen im Jahre 2005 die Ausgaben für betriebliche Weiterbildung hochgerechnet bei ca. 27 Mrd. Euro (Werner, 2006). Ein weiteres Ergebnis dieser Studie ist die Erwartung von Seiten der Unternehmen, dass der Bedarf und somit die Ausgaben für PE-Maßnahmen weiter steigen werden. An Ausgaben für PE-Maßnahmen ist jedoch auch die Erwartung an die Wirtschaftlichkeit dieser Maßnahmen gekoppelt, weshalb Phillips und Phillips (2001) in diesem Bereich von einem ebenfalls steigenden Interesse an Evaluationsmaßnahmen ausgehen. Das erklärte Ziel von Trainingsmaßnahmen liegt darin, erlerntes Wissen, Fertigkeiten und/oder Einstel-

² ca. 4 Mrd. €

³ ca. 18 Mrd. €

lungen auf den Arbeitsalltag erfolgreich zu übertragen, wodurch eine Leistungssteigerung erzielt wird oder erzielt werden soll (Ford & Kraiger, 1995). Um diese Übertragung sicherzustellen, sind systematische Evaluationen unerlässlich.

2.2 Evaluation

2.2.1 Begriffsklärung

Die Definition von Evaluation wird heutzutage aufgrund der Vielfältigkeit und der Dehnbarkeit des Begriffs als schwierig angesehen (Wottawa & Thierau, 2003). Den lateinischen Wurzeln zufolge lässt sich Evaluation vom Verb *valere* ableiten und bedeutet u.a. ‚sich wirksam erweisen‘. Verwirrend ist die Vielzahl der Begriffe, die im Zusammenhang mit Evaluation zum Teil gleichbedeutend benutzt werden oder mit denen speziellere Evaluationsformen beschrieben werden sollen – z.B. Erfolgskontrolle, Begleitforschung, Bewertungsforschung, Wirkungskontrolle, Qualitätskontrolle, Bildungscontrolling etc. (Wottawa & Thierau, 2003).

Ein Versuch zur Begriffsklärung wird darin gesehen, *Evaluation* von *Evaluationsforschung* zu unterscheiden (vgl. Wottawa & Thierau, 2003). Der Begriff der *Evaluation* wird hier etwas globaler als Bewertung gesehen, d.h. als einen Beurteilungsprozess, der den Wert eines Produktes, Prozesses oder Programms festlegt. Dabei müssen eine systematische Vorgehensweise und empirische Befunde nicht notwendigerweise vorhanden sein. Im Gegensatz dazu fordert die *Evaluationsforschung* ausdrücklich die Verwendung von wissenschaftlichen Forschungsmethoden bei der Durchführung eines Bewertungsprozesses. Hierbei wird statt einer bloßen Behauptung hinsichtlich des Nutzens der Beweis durch wissenschaftlich fundierte Methoden und Techniken in den Vordergrund gestellt.

Obwohl sie als zwei unterschiedliche Konstrukte betrachtet werden (Alvarez, Salas & Garofano, 2004), findet man oft einen synonymen Gebrauch der Begriffe *Trainings-evaluation* und *Trainingseffektivität*. *Trainingsevaluation* wird auf der einen Seite als Mess-technik gesehen, die das Ausmaß erfasst, in welchem bestimmte, a priori gesetzte Ziele durch das Training erreicht worden sind. Die verwendeten Methoden hängen dabei von eben diesen gesetzten Zielen ab und können verschiedene Bereiche abdecken, wie z.B. das Trainingsdesign bzw. die Trainingsinhalte, Veränderungen bei den Teilnehmern und Unter-

nehmungsergebnisse (organizational payoffs). Demgegenüber steht der Begriff der *Trainingseffektivität* für diejenigen Variablen, die einen Einfluss auf die Trainingsergebnisse vor, während und nach dem Training ausüben. Zur Messung dieser Effektivitätsvariablen, welche die Wirksamkeit bzw. den Erfolg des Trainings beeinflussen können, unterscheiden Alvarez et al. (2004) drei Kategorien: Variablen individueller Art, trainingsbezogene Variablen und organisationale Variablen. Genau diese Unterscheidung wird im Tannenbaum-Modell getroffen (Cannon-Bowers et al., 1995), welches an späterer Stelle noch genauer erläutert wird. Zusammenfassend sehen Alvarez et al. (2004) unter einer *Trainingsevaluation* den methodologischen Ansatz, um Lernergebnisse zu messen. Mit der Messung von Variablen der *Trainingseffektivität* hingegen wird die theoretische Annäherung abgebildet, die das Zustandekommen dieser Ergebnisse erklärt.

Scriven (1991) differenziert mit der summativen und der formativen Evaluation zwei Arten von Evaluation, bei denen neben dem Evaluationsziel auch der zeitliche Aspekt eine wichtige Rolle spielt. Ziel einer *summativen Evaluation* ist die Qualitätsbeurteilung nach Beendigung einer Weiterbildungsmaßnahme (Nork, 1991), d.h. erst am Ende einer erfolgten Maßnahme wird ihr Konzept auf Effektivität und Wirksamkeit hin überprüft. Hauptanliegen ist also die Messung des Einflusses auf individueller und organisationaler Ebene – während der Maßnahme erfolgt weder eine Messung noch eine Rückmeldung (Höft, 2001). Genau dieses Vorgehen zeichnet dagegen die *formative Evaluation* aus, bei der bereits während einer laufenden Maßnahme Informationen gesammelt werden. Die Rückmeldung dieser Zwischenergebnisse bildet die Basis für eine Analyse der laufenden Maßnahme und hat ihre Optimierung zum Ziel (Buchester, 2003; Höft, 2001). Weitere Kategorisierungsmöglichkeiten bieten Gülpen (1996) oder Scriven (1991).

2.2.2 Funktionen, Ziele und Bedeutung

Thierau (1991) zeigt verschiedene Aspekte von Evaluation auf, die als nutzenbringend für ein Unternehmen gesehen werden können. Im Kontext der vorliegenden Diplomarbeit treffen v.a. folgende zentrale Aspekte zu: Die *Bewertungs- und Beurteilungsfunktion* entspricht dem Prinzip der summativen Evaluation, es werden hierbei umfassende Informationen zu den Effekten der Maßnahme gesucht. In diesem Sinne können die gefundenen Ergebnisse eine *Optimierungsgrundlage* darstellen, d.h. die Maßnahme kann dem Unter-

nehmen sowie dem Seminaranbieter eine systematische Rückmeldung darüber geben, ob und in welchem Maße es verbesserungs- oder förderungswürdige Bereiche gibt. Eine detailliertere Übersicht der Funktionen von Evaluation findet sich in Thierau (1991) oder Wottawa und Thierau (2003). Für die an der Evaluation von PE-Maßnahmen beteiligten Personengruppen ergeben sich unterschiedliche Nutzenaspekte (vgl. Tabelle 2-1).

Tabelle 2-1. *Ziele und Motive der an einer Evaluation beteiligten Gruppen (Wottawa & Thierau, 2003, S. 57)*

Seminarteilnehmer	Trainer	Vorgesetzte	Unternehmensleitung
<ul style="list-style-type: none"> ▪ Lernerfolgsnachweis ▪ Karriereförderung ▪ Individuelles Feedback ▪ Lernmotivation 	<ul style="list-style-type: none"> ▪ Lehrerfolgsnachweis ▪ Bildungsbedarfsnachweis ▪ Feedback durch Teilnehmer ▪ Ressourcen-Gewinnung 	<ul style="list-style-type: none"> ▪ Berichterstattung über persönliche Eindrücke ▪ Entscheidungshilfen bei Personalfragen ▪ Beurteilung der Trainingsaktivitäten 	<ul style="list-style-type: none"> ▪ Beurteilung des Trainers ▪ Beurteilung der Teilnehmer ▪ Ressourcen-Bemessung ▪ Rechenschaftslegung ▪ Effizienz-Nachweis

Angesichts der Argumente, die für die Überprüfung der Wirksamkeit von Weiterbildungsmaßnahmen sprechen, müsste es theoretisch im Sinne eines jeden Trägers von Schulungsmaßnahmen, jedes Dozenten und jedes Weiterbildungsteilnehmers liegen, den Weiterbildungserfolg laufend zu überprüfen.

2.3 Evaluationsmodelle in der PE – Stand der Forschung

Obwohl das Stufenmodell von Kirkpatrick bereits vor über 40 Jahren erschien, gilt es trotz der Entwicklung neuerer Modelle als der bekannteste und in der PE-Praxis am meisten verwendete Evaluationsansatz (Kraiger, Ford & Salas, 1993; Salas & Cannon-Bowers, 2001).

2.3.1 Modell der vier Ebenen nach Kirkpatrick

Ursprünglich schlug Donald L. Kirkpatrick in den Jahren 1959 bis 1960 in vier nacheinander veröffentlichten Artikeln einen Evaluationsprozess vor, der als aufeinanderfolgende Schritte die Messung von Reaktionen (*reactions*), Lernen (*learning*), Verhalten (*behavior*) und Ergebnissen (*results*) beinhaltet (Kirkpatrick, (1959a, 1959b, 1960a, 1960b).

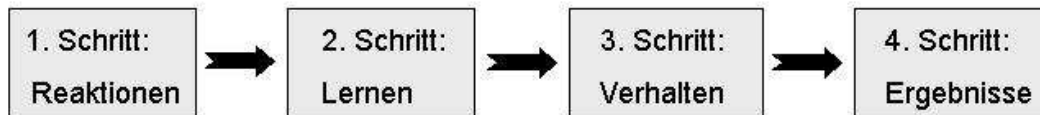


Abbildung 2-1. Schritte der Evaluation nach Kirkpatrick (z.B. 1996).

Sein Hauptanliegen war, Trainings- und Weiterbildungsverantwortliche auf die Bedeutung von Evaluation hinzuweisen. Die vier vorgeschlagenen Schritte (siehe Abbildung 2-1) sollten den komplexen Evaluationsprozess anschaulich aufgliedern und so zur Evaluation von Trainingsprogrammen bewegen.

Obwohl Kirkpatrick in seinen ursprünglichen Artikeln vier Schritte (*steps*) zur Evaluation nennt, wurde dieser Begriff in der nachfolgenden Forschung in Ebenen (*levels*) umbenannt und ihre Gesamtheit als das Vier-Ebenen-Modell von Kirkpatrick bezeichnet. Da Kirkpatrick selbst in späteren Arbeiten ebenfalls von ‚levels‘ spricht (z.B. Kirkpatrick, 1996), wird in der vorliegenden Arbeit ebenfalls von den vier Ebenen gesprochen. Diese sollen nun im Folgenden näher beschrieben werden.

Ebene 1: Reaktionen der Teilnehmer (reactions)

Unter Reaktionen sind nach Kirkpatrick keine Verhaltensreaktionen zu verstehen, sondern vielmehr die Zufriedenheit mit verschiedenen Aspekten der Maßnahme, etwa in Bezug auf Trainer, Trainingsinhalte und -methoden, zeitliche Aspekte und andere Rahmenbedingungen (Kirkpatrick, 1996, 1998). Für Kirkpatrick selbst stellen die Teilnehmerreaktionen ein Maß der Kundenzufriedenheit dar. Die Voraussetzungen für ein optimales Lernen sind ihm zufolge umso günstiger, je positiver die Reaktionen auf das Training ausfallen. Eine Messung auf dieser ersten Ebene erfolgt fast ausschließlich direkt am Ende einer Maßnahme durch einen Seminar-Feedbackbogen („paper-and-pencil“-Befragung).

Ebene 2: Lernen (learning)

Als einen zweiten Schritt betont Kirkpatrick (1996) die Bedeutung einer Evaluation auf der Lernebene, welche die Messung von erworbenem Wissen, verbesserten Fähigkeiten und/oder veränderten Einstellungen der Teilnehmer als Folge der Maßnahme beinhaltet.

Auch diese Ebene wird – meist zeitgleich mit der Erhebung der Teilnehmerreaktionen – am Ende der Maßnahme evaluiert. Inwiefern die einzelnen Inhalte der Veranstaltung gelernt wurden, kann durch einen Test oder eine Klausur zur Überprüfung des Wissensstands ermittelt werden. Empfohlen wird dabei eine Messung vor und am Ende der Maßnahme (Kirkpatrick, 1996). Die Differenz aus Vor- und Nachbefragung soll den Lernerfolg abbilden, der auf das Training zurückführbar ist. Verschiedene Autoren schlagen eine weitere Messung des Wissensstands zu späteren Zeitpunkten vor (z.B. Clement, 1982; Nork, 1991). Ist eine Vorher-Messung nicht möglich, sollte zur Überprüfung der Trainingseffekte eine Kontrollgruppe herangezogen werden – wobei ein Kontrollgruppen-Design mit Prä-Post-Messung den Idealfall darstellt (Kirkpatrick, 1996).

Ebene 3: Verhalten (behavior)

Auf dieser Ebene werden trainingsbedingte Verhaltensänderungen ermittelt, d.h. das Ausmaß, in dem die Teilnehmer nach der Maßnahme ihr Verhalten „on-the-job“ verändern. Diese Verhaltensänderung wird allgemein auch als Transfer bezeichnet, im Sinne einer Übertragung und Anwendung der gelernten Seminarinhalte auf das Verhalten am Arbeitsplatz (vgl. Alliger et al., 1997). Eine Evaluation auf dieser Ebene kann durch Verhaltensratings (Selbst- und/oder Fremdeinschätzung), Beobachtungen ‚on-the-job‘ etc. erfolgen.

Die Messung von Verhaltensänderung sollte erst mit einem gewissen Zeitabstand nach dem Training erfolgen, um den Trainees die Möglichkeit zur Umsetzung des Gelernten zu gewähren. Während Kirkpatrick (1996) von einem zeitlichen Abstand von mindestens drei Monaten zwischen Training und Messung ausgeht, ist für Sieber Bethke (2003) bereits nach 14 Tagen eine Messung möglich. Wie bei der Lernebene empfiehlt Kirkpatrick (1996) eine Prä-Post-Messung der Verhaltensebene mit Kontrollgruppensdesign.

Ebene 4: Ergebnisse (results)

Bei der Ebene der Unternehmensergebnisse steht die Erhebung von Kennzahlen im Vordergrund, die den Einfluss der Maßnahme auf Unternehmenskennzahlen widerspiegeln sollen. Als Beispiele hierfür können folgende Kennwerte dienen: Steigerung der Verkaufszahlen, Produktivitätssteigerung, Ergebnisverbesserung, Qualitätssteigerung, Kostensenkung, geringere Mitarbeiterfluktuation, geringere Absentismusrate.

Eine der bedeutendsten Arbeiten in Bezug auf das Modell von Kirkpatrick ist die Literaturübersicht von Alliger und Janak (1989), die mehrere in der Literatur zu findende implizite Annahmen überprüft, die dem Kirkpatrick-Modell seit seinem Erscheinen zugeschrieben werden. Neben den beiden Annahmen, eine höhere Ebene berge jeweils mehr Informationsgehalt als die vorangegangenen und die Erfüllung einer Ebene stelle die Bedingung für die nächsthöhere Ebene dar, ist eine dritte in der Literatur allgemein verbreitete Interpretation des Modells, alle Ebenen korrelierten positiv miteinander. Die Autoren geben allerdings zu bedenken, dass Kirkpatrick selbst (da eine entsprechende explizite Formulierung bei der ursprünglichen Beschreibung des stufenweise durchzuführenden Evaluationsprozesses fehlt) diese Annahmen wohl nicht beabsichtigt hat. Alliger und Janak (1989) nehmen des Weiteren an, dass diese Implikationen von den Forschern hineininterpretiert wurden, die sich seines Modells bedient haben (siehe auch Gülpen, 1996).

Als Konsequenz der ersten Annahme wird die vierte Ebene (im Hinblick auf ökonomische Aspekte) oftmals als die wichtigste angesehen, obwohl Kirkpatrick (1998) jede einzelne Ebene als wichtig sieht. Ihm zufolge sollte keine Ebene nur deswegen ausgelassen werden, weil die höhere als wichtiger betrachtet wird (Kirkpatrick, 1998). Andererseits betont Kirkpatrick in ebendieser Arbeit, dass der Evaluationsprozess mit jeder zusätzlichen Ebene komplizierter und kostenintensiver wird – und bedeutungsvoller und informativer. Dies ist dann richtig, wenn man auf allen Ebenen evaluiert. Es kommt daher nach Alliger und Janak (1989) darauf an, welches Ziel eine Maßnahme verfolgt: Eine Maßnahme, die lediglich zur Vermittlung von Informationen dienen soll (z.B. Informationen über die Organisation für Neueinsteiger oder Teilnehmer eines Trainee-Programms), wird kaum einen Einfluss auf Ergebnisebene haben, weshalb eine Evaluation auf dieser Ebene wenig sinnvoll ist.

Die zweite Annahme der Hierarchie untermauern Alliger und Janak in ihrer Arbeit (1989, S. 332) mit einem – allerdings unvollständigen – Zitat von Hamblin (1974), demzufolge „training leads to reactions which lead to learning which leads to changes in job behaviour which lead to changes in organizations (which lead to changes in the achievement of ultimate goals)“⁴ (Hamblin, 1974, S. 15). Es bleibt unerwähnt, dass sich Hamblin mit diesen Worten auf sein eigenes, als hierarchisch deklariertes Modell bezieht. Kirkpatrick

⁴ Der in der Arbeit von Alliger und Janak (1989) ausgelassene Teil des Zitats steht in Klammern.

selbst bezeichnete sein Modell zwar ursprünglich nicht explizit als hierarchisch, in neueren Arbeiten spricht jedoch auch er davon, dass jede Ebene „has an impact on the next level“ (Kirkpatrick, 1998, S. 19). Kirkpatrick sieht in positiveren Reaktionen auf das Training nicht nur eine gute Basis für den Wissenserwerb, sondern ebenso für die Umsetzung in Verhalten, und beides trägt zu Veränderungen auf Ergebnisebene bei.

Bezüglich der dritten Annahme, dass alle Ebenen positiv miteinander korrelieren, fanden Alliger und Janak (1989) nur eine geringe Anzahl an Studien, die überhaupt Korrelationen erwähnten. In Anbetracht der schwachen mittleren Korrelationen zwischen Reaktionen und Lernen ($r = .07$) sowie zwischen Reaktionen und Verhalten ($r = .05$) schlagen sie daher ein Modell vor, in dem die Reaktionen von den anderen Ebenen losgelöst sind, während direkte Zusammenhänge zwischen Lernen und Verhalten ($r = .13$), Lernen und Ergebnissen ($r = .40$) sowie zwischen Verhalten und Ergebnissen ($r = .19$) angenommen werden.

Neben dem vorgestellten Modell von Kirkpatrick gibt es für die Evaluation von PE-Maßnahmen noch eine Reihe weiterer Modelle (z.B. Hamblin, 1974; Warr, Bird & Rackham, 1970). Keines dieser Modelle erlangte jedoch die gleiche Bedeutung wie das von Kirkpatrick (Salas & Cannon-Bowers, 2001). Den Grund für den Erfolg und die Beliebtheit seines Modells sieht der Autor selbst in der Einfachheit und in der Praktikabilität (Bates, 2004; Kirkpatrick, 1996). Lediglich dem Modell von Phillips (2005), der dem Vier-Ebenen-Modell als fünfte Ebene den sogenannten Return on Investment (ROI) hinzufügt, kommt in der Praxis ebenfalls eine Bedeutung zu. In der Finanzsprache ist der ROI als Standardbegriff zur Bewertung von Kapitalinvestitionen bekannt. Dahinter steht die klare Erwartungshaltung, dass mit einer Investition ein Erlös aus eben diesem eingesetzten Kapital erfolgt (Phillips, 2005). Der Erfolg einer Maßnahme wird demnach dahingehend gemessen und beurteilt, ob sie sich finanziell auszahlt.

Darüber hinaus gibt es ein weiteres Modell, das im Bereich der Evaluation und Effektivitätsforschung von PE-Maßnahmen ebenfalls als grundlegend angesehen wird. So wie Kirkpatrick in seinem Modell erstmals mögliche Kriterien für eine Evaluation spezifizierte, entwickelten Baldwin und Ford (1988) ein Modell zur Beschreibung der Bedingungen, unter denen eine Übertragung der Trainingsinhalte in die Praxis optimal erfolgen kann.

2.3.2 Modell des Transfers nach Baldwin und Ford

Baldwin und Ford (1988) beziehen in ihrem Transfermodell als erste Autoren die Faktoren der Arbeitsumgebung ein, weshalb es im angloamerikanischen Raum als das erste Rahmenmodell für Transfer und Transferbedingungen gilt (Piezzi, 2002).

Das Wort Transfer ist vom lateinischen Verb *transferre* abzuleiten und bedeutet unter anderem ‚übertragen‘, ‚verschieben‘, ‚transportieren‘. Diese lässt an eine räumliche Bewegung von einem Ausgangspunkt zu einem Zielort denken, eine weitere Übersetzung bedeutet jedoch auch ‚etwas auf etwas oder jemanden übertragen‘: So wird z. B. der Transfer in der betrieblichen Weiterbildung als ein Prozess gesehen, durch den das in einem Lernfeld Gelernte in das betriebliche Funktionsfeld effektiv ein- und umgesetzt wird (Piezzi, 2002). Alliger et al. (1997) definieren Transfer als Übertragung und Anwendung im Seminar gelernter Inhalte auf das aktuelle Verhalten „on-the-job“. Für Praktiker ist genau dieser Schritt der Umsetzung wichtig – die Ergebnisse der Maßnahme sollen im Arbeitsumfeld sichtbar werden und sich auf Unternehmensebene auswirken.

Das Modell von Baldwin und Ford (1988) sieht dabei als Bedingung für erfolgreich verlaufenen *Transfer* einerseits die Generalisierung neu erlernter Inhalte auf den Arbeitsplatz und andererseits die Aufrechterhaltung des Gelernten über einen gewissen Zeitraum.

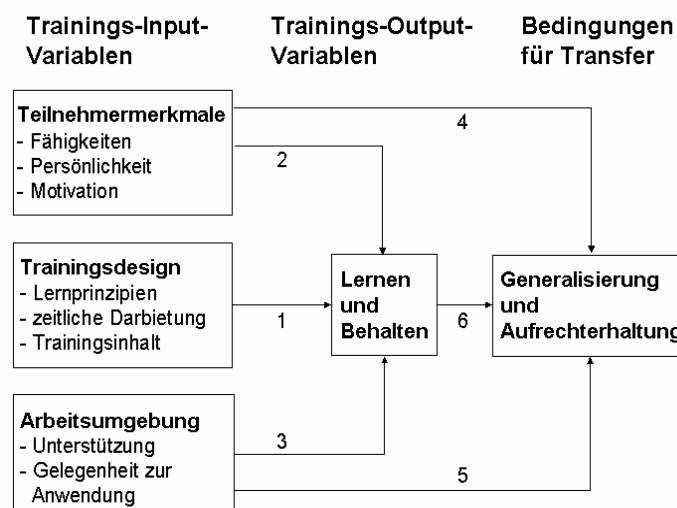


Abbildung 2-2. Transfermodell von Baldwin & Ford (1988).

Unter den Trainings-Input-Variablen (vgl. Abbildung 2-2) verstehen Baldwin und Ford folgende Faktoren:

- die *Merkmale der Teilnehmer* (individuelle Fähigkeiten, Persönlichkeitsmerkmale, Motivation)
- das *Trainingsdesign* (pädagogisch-psychologische Lernprinzipien, zeitliche Darbietung der Inhalte, Relevanz der Inhalte)
- und die *Arbeitsumgebung* (Unterstützung durch Kollegen oder Vorgesetzte, Gelegenheit zur Anwendung).

Die Trainings-Output-Variablen *Lernen* und *Behalten* werden als Ergebnisse des Trainings betrachtet, d.h. wie viel Trainingsstoff gelernt und nach Abschluss des Trainings behalten wird.

Die Transferleistung wird nach Meinung der Autoren direkt und indirekt durch die Trainings-Input- und Trainings-Output-Variablen beeinflusst. Diese Einflüsse werden anhand von sechs Verbindungen (siehe Abbildung 2-2) spezifiziert, die ihrerseits als grundlegend für das Verständnis des Transferprozesses gesehen werden (Baldwin & Ford, 1988). So verdeutlichen die Pfeile 1, 2 und 3 den direkten Einfluss aller drei Input-Variablen (*Teilnehmermerkmale*, *Trainingsdesign* und *Arbeitsumgebung*) auf die Output-Variablen *Lernen* und *Behalten*. Darüber hinaus wird für die *Teilnehmermerkmale* und die *Arbeitsumgebung* ein direkter Einfluss auf den *Transfer* angenommen (Pfeile 4 und 5): Mangelnde Motivation oder fehlende Unterstützung bei der Umsetzung der Inhalte am Arbeitsplatz kann beispielsweise eine Aufrechterhaltung selbst gut gelernter Inhalte verhindern. Der *Transfer* wird des Weiteren direkt durch das Ausmaß der beiden Output-Variablen bedingt: Nur das, was gelernt und auch eine Zeitlang behalten wird, kann im Endeffekt generalisiert werden (siehe Pfeil 6). Alle drei Input-Variablen üben außerdem über die Output-Variablen noch einen indirekten Einfluss auf den Transfer aus.

Seit dem Erscheinen des Modells von Baldwin und Ford hat sich im Forschungsbereich der Bedingungen für Transfer und Trainingseffektivität viel getan. Vor allem die Forschergruppe um Scott Tannenbaum befasste sich mit der Untersuchung von möglichen Einflussvariablen auf diejenigen Messkriterien, die bis dahin zur Evaluation von Trainingseffektivität eingesetzt wurden: die Kriterien von Kirkpatrick.

2.3.3 Modell der Trainingseffektivität nach Tannenbaum

Ein Hauptkritikpunkt am Modell von Kirkpatrick liegt z.B. nach Bates (2004) in dessen Einfachheit, da es weder Faktoren berücksichtigt, die den Transferprozess betreffen, noch die individuellen oder kontextuellen Faktoren, die allgemein auf die Trainingseffektivität einwirken können.

Den Ansatz von Kirkpatrick sehen Cannon-Bowers, Salas, Tannenbaum und Mathieu (1995) zwar als einen guten Versuch, die Multidimensionalität des Trainingskonzepts anhand der verschiedenen möglichen Trainingsergebnisse (die vier Schritte bzw. Ebenen) aufzuzeigen und zu untergliedern. Für ein vollständiges Verständnis der Trainingseffektivität ist es aber aus ihrer Sicht notwendig, die verschiedenen Effektivitätskomponenten zu identifizieren und zu untersuchen. Trainingseffektivität wird dabei als komplexes Phänomen gesehen (Cannon-Bowers et al., 1995), was die Forderung impliziert, ein besseres Verständnis über diese vielen verschiedenen Einflussfaktoren zu erhalten, welche die Effektivität nicht nur steigern, sondern sie auch beeinträchtigen können. Daher entwickelte die Forschungsgruppe um Tannenbaum ein integratives Rahmenmodell, das zum einen individuums-, trainings- und organisationsbezogene Charakteristika enthält und zum anderen den zeitlichen Aspekt berücksichtigt (siehe Abbildung 2-3).

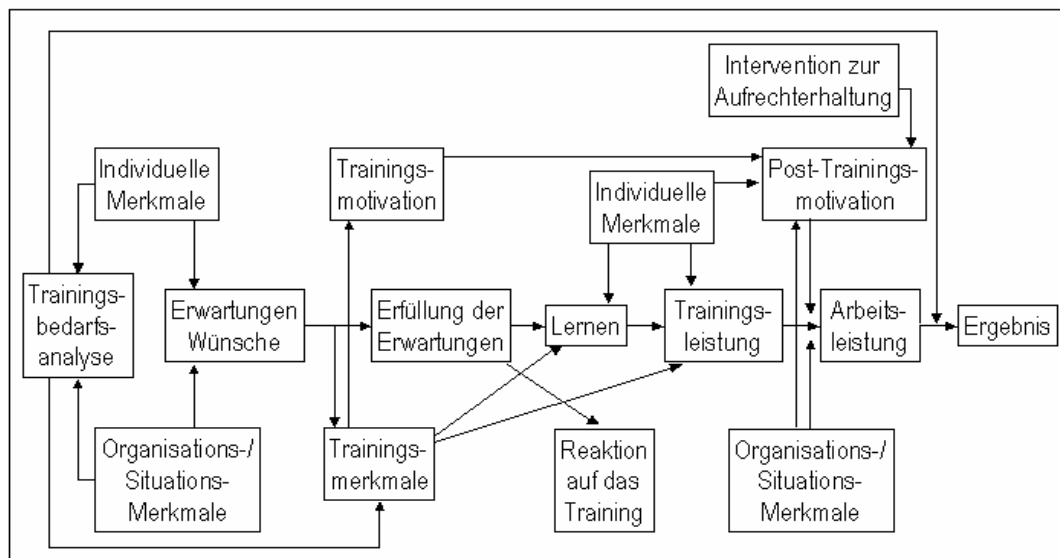


Abbildung 2-3. Rahmenmodell der Trainingseffektivität nach Tannenbaum (aus Höft, 2001).

Wie Abbildung 2-3 zeigt, werden die von Baldwin und Ford (1988) eingeführten Randbedingungen hier aufgenommen und hinsichtlich ihrer Wirkung vor, während und nach dem Training differenziert. Das Modell nimmt zwar die Kriterien nach Kirkpatrick auf, weicht aber in Anlehnung an die Ergebnisse von Alliger und Janak (1989) von den damit verbundenen Konzeptionen ab. Die Kriterien werden demnach nicht als hierarchisch angeordnete Ebenen betrachtet, vielmehr stehen die Teilnehmerreaktionen, wie in Abbildung 2-3 zu sehen ist, unverbunden zur zusammenhängenden Kette aus Lernen, Verhalten und Ergebnissen. Cannon-Bowers et al. (1995) nehmen für das Lernen, das Verhalten und die Ergebnisse eine jeweilige Beeinflussung von verschiedenen Trainingsprogrammkomponenten und Input-Faktoren an.

Die Verhaltensebene wird außerdem in *Trainingsleistung* und *Arbeitsleistung* „on-the-job“ aufgeteilt. Während die Trainingsleistung anzeigt, ob der Teilnehmer Gelerntes anwenden kann (im oder unmittelbar nach dem Training), stellt dieses Modell die Arbeitsleistung als die eigentliche Transferleistung dar: Hier zeigt sich, ob ein Teilnehmer in der Lage ist, das Gelernte außerhalb des Trainingsrahmens anzuwenden und auf ähnliche Situationen zu übertragen. Hinter dieser Aufteilung steht die Überlegung, dass ein Teilnehmer während oder direkt nach dem Training zwar durchaus gute Leistungen erzielen kann, diese dann aber an seinem Arbeitsplatz nicht mehr zeigt. Diese Diskrepanz kann sich beispielsweise dann zeigen, wenn ein zu großer Unterschied zwischen Trainings- und Arbeitsplatzbedingungen eine Übertragung des Gelernten „on-the-job“ verhindert. Nicht selten jedoch liegen ungünstige Voraussetzungen seitens des Unternehmens vor. Selbst ein gut konzipiertes und an sich wirksames Training wird scheitern, wenn den Teilnehmern durch äußere Umstände die Möglichkeit genommen wird, das Erlernete umzusetzen.

Cannon-Bowers et al. (1995) unterteilen die auf die Trainingseffektivität wirkenden Variablen in individuelle, trainingsbezogene und organisationale/ situationale Faktoren.

Unter *individuellen Einflussfaktoren* sind solche Faktoren zusammenzufassen, die ein Teilnehmer in die Trainingssituation mitbringt: Persönlichkeitsvariablen, Einstellungen, kognitive Fähigkeiten, demographische Variablen, Selbstwirksamkeit, Motivation sowie Erwartungen (Cannon-Bowers et al., 1995; Salas & Cannon-Bowers, 2001). Tannenbaum, Mathieu, Salas und Cannon-Bowers (1991) konnten in ihrer Prä-Post-Studie mit US-Navy-Rekruten zeigen, dass Trainees, deren Erwartungen erfüllt worden waren, nach dem

Training eine höhere akademische und körperliche Selbstwirksamkeit sowie eine höhere Motivation aufwiesen. Darüber hinaus übten sowohl Motivation als auch Selbstwirksamkeit einen Einfluss auf die Teilnehmerreaktionen aus. In einer anderen Untersuchung mit Studenten, die an einem 8-wöchigen Bowlingkurs der Universität teilnahmen, fanden Mathieu, Martineau und Tannenbaum (1993) einen Hinweis für die Relevanz der Selbstwirksamkeit: Diese korrelierte positiv mit der nachfolgenden Bowlingleistung sowie mit den Teilnehmerreaktionen. Hierbei ist jedoch anzumerken, dass die Rekruten lediglich durch eine 2-Item-Skala und die Studenten anhand einer 4-Item-Skala nach ihrer Zufriedenheit (*affective reactions*) befragt wurden. Cannon-Bowers et al. (1995) erhoben in ihrer Studie an Rekruten die Motivation durch einen Valenz-Instrumentalitäts-Erwartungs-Fragebogen in Anlehnung an Vroom (1964) und konnten zeigen, dass motivierte Rekruten nach dem Seminar positivere Teilnehmereinschätzungen (sowohl *affective* als auch *utility reactions*) abgaben. Außerdem führt die Motivation, an einem Training teilzunehmen und Neues zu erlernen, bei den höher motivierten Seminarteilnehmern zu einem besseren Lernen, zu einer besseren Leistung und letztlich zum erfolgreichen Abschluss der Maßnahme (Baldwin, Magjuka & Loher, 1991).

Auf der Seite der *trainingsbezogenen Charakteristika* gilt es beispielsweise, Trainingsmethode und -inhalt, Instruktionsstil, Übungen und das dazugehörige Feedback sowie die verwendeten Medien und Materialien zu berücksichtigen. Ebenso ist die zeitliche Darbietung der Inhalte bedeutsam (s.a. Baldwin & Ford, 1988).

Organisationale und situationale Einflüsse wirken vor allem vor und nach dem Training, da sie durch ihre Beschaffenheit einen direkten Einfluss auf Trainingserwartungen, Wünsche und Trainingsmotivation haben und somit einen indirekten Einfluss auf die Effektivität eines Trainings haben können. Unter organisationale Variablen fallen z.B. die Unternehmenskultur, -geschichte, -grundsätze, das Lern- und Transferklima und die Auswahl der Teilnehmer sowie deren Benachrichtigung. All dies kann bereits im Vorfeld die Erwartungen der Teilnehmer prägen und somit indirekt auf die Effektivität einwirken. Nach erfolgtem Training können Transfermotivation und Arbeitsleistung durch das herrschende Transferklima, die Unterstützung durch die Vorgesetzten sowie durch die gegebenen Ressourcen beeinflusst werden. So zeigten Tracey, Tannenbaum und Kavanagh (1995) die Bedeutung des Arbeitsumfeldes auf, wobei einerseits nicht nur das Transferklima direkte

Auswirkungen darauf hat, wie oft später gewünschtes Verhalten „on-the-job“ gezeigt wird (Transferleistung), sondern andererseits der Aspekt der „learning culture“. Dieser Begriff umschreibt, ob in einem Unternehmen eine „Kultur“ des ständigen Lernens im Sinne eines stetigen Verbesserungsprozesses herrscht.

Beim Transferprozess spielt jedenfalls neben der Unterstützung durch das soziale Umfeld auch eine Rolle, inwiefern den Trainees die Möglichkeit eingeräumt wird, neu gelerntes Wissen oder neu erworbene Fertigkeiten in ihrer Arbeitsumgebung einzusetzen. Sowohl das Modell von Noe (1986) als auch die Arbeit von Rouiller und Goldstein (1993) legen hierfür bei der Betrachtung transferrelevanter Faktoren eine Zweiteilung nahe. Ein *wohlwollendes Transferumfeld* (Noe, 1986; Noe & Schmitt, 1986) bzw. ein günstiges *Transferklima* (Rouiller & Goldstein, 1993) muss zum einen situationalen Faktoren umfassen – etwa zeitliche und materielle Ressourcen, Anwendungsmöglichkeit etc. Zum anderen erscheint das Ausmaß, in dem die soziale Arbeitsumgebung (z.B. Kollegen, Vorgesetzte) die Anwendung neu erworbener Kenntnisse bzw. Fertigkeiten unterstützt und fördert, ausschlaggebend für einen erfolgreichen Transfer zu sein. Facticeau, Dobbins, Russell, Ladd und Kudisch (1995) gehen noch einen Schritt weiter und unterteilen das soziale Umfeld neben Arbeitskollegen und direkten Vorgesetzten in untergeordnete Mitarbeiter sowie Top-Führungskräfte. Im Rahmen einer Trainingsbedarfsanalyse wurden Daten von 967 Managern erhoben, bei denen lediglich diejenigen Manager ausgewählt wurden, die bereits ein Training absolviert hatten. Die Studie hatte die Untersuchung des Einflusses allgemeiner Wahrnehmungen des Trainingsumfelds auf den wahrgenommenen Transfer zum Ziel. Neben der Unterstützung durch Kollegen zeigte sich hierbei als weiterer signifikanter Einfluss die Unterstützung durch die untergebenen Mitarbeiter. Im Gegensatz zu Noe (1986) und Rouiller und Goldstein (1993) zeigte sich jedoch kein Einfluss der Vorgesetztenunterstützung oder der situationalen Faktoren.

Im Vorfeld zum Training wird der *Bedarfsanalyse* eine zentrale Rolle zugesprochen. Alvarez et al. (2004) sehen in einer solchen Analyse einen wichtigen ersten Schritt hinsichtlich der Effektivität einer Maßnahme. Die Bedarfsanalyse ist ihrerseits ein Prozess, der folgende drei Schritte umfasst: die Organisationsanalyse (Welche Organisationsziele können durch die PE-Maßnahme erreicht werden? In welchem Bereich des Unternehmens ist eine Weiterbildung notwendig?), die Aufgabenanalyse (Welche Elemente könnten dem

Mitarbeiter helfen, seine Arbeit effektiver auszuführen? Was muss das Training beinhalten?) sowie die Personenanalyse (Welche Mitarbeiter benötigen eine Weiterbildung und in was?). Eine solche systematische Analyse liefert demnach nicht nur die Basis für die konzeptuelle Entwicklung und Durchführung von PE-Maßnahmen. Sie dient darüber hinaus als Anknüpfungspunkt für die Evaluation - durch den Vergleich der Inputs vor Beginn der Maßnahme mit den Ergebnissen nach deren Abschluss (Arthur et al., 2003). Trotz der Wichtigkeit einer Bedarfsanalyse erscheint in diesem Zusammenhang der Befund der Meta-Analyse von Arthur et al. (2003) bemerkenswert, wonach lediglich 22 von 397 Studien die Durchführung einer Bedarfsanalyse berichteten. Es wird jedoch von den Autoren angenommen, dass eine weitaus größere Anzahl der untersuchten Studien eine solche Analyse zwar beinhaltete, sie jedoch nicht beschrieben wurde.

Weitere postulierte Einflussgrößen sind z.B. das Vorwissen und die Praxiserfahrung, die ein Teilnehmer mitbringt (Piezzi, 2002). Bei Piezzi (2002) korrelierte die Höhe an Praxiserfahrung mit der mittleren Transferleistung negativ ($r = -.33, p < .05$), wonach sehr erfahrene Seminarteilnehmer den geringsten Transfer zeigten. Dieser negative Zusammenhang wird von Piezzi (2002) dahingehend interpretiert, dass sich bei einem hohen Transfer starke Veränderungen der Arbeitsleistung in den behandelten Seminarthemen zeigen würde, was bei den erfahrenen Teilnehmern jedoch nicht zutrifft. Je neuer die Thematik für die Seminarteilnehmer und entsprechend geringer ihre Erfahrung mit jobspezifischen Themen ist, umso wahrscheinlicher erscheint eine Veränderung im Arbeitsverhalten.

2.4 Evaluationsmodelle der PE und ihre Anwendung in der Praxis

Das Tannenbaum-Modell beschreibt auf sehr umfassende Weise die möglichen Faktoren, die einen Einfluss auf die Trainingseffektivität ausüben könnten. Obwohl in diesem Modell die von Kirkpatrick formulierten Ebenen enthalten sind, hat es sich vermutlich aufgrund dieser Komplexität noch nicht als Standard-Modell für Evaluationsmaßnahmen durchsetzen können. Der Vier-Ebenen-Ansatz bleibt somit weiterhin der bekannteste und am häufigsten angewandte Evaluationsansatz (z.B. Alliger et al., 1997; Kraiger et al., 1993; Salas & Cannon-Bowers, 2001).

2.4.1 Anwendung der einzelnen Ebenen nach Kirkpatrick

Betrachtet man diese „häufige Anwendung“ differenzierter, fällt die unterschiedliche Handhabung in der Verwendung der einzelnen Ebenen auf: Entgegen der ursprünglichen Empfehlung Kirkpatricks werden oftmals nicht alle Ebenen im Evaluationsprozess berücksichtigt. Wie Tabelle 2-2 verdeutlicht, zeigen sich im angloamerikanischen Raum und bei Unternehmen in Deutschland große Unterschiede bei der Anwendung der einzelnen Ebenen.

Tabelle 2-2. Anwendungshäufigkeit der Evaluationsebenen in den USA und in Deutschland

Ebene nach Kirkpatrick	USA ^a	Deutschland ^b
Reaktion	78%	69% (88%) ^c
Lernen	32%	39% (39%)
Verhalten	19%	33% (21%)
Ergebnisse	7%	42% (25%)

Anmerkungen: ^aBericht der American Society for Training and Development (ASTD; Van Buren & Erskine, 2002). ^bUmfrage bei 260 Unternehmen mit mind. 500 Mitarbeitern (Borchert & Rutschke, 2005). ^cIn Klammern zum Vergleich: Umfrage vom GABAL e.V. (Reiter, 2005) bei Seminaranbietern und Trainern.

Tabelle 2-2 zufolge evaluieren die meisten befragten Unternehmen mittels Teilnehmerreaktionen. Mit jeder weiteren Ebene sinkt jedoch die Anzahl an Unternehmen, die diese Ebene einsetzt. Vergleicht man die Häufigkeit, mit der z. B. die Reaktions- und die Verhaltensebene gemessen werden, wird der Abfall in angloamerikanischen Unternehmen von 78% auf 19% bzw. in deutschen Unternehmen von 88% auf 21% sehr deutlich. Eine Umfrage zur Bildungs-Effizienz vom GABAL e.V. (Reiter, 2005) bei Personalentwicklern/Seminaranbietern, woran sie den Erfolg von Seminaren messen, ergab dass diese Diskrepanz nicht daran liegt, dass die weiteren Ebenen als bedeutungslos angesehen werden (Tabelle 2-3).

Tabelle 2-3. Theoretische Bedeutung der Evaluationsebenen und ihre tatsächliche Umsetzung

Ebene nach Kirkpatrick	theoretische Relevanz	tatsächliche Messung
Reaktion	92%	88%
Lernen	97%	39%
Verhalten	86%	21%
Ergebnisse	79%	25%

Aus Tabelle 2-3 wird die Diskrepanz zwischen der Relevanz, die den jeweiligen Ebenen zugeschrieben wird, und der tatsächlichen Messung ab der zweiten Evaluationsebene deutlich. Während die Reaktionsebene von 92% aller Befragten als wichtig eingeschätzt wird und 88% eine solche Messung tatsächlich durchführen, wird auf Lern- und Verhaltensebene schon deutlich weniger oft evaluiert. Obwohl 97% dies auf Lernebene für wichtig halten, realisieren es nur noch 39%. Eine Evaluation auf Verhaltensebene halten 86% der befragten Seminaranbieter für bedeutsam, nur 21% setzen diese aber in die Tat um. Eine Evaluation auf vierter Ebene wird noch von 79% als wichtig angesehen, gemessen werden die Ergebnisse aber tatsächlich von 25% der Seminaranbieter.

Auch in der einschlägigen Fachliteratur lässt sich eine Diskrepanz in der Anwendungs- bzw. Untersuchungshäufigkeit der Ebenen aufzeigen. Von den untersuchten 201 Studien⁵ einer Literaturrecherche von Alliger und Janak (1989) befassten sich 149 Studien (74%) mit der Evaluation auf nur einer Ebene. Insgesamt 44 Studien (22%) untersuchten gleichzeitig zwei Ebenen, wohingegen in fünf Studien (2.5%) auf drei Ebenen gemessen wurde und nur drei Studien (1.5%) alle vier Ebenen berücksichtigten. Interessanterweise befasst sich die Mehrheit der 149 Studien, die nur eine einzige Ebene messen, nicht mit der Reaktionsebene, wie die Befunde aus der Praxis vermuten lassen könnten, sondern mit der Messung auf den Ebenen Lernen und Verhalten.

2.4.2 Hindernisse bei der Durchführung von Evaluation

Im Hinblick auf den hohen finanziellen Aufwand von PE-Maßnahmen überrascht der weitverbreitete Verzicht auf eine Evaluation. Nork (1991) zählt neben den Kosten eine Vielzahl anderer Gründe für diese Vernachlässigung von Evaluation auf: Glaube an die Wirksamkeit von Weiterbildungsmaßnahmen, fehlendes Evaluationsbewusstsein, Ängste der Beteiligten, Mangel an Ressourcen, Konzepten sowie Instrumenten zur Durchführung, Problem der Zurechenbarkeit und eine komplizierte Operationalisierung. Vor allem der Kostenaspekt ist neben dem zeitlichen Aufwand vermutlich der wichtigste Hinderungsgrund und wird von

⁵ In der Originalarbeit von Alliger und Janak (1989) wird in der entsprechenden Tabelle auf S. 336 eine Gesamtzahl von 203 Studien angegeben, nach Addieren der entsprechenden Studien erhält man jedoch 201 Studien.

Eichenberger (1990, S. 35) etwas plakativ, aber treffend unter dem Titel „Millionen für Bildung, Pfennige für Evaluierung“ subsummiert. Dieser gern zitierte Satz wird von Stiefel (1997, S. 27) bekräftigt, der von einer „Evaluierungsmisere“ in der Personalentwicklung spricht. Der potentielle Nutzen, den eine systematische Evaluation für ein Unternehmen bergen kann, wird dabei oftmals übersehen (Hertel, Orlikowski, Jokisch, Schöckel & Haardt, 2004). Hertel et al. (2004) sehen die Vorteile in einer Steigerung der Effektivität: Durch einen umfassenden Evaluationsprozess kann u.a. der Trainingsinhalt auf die Bedürfnisse der Teilnehmer angepasst werden (*customizing*), eine Erfolgskontrolle der Trainer durchgeführt werden und ein nachhaltiges Qualitätsmanagement sichergestellt werden.

Obwohl die Bedeutung einer umfassenden und systematischen Evaluation weitestgehend erkannt wird, bleibt ihre Umsetzung in der Praxis oftmals unvollständig (siehe Ergebnisse von Borchert & Rutschke, 2005). Dass jedoch die erste Kirkpatrick-Ebene fast immer eingesetzt wird, ergibt sich aus den offensichtlichen Vorteilen: Eine Reaktionsmessung ist leicht, unkompliziert und ohne großen finanziellen oder zeitlichen Aufwand durchführbar.

2.4.3 Reaktionen als „Happiness index“? – Kritik an Kirkpatrick

Die in der Praxis vorherrschende Evaluation auf der Reaktionsebene ist nicht unumstritten. Zum einen birgt eine unmittelbare Messung nach dem Seminar aufgrund der zeitlichen Nähe die Gefahr von Verzerrungen im Sinne zu optimistischer Einschätzungen (Konradt, Hertel & Behr, 2002; Nork, 1991). Eine solche „Positiv-Färbung“ der Ergebnisse kann durch den Einfluss des Trainers oder Einflüsse der Umgebung bzw. der Räumlichkeiten, der Trainingsatmosphäre etc. entstehen. Um dies zu vermeiden, erachtet Nork (1991) eine weitere Messung für sinnvoll, wenn die Teilnehmer wieder an ihren Arbeitsplatz und ihre gewohnte Umgebung zurückgekehrt sind. Eine zu positive unmittelbare Seminar-Bewertung könnte sich somit im Rückblick relativieren. Bei einer Trennung der Reaktionen in verschiedene Aspekte sind Alliger et al. (1997) im Hinblick auf die *utility reactions* der Ansicht, die Teilnehmer könnten hier zu Spekulationen verleitet werden, welchen zukünftigen Nutzen sie in der Maßnahme sehen. Sie erachten es daher als sinnvoll, weitere Reaktionsmessungen beispielsweise nach einem, drei oder sechs Monaten zu erheben (Alliger et al., 1997). Dann erst hätten die Trainees erlebt, welche Inhalte sie tatsächlich

umsetzen konnten und sind folglich besser in der Lage, zu beurteilen, ob und in welchem Ausmaß das Training für sie von Nutzen war

Auf der anderen Seite sieht Holton (1996) die Rolle der Teilnehmerreaktionen als eines der vier Evaluationskriterien als fraglich. Die von Alliger und Janak (1989) gefundenen fehlenden bzw. geringen gefundenen Korrelationen mit den anderen Ebenen verleiten ihn dazu, in der Berücksichtigung der Reaktionen eines der größten Schwächen des Vier-Ebenen-Modells zu sehen (Holton, 1996). Im Hinblick auf die Ergebnisse von Noe und Schmitt (1986), die ebenso wenig wie Alliger und Janak (1989) einen bedeutsamen direkten linearen Zusammenhang zwischen den Reaktionen und der Lernebene fanden, betrachtet er diese erste Ebene lediglich als einen Einflussfaktor auf das Lernen – nicht jedoch auf das Verhalten. Dadurch verliert die Reaktionsebene nach Holton (1996) den Status eines Trainingsergebnisses, weshalb er sogar für ihren Ausschluss aus den Evaluationsmodellen plädiert.

Ein weiterer Vorwurf an der Reaktionsebene besteht in der Undifferenziertheit ihrer Messung, wodurch die fehlenden bzw. geringen Korrelationen vermutet werden. Demgegenüber stehen jedoch die Befunde einer Untersuchung von Warr, Allan und Birdi (1999), bei der sich größere Zusammenhänge zwischen der Reaktionsebene und den anderen Ebenen zeigten. Die Teilnehmerreaktionen wurden dabei wie schon bei Warr und Bunce (1995) in die Facetten Zufriedenheit (*affective reactions*), Nutzen (*utility reactions*) und Schwierigkeit (*difficulty reactions*) unterteilt. In der Arbeit von 1999 kam zusätzlich die vierte Facette der Transfermotivation hinzu. Eine Ausdifferenzierung der Reaktionsebene in verschiedene Facetten befürworteten auch Morgan und Casper (2000), die in ihrer Analyse der Faktorstruktur gleich sechs Faktoren fanden (Zufriedenheit mit Trainer, Trainingsmanagement und Trainingsorganisation, Klausur, Anwendbarkeit/ Nutzen des Seminars, Material und Kursstruktur). Morgan und Casper (2000) heben die Bedeutung der Teilnehmerreaktionen insofern hervor, als dass sie u.a. als potentielle Prädiktoren für die weiteren Ebenen dienen können.

Diesen Aspekt hebt auch ein Befund der Meta-Analyse von Alliger et al. (1997) hervor. Hier differenzierten die Autoren in *affective reactions* und *utility reactions*, wobei die nutzenbezogenen Einschätzungen stärker als die affektiven Zufriedenheitseinschätzungen mit dem Verhalten korrelierten. Alliger et al. (1997) sehen dies als Folge der Spezifität der

utility reactions: Je spezifischer und verhaltensbezogener die nutzenbezogenen Einschätzungen ausformuliert werden, umso größer wird ihre Vorhersagekraft für eine spätere Transferleistung der Inhalte on-the-job. Eine ausbleibende bzw. niedrige Korrelation zwischen Reaktions- und Verhaltensebene ist demzufolge durch eine Inkonsistenz der Messungen auf diesen beiden Ebenen erklärbar. Zu dieser Ansicht kommen auch Ajzen und Fishbein (1977) in ihrem Review: „the relations between attitudes and behaviors tend to increase in magnitude as the attitudinal and behavioral entities come to correspond more closely in terms of their target and action elements“ (S. 911).

Ein weiteres Ergebnis von Alliger et al. (1997) ist die höhere Korrelation des wahrgenommenen Nutzens mit dem Verhalten ($r = .18$) im Gegensatz zum Lernen ($r = .07$). Alliger et al. (1997) diskutieren für dieses Ergebnis den Einfluss der Kenntnis eines Teilnehmers im Hinblick auf das Arbeitsumfeld, in das er später zurückkehren wird. Werden Teilnehmer also nach Seminarende nach dem wahrgenommenen Nutzen gefragt, könnten diese *utility reactions* durch das situationale Umfeld implizit beeinflusst werden. Nutzeneinschätzungen sind demzufolge bessere Prädiktoren oder Indikatoren für Transfer als für ein Ergebnis auf Lernebene (Alliger et al., 1997). Eine gute Erklärung für den Zusammenhang zwischen Nutzenreaktionen und Transfer sehen Alliger et al. (1997) in den Worten, „what we think is useful may correlate with what we use“ (S. 352).

Eine inhaltliche Überlappung der Evaluationskriterien (Ebenen) spiegelt sich dementsprechend in den Korrelationen wider. Zu beachten ist daher die Übereinstimmung zwischen dem Inhalt der Messebene und dem Inhalt des Fragebogens, der zur Messung dieser Ebene bestimmt ist. Objektive Messungen der Lernebene, z.B. in Form eines Wissenstests, geben wieder, inwiefern Teilnehmer vermitteltes Wissen direkt abrufen können. Das abgefragte Wissen (also der Inhalt der Messung) sollte sich dabei mit den tatsächlichen Inhalten des Seminars decken. Jedoch mangelt es an Überschneidungen mit der Verhaltensebene – anders als bei den antizipierten Nutzeneinschätzungen können reine Wissensergebnisse kein späteres Verhalten vorhersagen, weshalb hier geringe Korrelationen plausibel erscheinen.

Obwohl die Anzahl an Reaktionsfacetten in der Literatur durch die unterschiedlichen Operationalisierungen variiert (Brown, 2005), sollten Reaktionsmessungen neben den

affektiven Urteilen mindestens Bewertungen hinsichtlich des Nutzens beinhalten (Alliger et al., 1997; Warr et al., 1999; Warr & Bunce, 1995).

Neben den Teilnehmerreaktionen wird nach Kraiger et al. (1993) auch die Lernebene als multidimensional angesehen. Kraiger et al. (1993) gliedern die Ergebnisse auf der Lernebene auf in die sogenannten KSAs: Wissen (*knowledge*), Fertigkeiten (*skills*) und Einstellungen (*attitudes*), vgl. Abbildung 2-4.

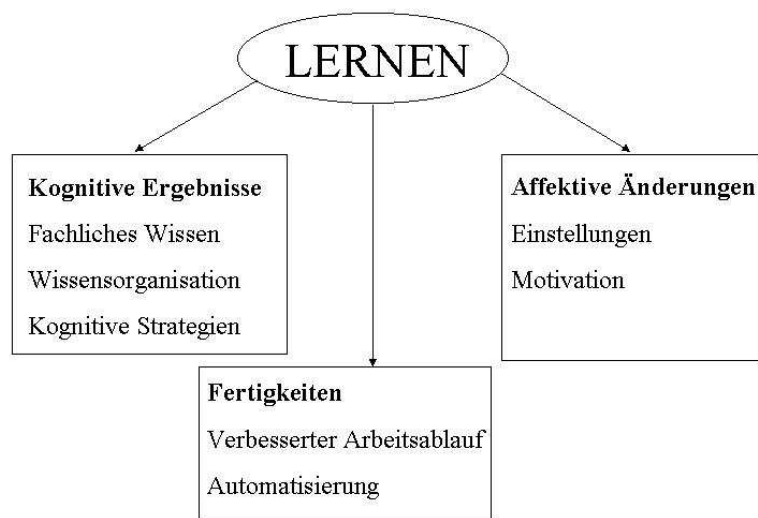


Abbildung 2-4. Komponenten der Lernebene nach Kraiger, Ford und Salas (1993).

Auf der Seite der *kognitiven Ergebnisse* sehen Kraiger et al. (1993) beispielweise ein verbessertes deklaratives Wissen, die Entwicklung bzw. Anwendung von adäquaten mentalen Strategien (zur Speicherung, Organisation und Anwendung von Wissen) sowie die Entwicklung von kognitiven Strategien und Funktionen zur Überprüfung und Regulierung von Leistung. Im zweiten Komplex steht die Entwicklung bzw. Verbesserung von Fertigkeiten für einen schnelleren und fehlerfreien Arbeitsablauf und die Automatisierung von Abläufen. Zuletzt wird die Entwicklung und Verinnerlichung angemessener Einstellungen angestrebt, etwa Veränderungen von Motivation, Selbstwirksamkeit, Zielsetzung etc. Einstellungen als Bestandteile eines Lernprozesses können somit als wichtige Voraussetzungen für Anpassungen im Verhalten bzw. im Handeln gesehen werden. Auch laut Piezzi (2002) kann in diesem Sinne eine nach dem Seminar gemessene Veränderung der Einstellung als Indikator für den Lernerfolg gelten.

2.4.4 Vorteile einer Reaktionsmessung zu mehreren Zeitpunkten

Clement (1982) sieht einen wichtigen Beitrag der Teilnehmerreaktionen darin, bei gegebenem Anlass neben dem unmittelbaren Seminar-Feedback weitere Reaktions-Messungen durchzuführen. Dieser Anlass besteht dann, wenn es darum geht, die Gründe für ausbleibende Erfolge auf den höheren Evaluationsebenen (Lernen, Verhalten, Ergebnisse) zu identifizieren. Die Aufmerksamkeit soll hierbei auf interferierende oder beeinflussende Variablen gelenkt werden, und „one good way to maintain this awareness is by getting additional reactions from the trainee [...] on why the training did or did not succeed“ (Clement, 1982, S. 184). Nork (1991) betont gleichfalls die Bedeutung einer nachgelagerten Reaktionsmessung, um z. B. zu positiv ausgefallene Einschätzungen zu relativieren. Zeitlich nachgelagerte Reaktionsmessungen können demzufolge im Sinne Norks (1991) dieselben Inhalte umfassen, um zu überprüfen, ob sich die Höhe der Einschätzungen verändert. Diese Einschätzungen können aber auch durch unterschiedliche Instrumente erfasst werden, um im Sinne Clements (1982) Informationen zu erhalten, die über das unmittelbare Feedback hinausgehen und die daher andere Reaktions-Facetten messen. Sieber Bethke (2003) schildert z.B. den Versuch eines Unternehmens, durch mehrfache Reaktionsmessungen mehr Informationen aus dem Feedbackbogen zu erhalten als über den Einsatz eines „happiness sheet“: Mit der sogenannten „Heißabfrage“ wurden die Teilnehmer direkt am Ende der Maßnahme nach ihren Eindrücken gefragt, solange diese noch frisch waren. Die „Warmabfrage“ wurde zwei bis vier Wochen nach der Maßnahme durchgeführt, es erfolgte also eine Beurteilung im Rückblick und mit mehr kritischer Distanz. Mit der „Kaltabfrage“ sollte letztlich der Frage nachgegangen werden, was denn nun vom Seminar tatsächlich an Verwertbarem mitgenommen worden ist, d.h. es sollte das Ausmaß an Transfer bestimmt werden. Leider geht Sieber Bethke (2003) nicht näher darauf ein, ob zusätzlich eine Evaluation auf den anderen Ebenen stattfand und welche Zusammenhänge sich durch diese zeitlich versetzten Reaktionsmessungen ergaben.

2.4.5 Bedeutung von Feedback im Evaluationsprozess

Mit der Aktionsforschung legte Lewin (1947) einen der Grundsteine für moderne Organisationsentwicklung. Ein zentrales Merkmal dieser Art von Forschung besteht in der implizierten Kooperationsbereitschaft zwischen Wissenschaft und Praxis (Gebert & v. Rosenstiel, 2002), denn es wird gemeinsam versucht, ein Problem zu lösen – immer im

Wechsel zwischen Aktion und Forschung. Die in der empirischen Sozialforschung übliche strikte Trennung zwischen Forschendem und Forschungsobjekt solle aufgehoben werden, damit beide ein einheitliches Handlungssystem bilden, in welchem Feedback eine wesentliche Rolle spielt.

Die Survey-Feedback-Methode bezeichnet dabei das systematische Sammeln von Daten über den Zustand einer Organisation bzw. eines Bereiches sowie die Rückmeldung (*feedback*) dieser Daten an alle Beteiligten des Veränderungsprozesses (v. Rosenstiel, Molt & Rüttinger, 2005). Lewin (1947) unterteilt diese sozialen Veränderungsprozesse, wie sie die Aktionsforschung vorsieht, in drei Phasen. In der erste Phase (*unfreezing*) gilt es, fehlerhafte Verhaltensweisen zu identifizieren und eine Veränderungsmotivation zu schaffen. Nach einer Momentaufnahme wird also Feedback über den Ist-Zustand gegeben, den es in Richtung erarbeiteter Ziele zu verändern gilt. Mit Phase 2 (*moving*) werden diese Veränderungen eingeführt, z.B. neue Verhaltensweisen erlernt, um dieses Verhalten dann in der dritten Phase (*refreezing*) auf dem neu erlernten Level zu stabilisieren und zu erhalten. Dabei ist nach Jöns (1997) Feedback als fundamentale Voraussetzung für individuelles Lernen anzusehen, denn auch im Rahmen eines Lernprozesses gilt es, Ist- und Sollzustand zu vergleichen und Entwicklungen zu verfolgen. Durch entsprechendes Feedback kann demnach vermittelt werden, ob und in welche Richtung weitergearbeitet werden sollte.

Auch im Modell von Miller, Galanter und Pribram (1973) wird solange verändert und getestet, bis ein gewünschter Zustand erreicht ist. So beschreiben sie etwa menschliches Verhalten mit Hilfe von Feedback-Schleifen in ihrem TOTE (Test-Operate-Test-Exit)-Modell, wonach jede Handlung einen Regelkreis durchläuft. Nach Bestimmung eines Handlungszieles wird zunächst in einem Test (T) überprüft, ob der Ist-Zustand vom gewünschten Zielzustand abweicht. Bei vorhandener Diskrepanz erfolgt eine Handlung (O). Ein weiterer Test (T) überprüft, ob der Zielzustand erreicht wurde. Ist dieser erreicht, ist die Verhaltenseinheit beendet (E). Auch innerhalb der Zielsetzungstheorie von Locke und Latham (1990), die sich mit der Wirkung von Zielen auf Motivation und Handeln -befasst, dient Feedback dazu, das Verhalten im Hinblick auf die gesetzten Ziele zu evaluieren. Laut Nadler (1979) kommt dem Feedback allgemein eine Orientierungs- bzw. Hinweisfunktion zu, oder aber es kann als Belohnung, als Lernfunktion oder Motivationsfunktion wirken.

Zusammenfassend ist Feedback, bezogen auf PE-Maßnahmen in Unternehmen, entsprechend sowohl für das Lernen einzelner Mitarbeiter als auch für das Lernen von Organisationen als soziale Gemeinschaft erforderlich. Dies impliziert für die Durchführung einer Evaluation, dass ihre Funktion nicht auf den Wirksamkeitsnachweis einer Maßnahme reduziert werden sollte, sondern dass diese Maßnahme als Ausgangspunkt für einen Lernprozess erkannt wird. In diesem Prozess, der sich an das Training anschließt, ist Feedback von großer Bedeutung, da sich durch entsprechende Messungen Aussagen über Veränderungen treffen lassen, Hindernisse bei der Umsetzung angesprochen und behoben werden können. Dabei ist neben den Trainees auch das Arbeitsumfeld miteinzubeziehen, also Kollegen, Vorgesetzte und die Personalverantwortlichen. Durch das Zusammenspiel von Aktion und Rückmeldung wird dafür gesorgt, dass innerhalb des Evaluationsprozesses nachhaltigere Veränderungen stattfinden als bei reinen Erfolgsüberprüfungen der Fall ist.

2.5 Schlussfolgerungen

Heutzutage erfolgen Evaluationen vorrangig in Anlehnung an das Modell von Kirkpatrick (1996). Ungeachtet der Kritik, die daran geäußert wurde (Bates, 2004; Holton, 1996), sieht eine Evaluation in der Praxis oftmals nur die Berücksichtigung der ersten Ebene vor (vgl. Borchert & Rutschke, 2005). Die Bedeutung der übrigen Ebenen wird dabei zwar ebenfalls gesehen, meist scheitert aber eine umfassendere Evaluation aus verschiedenen Gründen (Nork, 1991). Teilnehmerreaktionen sollten innerhalb eines Evaluationsprozesses jedoch aus mehreren Gründen nicht unterschätzt werden. Ihr größter Vorteil liegt in der einfachen und in jeder Hinsicht ökonomischen Messung. Darüber hinaus lassen sich aus der Literatur verschiedene weitere mögliche Vorteile ableiten: Eine entsprechend differenziertere Messung, d.h. bei einer Aufteilung in affektive und nutzenbezogene Reaktionen (Alliger et al., 1997; Warr & Bunce, 1995), lässt höhere Zusammenhänge zu den anderen Ebenen erwarten – v.a. zur Verhaltensebene, deren besondere Relevanz im darin abgebildeten Transfer liegt. Da zu verschiedenen Zeitpunkten unterschiedliche Aspekte relevant sein können, empfiehlt Clement (1982) die Erhebung verschiedener Reaktionsfacetten, wodurch sich wertvolle Hinweise auf mögliche Hindernisse oder Schwierigkeiten finden lassen (s. a. Nork, 1991; Sieber Bethke, 2003). Außerdem kann eine mehrfache Messung von Reaktionen nach Clement (1982) Anhaltspunkte darüber geben, warum sich auf der einen

oder anderen Ebene Schwierigkeiten ergeben. Von Morgan und Casper (2000) werden Teilnehmerreaktionen sogar als Prädiktoren für die höheren Ebenen angesehen.

Trotz dieser positiven Befunde in Bezug auf die Reaktionsebene sollte sich eine Evaluation nicht nur auf die Messung dieser Ebene beschränken, wie es überwiegend in der Evaluationspraxis geschieht (Borchert & Rutschke, 2005). Wie Kirkpatrick (1996) formuliert, sind positive Reaktionen auf eine Maßnahme zwar sehr förderlich für nachfolgenden Lernerfolg, sie sind allerdings dafür keine Garantie. Umgekehrt erscheint es im Zuge einer umfassenden Evaluation nicht sinnvoll, die Reaktionsebene zu umgehen und z.B. ausschließlich das Verhalten (Transferleistung) zu erfassen. Bleibt eine Transferleistung tatsächlich aus, fehlen bei diesem einzelnen Befund mögliche Hintergrundinformationen darauf, weshalb die Inhalte nicht umgesetzt werden können. Über unmittelbare Reaktionsmessungen lassen sich zumindest Hinweise gewinnen, wie die Maßnahme akzeptiert und ob sie als nützlich empfunden wurde, ob es Schwierigkeiten bei der Seminardurchführung oder Erarbeitung der Themen gab, ob der Trainer die Inhalte gut vermitteln konnte etc.

Für die Evaluation einer PE-Maßnahme reicht es weiterhin nicht aus, nur die Ebenen des Kirkpatrick-Modells zu berücksichtigen. Es muss mindestens eine Ausdifferenzierung der Reaktionen erfolgen und idealerweise auch der Lernebene in Wissen, Fertigkeiten und Einstellungen. Auch muss der Vielzahl an Variablen, die im Hinblick auf die Effektivität bei Trainingsprozessen einwirken, Rechnung getragen und miteinbezogen werden (Cannon-Bowers et al., 1995). Die vorliegende Untersuchung berücksichtigt deswegen einerseits die Rolle der Reaktionsebene, in dem diese zu mehreren Zeitpunkten und differenziert gemessen wurde. Andererseits wurden zusätzlich Einflussvariablen individueller, organisationaler und situationaler Art erhoben, die nicht nur die Ergebnisse auf Lern- und Verhaltensebene beeinflussen können, sondern bereits auch die Reaktionsebene.

Durch die Online-Durchführung bestand die Möglichkeit, den Seminarteilnehmern direkt im Anschluss an die jeweilige Befragung einige Ergebnisse individuell zurückzumelden, wodurch Vergleiche der Ergebnisse zu den anderen Zeitpunkten gezogen werden konnten. Trotz der erläuterten Bedeutung von Feedback innerhalb eines Evaluationsprozesses wurde in der vorliegenden Untersuchung aufgrund der hohen Komplexität der Daten (u.a. durch die mehrmaligen Messungen und die verschiedenen Ebenen) davon abgesehen, diese Thematik als Fragestellung zu behandeln.

3 Fragestellungen und Hypothesen

Ziel der vorliegenden Arbeit ist es, neue Erkenntnisse über die oftmals kritisierte Reaktionsebene zu gewinnen. Die im vorigen Kapitel dargestellten Befunde zum Nutzen der Teilnehmerreaktionen weisen daraufhin, dass diese unter Berücksichtigung bestimmter Bedingungen durchaus einen Beitrag zur Evaluation des Erfolgs einer Maßnahme leisten können.

Zusammenfassend lässt sich aus der gesichteten Literatur ableiten, dass die Erfassung differenzierter Einschätzungen als sinnvoll erachtet wird (z. B. Warr & Bunce, 1995) und eine mehrfache Messung dieser Ebene ebenfalls befürwortet wird (Clement, 1982; Nork, 1991). Zudem legen die Arbeiten rund um das Trainingseffektivitätsmodell von Tannenbaum (Cannon-Bowers et al., 1995) eine Berücksichtigung von organisationalen, individuellen und trainingsbezogenen Faktoren nahe.

Auf der Basis des im Vorfeld dargestellten Forschungsstandes und der bisher diskutierten Theorie sollen die im Folgenden beschriebenen Fragestellungen untersucht werden.

3.1 Korrelationen zwischen den Evaluationsebenen nach Kirkpatrick

Im Kontext des Modells von Kirkpatrick (1996) und den in der Literatur berichteten schwachen Korrelationen (z. B. Alliger & Janak, 1989) soll mit dieser ersten Fragestellung untersucht werden, wie die hier betrachteten Ebenen 1 bis 3 zusammenhängen, wenn die Reaktionsebene a) differenziert und b) zu mehreren Zeitpunkten gemessen wird. Gemäß der Annahme, die niedrigen Korrelationen der Reaktionsebene mit den weiteren Ebenen seien ein Ergebnis undifferenzierter Reaktionsmessungen (z.B. Warr & Bunce, 1995), müssten sich im Gegensatz zu einer allgemeinen Zufriedenheitsmessung höhere Zusammenhänge zwischen den differenzierten Messungen (bspw. der Anwendbarkeit) und der Lern- und Verhaltensebene zeigen. Dem Vorschlag von Clement (1982) folgend werden in dieser Untersuchung die Reaktionen zu insgesamt drei Zeitpunkten gemessen, wobei diese Einschätzungen unterschiedliche Aspekte darstellen und sich bis auf wenige Ausnahmen voneinander unterscheiden. Die erste Messung (t1) erfolgt mittels des unmittelbaren Seminar-Feedbackbogens, die zweite Messung (t2) nach ca. zwei Wochen und die dritte Messung

(t3) wird etwa drei Monate nach Seminarende durchgeführt. Zu dieser Fragestellung werden folgende Hypothesen formuliert:

Hypothese 1-A:

Es besteht ein positiver Zusammenhang zwischen den Kennwerten der Reaktionsebene und den Kriterien der Lernebene.

Hypothese 1-B:

Es besteht ein positiver Zusammenhang zwischen den Kennwerten der Reaktionsebene und den Kriterien der Verhaltensebene.

Des Weiteren wird in diesem Kontext auch der Zusammenhang zwischen Lern- und Verhaltensebene untersucht, und folgende Nebenhypothese dazu aufgestellt.

Nebenhypothese N-1:

Es besteht ein positiver Zusammenhang zwischen den Kennwerten der Lernebene und den Kriterien der Verhaltensebene.

3.2 Vorhersage des Seminarerfolgs anhand der Reaktionen

Ihrem (Ver-)Ruf als „happiness index“ steht die Annahme gegenüber, Teilnehmerreaktionen hätten doch einen gewissen Nutzen – und zwar möglicherweise als Prädiktoren für die anderen Ebenen (z.B. Morgan & Casper, 2000). Es stellt sich daher die Frage, inwiefern die Messungen der Reaktionsebene tatsächlich als Prädiktoren für die Ergebnisse auf den Ebenen Lernen und Verhalten herangezogen werden können. Früheren Befunden zufolge (z.B. Alliger & Janak, 1989) korrelieren *utility reactions* stärker mit der Verhaltensebene als *affective reactions*. Entsprechend sollen aus den hier vorgenommenen drei Reaktionsmessungen neben der Angabe von *affective reactions* (Durchführung) vor allem die *utility reactions* (Anwendbarkeit, tatsächliche Anwendung der Inhalte, Nutzen des Seminars und Umsetzung persönlicher Ziele) als Prädiktorvariablen verwendet werden. Zu dieser Fragestellung werden folgende Hypothesen aufgestellt:

Hypothese 2-A:

Die Reaktionsmessungen Durchführung (t1), Anwendbarkeit (t1), Anwendung (t2), Nutzen (t2), Anwendung (t3), Nutzen (t3) und Umsetzung persönlicher Ziele (t3) tragen zur Vorhersage der Ergebnisse auf der Lernebene bei.

Hypothese 2-B:

Die Reaktionsmessungen Durchführung (t1), Anwendbarkeit (t1), Anwendung (t2), Nutzen (t2), Anwendung (t3), Nutzen (t3) und Umsetzung persönlicher Ziele (t3) tragen zur Vorhersage der Ergebnisse auf der Verhaltensebene bei.

3.3 Auswirkung von Einflussgrößen auf die verschiedenen Ebenen

Es ist von einer gewissen Variabilität der Teilnehmereinschätzungen sowie der Ergebnisse auf Lern- und Verhaltensebene auszugehen. In Anlehnung an das Tannenbaum-Modell (Cannon-Bowers et al., 1995) soll daher überprüft werden, ob die Einschätzungen der Reaktionsebene sowie die Ergebnisse auf Lern- und Verhaltensebene einem systematischen Einfluss von individuellen sowie organisationalen bzw. situationalen Variablen (vgl. Cannon-Bowers et al., 1995; Mathieu, Tannenbaum & Salas, 1992; Tannenbaum et al., 1991) unterliegen.

Verschiedene Autoren (z.B. Cannon-Bowers et al., 1995; Baldwin et al., 1991) zeigten die Relevanz von Motivation bei einem Trainingsprozess auf. Demnach erzielten motivierte Seminarteilnehmer bessere Werte im Lernen und in der Arbeitsleistung, d.h. im Transfer. Ähnliches zeigte sich für die Selbstwirksamkeit (Tannenbaum et al., 1991), weshalb für diese beiden individuellen Einflussvariablen folgende Hypothesen benannt werden.

Hypothese 3-A:

Hoch motivierte Teilnehmer unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von weniger motivierten Teilnehmern.

Hypothese 3-B:

Teilnehmer mit einer hohen Selbstwirksamkeit unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit geringerer Selbstwirksamkeit.

In der Bedarfsanalyse wird eine zentrale Voraussetzung für den späteren Erfolg einer Trainingsmaßnahme gesehen (Alvarez et al., 2004). Da eine solche Analyse nicht durchgeführt werden konnte, sollte statt dessen geprüft werden, welchen Effekt die Höhe des subjektiv empfundenen Bedarfs, den die Teilnehmer für die durchgeführte Maßnahme sahen, auf die erhobenen Ebenen haben.

Hypothese 3-C:

Teilnehmer, die für sich selbst einen hohen subjektiven Bedarf für die Maßnahme sehen, unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit mittlerem bis geringem Bedarf.

Für eine erfolgreiche Übertragung der erlernten Inhalte auf die Arbeitssituation zeichnen verschiedene organisationale und situationale Faktoren verantwortlich, die Rouiller und Goldstein (1993) im sogenannten Transferklima zusammenfassen. Darunter fällt etwa die Anwendungsmöglichkeit der Seminarinhalte im tatsächlichen Arbeitsumfeld (wie z.B. das Ausmaß an Kundenkontakt als Indikator für die Anwendung von Vertriebstechiken) oder aber die Transferunterstützung, die ein Trainee nach der Rückkehr an den Arbeitsplatz von Seiten seiner Kollegen und Vorgesetzten erhält. Doch nicht nur das Verhaltensebene, also die Transferleistung kann somit begünstigt oder gehemmt werden, Facticeau et al. (1995) zufolge kann die Wahrnehmung der Unterstützung des Arbeitsumfeldes auch die Teilnehmereinschätzungen beeinflussen. Eine weitere berücksichtigte Variable ist die Vorerfahrung, die die Teilnehmer aufgrund ihrer Tätigkeit mitbringen (Piezzi, 2002). Folgende Hypothesen werden in Bezug auf organisationale/ situationale Variablen formuliert.

Hypothese 3-D:

Teilnehmer mit sehr viel Kundenkontakt (Anwendungsmöglichkeit) unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit mittleren oder nur wenig Kundenkontakt.

Hypothese 3-E:

Teilnehmer mit Vorerfahrung unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern ohne Vorerfahrung.

Hypothese 3-F:

Teilnehmer mit einem positiven Transferklima (Unterstützung durch Arbeitsumfeld sowie durch Führungskräfte) unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit einem negativen bzw. einem uneinheitlichen Transferklima.

Da in der Evaluationspraxis vorwiegend auf der Reaktionsebene evaluiert wird (vgl. Borchert & Rutschke, 2005), sind oft vor allem diese Messungen Basis weiterer Entscheidungen. Es soll daher abschließend geprüft werden, inwiefern sich Teilnehmer mit positiven bzw. negativen Bewertungen im unmittelbaren Seminar-Feedbackbogen auch entsprechend im Verlauf der Messungen auf Lern- und Verhaltensebene besser bzw. schlechter abschneiden. Analog zur Trennung in *affective* und *utility reactions* soll hier sowohl der Einfluss der allgemeinen Seminarbewertung (im Sinne allgemeiner affektiver Zufriedenheit mit dem Seminar) als auch der Einfluss einer spezifischen Seminarbewertung (im Sinne einer Anwendungs-Einschätzung der behandelten Themen) untersucht werden. Es werden daher folgende Hypothesen aufgestellt.

Hypothese 3-G:

Teilnehmer mit einer positiven allgemeinen Seminarbewertung (allgemeine Zufriedenheit mit der Veranstaltung) unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit einer weniger positiven allgemeinen Bewertung.

Hypothese 3-H:

Teilnehmer mit einer positiven spezifischen Seminarbewertung (antizipierte Anwendbarkeit der behandelten Themen) unterscheiden sich in den Kennwerten der Reaktions-, Lern- sowie Verhaltensebene von Teilnehmern mit einer weniger positiven spezifischen Bewertung.

4 Methodische Umsetzung

In der deutschen Niederlassung eines Elektronik- und Telekommunikationskonzerns wurde im Rahmen einer umfassenden Qualifizierungsmaßnahme ein Vertriebstraining durch einen internationalen Trainings- und Weiterbildungsanbieter durchgeführt. Dabei nahm die Abteilung „Global Services“ des Unternehmens an dieser Maßnahme teil. Erklärtes Ziel dieser Maßnahme lag für das Unternehmen darin, die Verkaufsfertigkeiten der Service-Mitarbeiter zu verbessern und deren Fokus stärker als bisher auf den Vertriebsaspekt innerhalb der Servicetätigkeit auszurichten. In Zusammenarbeit mit dem Trainingsanbieter, der Personalabteilung des Unternehmens und dem Institut für Begleitforschung als Projektpartner für die externe Evaluation sollte diese Qualifizierungsmaßnahme durch ein umfassendes Evaluationsprojekt begleitet werden. Hierfür wurden Daten der Seminarteilnehmer sowie der entsprechenden Führungskräfte und einer Kontrollgruppe erhoben. Ein Teil der in diesem Projekt erhobenen Daten wurde für die vorliegende Arbeit verwendet.

4.1 Stichprobenbeschreibung

Zur Beschreibung der Stichprobe wurden verschiedene biographische und soziodemographische Daten erfragt. Neben Angaben zu Geschlecht, Alter, Familienstand, Schulabschluss und Berufsausbildung wurden noch folgende Daten zur Arbeitssituation der Teilnehmer sowie der Kontrollgruppe erhoben: Dauer im Unternehmen, Dauer in derzeitiger Tätigkeit und Titel bzw. Aufgabenbereich innerhalb des Unternehmensbereichs. Zusätzlich sollten die Teilnehmer angeben, wie sie ihre berufliche Vorerfahrung im Vertrieb einstufen, wie sie ihr eigenes Vertriebspotenzial einschätzen und ob sie bereits in der Vergangenheit an einer Vertriebschulung teilgenommen haben.

4.1.1 Seminarteilnehmer (Trainees)

Obwohl insgesamt $N = 72$ Service-Mitarbeiter (7 Frauen und 65 Männer) am dreitägigen Vertriebstraining teilnahmen, mussten für die Auswertung mehrere Datensätze ausgeschlossen werden. Sechs Seminarteilnehmer bearbeiteten neben der Basisbefragung nur den Seminar-Feedbackbogen, weshalb ihre Daten für die weitere Auswertung nicht heran-

gezogen werden konnten. Bei vier Trainees, die die ersten beiden Befragungen bearbeitet hatten, konnten die Datensätze nicht zusammengefügt werden. Insgesamt 15 Trainees hatten nur den Seminar-Feedbackbogen bearbeitet und fielen für die weitere Analyse ebenfalls aus. Der Ausfall bei den Folgebefragungen lässt sich hauptsächlich durch die Überschneidung des Bearbeitungszeitraums der beiden Nachbefragungen mit den Urlaubszeiten der Trainees erklären.

Durch den Ausfall reduzierte sich die endgültige Stichprobe der Seminarteilnehmer auf $N = 47^6$. In die Auswertung gingen die Daten von insgesamt 5 weiblichen (11%) sowie 42 männlichen (89%) Trainees ein. Die Trainees waren zwischen 26 und 54 Jahre alt, wobei das Durchschnittsalter bei 38 Jahren lag ($M = 37.81$, $SD = 4.92$). Die Mehrzahl besaß einen akademischen Abschluss (74%), fünf Seminarteilnehmer (11%) hatten eine Berufsausbildung absolviert und sieben Seminarteilnehmer (15%) machten keine Angabe.

Die Unternehmenszugehörigkeit der Trainees betrug durchschnittlich ca. 9 Jahre ($M = 8.50$, $SD = 4.15$). In ihrem derzeitigen Job waren sie im Durchschnitt ca. 4 Jahre tätig ($M = 3.88$, $SD = 3.46$). Von den Trainees gaben fünf (10%) an, eine reine Vertriebstätigkeit auszuführen und 16 (34%) gaben an, einen Anteil an Vertriebsaufgaben in ihrer derzeitigen Tätigkeit zu haben. Insgesamt 22 Trainees (47%) berichteten, keinerlei Vertriebstätigkeit in ihrem Job auszuüben. Von vier Trainees (9%) fehlte diese Angabe. Ein ähnliches Bild zeigte sich bei der Frage nach einer früheren Teilnahme an einer Vertriebschulung: Vier Trainees (8%) gaben an, an einer reinen Vertriebschulung teilgenommen zu haben, 12 (26%) berichteten über die Teilnahme an einer Schulung mit Vertriebsanteilen, wohingegen 28 Trainees (60%) noch nie an einer Vertriebschulung teilgenommen hatten. Drei Trainees (6%) machten diesbezüglich keine Angabe.

4.1.2 Kontrollgruppe

Neben der Stichprobe der Seminarteilnehmer konnte noch eine zweite Stichprobe von $N_{KG} = 21$ Mitarbeitern als Kontrollgruppe gewonnen werden. In der vorliegenden Untersuchung war es nicht möglich, den Wissenstand der Trainees vor dem Training zu erfassen. Zur Überprüfung des Trainingseffekts auf dieser Ebene wird daher die Kontrollgruppe herangezogen, die den Wissenstest ebenfalls bearbeitete, ohne am Seminar teilgenommen

⁶ In Kapitel 4.5.2.3 soll geprüft werden, ob hier ein Selektionseffekt vorliegt.

zu haben. Dabei sollte diese untrainierte Gruppe eine signifikant *geringere* Anzahl an richtigen Antworten erzielen als die Trainees, die diese Inhalte im Training durchgenommen hatten.

Die Mitarbeiter der Kontrollgruppe wurden aufgrund ihres ähnlichen Tätigkeits- und Aufgabenfeldes ausgewählt, obwohl sie nicht aus demselben Service-Bereich stammten wie die Seminarteilnehmer. Eine Teilnahme an diesem oder einem anderen Vertriebsstraining war für diese 21 Mitarbeiter zum Zeitpunkt der Untersuchung nicht vorgesehen.

Die Stichprobe der Kontrollgruppe bestand aus 5 Frauen (23%) und 17 Männern (77%), die zwischen 31 und 50 Jahre alt waren ($M = 38.05$, $SD = 4.85$). 15 Personen (68%) gaben an, einen akademischen Abschluss zu haben, sechs Personen (27%) besaßen eine berufliche Ausbildung und eine Person machte diesbezüglich keine Angabe. Die durchschnittliche Zugehörigkeit zum Unternehmen betrug bei den Mitarbeitern der Kontrollgruppe ca. 10 Jahre ($M = 9.50$, $SD = 3.65$), im derzeitigen Job waren sie seit ca. fünf Jahren tätig ($M = 4.66$, $SD = 2.91$). Insgesamt gaben vier Mitarbeiter der Kontrollgruppe (18%) an, eine reine Vertriebstätigkeit auszuführen. Fünf Mitarbeiter (23%) berichteten, in ihrer Tätigkeit Vertriebsanteile zu haben, während 10 Mitarbeiter (45%) berichteten, keine Vertriebstätigkeit innezuhaben. Drei Mitarbeiter (14%) machten hierüber keine Angabe. Zwei Mitarbeiter (9%) hatten bereits an einer Vertriebschulung teilgenommen, vier Mitarbeiter (18%) berichteten über eine Schulung mit Vertriebsanteilen, 14 Mitarbeiter (64%) hatten noch an keiner Vertriebschulung teilgenommen und von zwei Mitarbeitern (9%) fehlte diese Angabe.

4.1.3 Führungskräfte

Parallel zu den Selbsteinschätzungen der Seminarteilnehmer sollten Fremdeinschätzungen durch die jeweiligen Vorgesetzten erhoben werden, weshalb im Verlauf der Datenerhebung insgesamt 17 Führungskräfte angeschrieben wurden. Die Anzahl an einzuschätzenden Trainees variierte dabei je nach Führungskraft zwischen 1 und 17. Um den Aufwand für die Führungskräfte gering zu halten, musste auf die Fremdeinschätzung der Mitarbeiter der Kontrollgruppe verzichtet werden.

4.1.4 Kontaktaufnahme

Der Kontakt der Diplomandin zu den Seminarteilnehmern, zur Kontrollgruppe sowie zu den Führungskräften verlief über E-Mail. Alle Einladungen zu den Online-Befragungen sowie die Erinnerungen wurden auf diese Weise verschickt. Jede E-Mail wurde individuell verfasst und berücksichtigte eine eventuell vorangegangene Korrespondenz mit den Studienteilnehmern. In den E-Mails wurden spezifische Informationen (Inhalt und Bearbeitungsdauer) sowie die Zugangsdaten zur jeweiligen Online-Befragung mitgeteilt. Beispiele für die Einladungs-E-Mails an Trainees, Kontrollgruppe und Führungskräfte sowie die beigefügten Informationen finden sich in Anhang A, S. 3ff.

Die Teilnahme an den Befragungen war freiwillig. In einer per E-Mail verschickten Bekanntmachung wies der Personalverantwortliche vor Beginn des Evaluationsprojektes alle beteiligten Mitarbeiter und Führungskräfte auf die Bedeutung des Projekts für das Unternehmen hin und warb um Unterstützung und möglichst hohe Partizipation.

4.1.5 Datenschutz

Für die Befragungen erhielten Seminarteilnehmer und Kontrollgruppe neben den allgemeinen Zugangsdaten (Benutzername und Passwort) individualisierte Codes. Die Führungskräfte erhielten analog dazu pro einzuschätzenden Mitarbeiter einen Code. Dadurch wurde die Zuordnung der Daten der einzelnen Befragungen zum jeweiligen Teilnehmer unter Wahrung des Datenschutzes ermöglicht. Zugeteilt wurden die Codes durch die Diplomandin, und nur sie und der Projektleiter vom Institut für Begleitforschung kannten die persönliche Zuordnung. Es wurden keine persönlichen Daten oder Codes an das Unternehmen oder den Seminaranbieter weitergegeben. Die Rohdaten der Untersuchung verblieben beim Institut für Begleitforschung und bei der Diplomandin. Eine Rückmeldung der Ergebnisse an das Unternehmen sowie den Seminaranbieter erfolgte aus Datenschutzgründen ausschließlich in aggregierter Form.

Sobald sich ein Teilnehmer mit seinem Code in die Datenbank eingeloggt hatte, wurde dieser Vorgang abgespeichert – der spezifische Code war hiermit für diese aktuelle Befragung vergeben. Einem Missbrauch bzw. einer Verzerrung der Daten durch mehrfache Bearbeitung wurde dadurch entgegengewirkt. Im Falle einer willentlichen oder technisch verur-

sachten, unbeabsichtigten Unterbrechung, konnte sich der Teilnehmer anhand einer PIN⁷, die ihm auf Nachfrage zugeschickt wurde, erneut in die Befragung einloggen und die noch unbearbeiteten Teile bearbeiten. Dies war möglich, da die Befragungen in mehrere Teile aufgegliedert waren und bei erfolgter Bearbeitung nacheinander in der Datenbank abgespeichert wurden.

4.2 Untersuchungsdesign

Die Datenerhebung fand im Rahmen einer Felduntersuchung statt. Dabei nahm die Abteilung des „Global Services“ des Unternehmens an der Vertriebs-Qualifizierungsmaßnahme teil. Eine Zuordnung der Seminarteilnehmer zu den einzelnen Seminarterminen erfolgte nur in dem Sinne „zufällig“, als dass die Trainees von der Personalabteilung je nach individueller Terminlage auf die neun Seminartermine verteilt wurden.

Der Vorteil einer solchen Vorgehensweise liegt in der externen Validität der Ergebnisse. Da diese Art von Daten im Vergleich zu Laborstudien in kaum oder nicht veränderten Umgebungen erhoben werden, zeigen sie nach Ansicht von Bortz und Döring (1995) ein Stück unverfälschter Realität. Zu bedenken ist dabei jedoch, dass bereits die Durchführung einer Untersuchung in einem „natürlichen“ Umfeld eine Veränderung und somit eine Verfälschung darstellt. Außerdem konnten in der vorliegenden Untersuchung keinerlei Störvariablen (z.B. Unterschiede zwischen den Teilnehmern der einzelnen Seminartermine) kontrolliert werden, was zu Lasten der internen Validität geht.

Ausgehend von den in Kapitel 3 formulierten Fragestellungen sind die von Kirkpatrick (1959, 1996) vorgeschlagenen Evaluationskriterien auf der Reaktions-, Lern- und Verhaltensebene zu operationalisieren.

Grundlegend orientiert sich die vorliegende Arbeit am Modell Kirkpatrick's (1959, 1996), es gibt jedoch einige Abweichungen.

So erfolgte beispielsweise eine dreifache Messung der Reaktionsebene anstelle der üblichen einmaligen Messung (siehe Tabelle 4-1). Auch die Lernebene wurde unterteilt in Wissen und Einstellungen (Kraiger et al., 1993).

⁷ PIN = Persönliche Identifikationsnummer.

Tabelle 4-1. *Untersuchungsplan*

Ebene nach Kirkpatrick	t0 Basisbefragung (2 Wochen vor Seminar)	t1 Seminar	t2 1. Nachbefragung (nach 2-3 Wochen)	t3 2. Nachbefragung (nach 3 Monaten)
Reaktionen		X ₁	X ₂	X ₃
Lernen:			X ₁	
- Wissen				
- Einstellung zum Vertriebsverhalten	X ₁			X ₂
Subjektive Verhaltenseinschätzung				X ₁

Anmerkungen. Die „X“ kennzeichnen die Durchführung einer Messung der jeweiligen Ebene zu dem entsprechenden Zeitpunkt, der Index entspricht der Anzahl an Messungen.

Der Wissenstest wurde dabei nicht, wie sonst üblich, parallel zum unmittelbaren Feedback erhoben, sondern zusammen mit der zweiten Reaktionsmessung nach zwei Wochen. Um Einstellungsänderungen zu erhalten, wurden die verhaltensbezogenen Einstellungen vor dem Seminar und zusammen mit der dritten Reaktionsmessung nach drei Monaten erfasst. Eine Messung der Ergebnisse (Ebene 4) konnte entgegen der ursprünglichen Planung nicht realisiert werden. Schließlich wurden noch weitere, im Trainingseffektivitätsmodell nach Tannenbaum (Cannon-Bowers et al., 1995) bedeutsame Kriterien erhoben.

Tabelle 4-2 gibt zum besseren Verständnis einen Überblick über den zeitlichen Verlauf der Messungen sowie über die verwendeten Instrumente.

Tabelle 4-2. *Übersicht Messzeitpunkte und eingesetzte Instrumente bei Seminarteilnehmern (TN), Kontrollgruppe (KG) sowie Führungskräften (FK)*

	t0 Basisbefragung (2 Wochen vor Seminar)	t1 Seminar	t2 1. Nachbefragung (nach 2-3 Wochen)	t3 2. Nachbefragung (nach 3 Monaten)
TN (N=47)	<ul style="list-style-type: none"> ▪ Biographische Daten ▪ Evaluationsbogen ▪ Offene Fragen 	<ul style="list-style-type: none"> ▪ <i>seminarCheck</i> 	<ul style="list-style-type: none"> ▪ <i>seminarUpdate</i> ▪ Wissenstest 	<ul style="list-style-type: none"> ▪ <i>transferCheck</i> ▪ Evaluationsbogen ▪ Offene Fragen
KG (N=22)	<ul style="list-style-type: none"> ▪ Biographische Daten ▪ Evaluationsbogen 		<ul style="list-style-type: none"> ▪ Wissenstest 	<ul style="list-style-type: none"> ▪ Evaluationsbogen
FK (N=17)	<ul style="list-style-type: none"> ▪ Führungskraft-Evaluationsbogen ▪ Offene Fragen 			<ul style="list-style-type: none"> ▪ Führungskraft-Evaluationsbogen ▪ Offene Fragen

Anmerkungen. In fett-kursiver Schrift sind die drei Messungen der Teilnehmerreaktionen hervorgehoben. Die graue Schrift bei der Erhebung der Führungskräfte-Daten verdeutlicht, dass eine solche Messung zwar geplant und durchgeführt wurde, aufgrund der niedrigen Rücklaufquote im Laufe der Untersuchung aber nicht weiter verfolgt wurde.

Außer dem unmittelbaren Feedbackbogen für die Teilnehmer am Ende des Seminars, der als Papierfragebogen durchgeführt wurde, erfolgten alle Befragungen online. Die Instrumente, die sowohl bei den Teilnehmern als auch bei der Kontrollgruppe eingesetzt wurden, waren inhaltlich identisch. Für die Führungskräfte wurden jeweils dieselben Items verwendet und lediglich zur Fremdeinschätzung umformuliert.

Eine Beschreibung der in Tabelle 4-2 dargestellten einzelnen Instrumente erfolgt nach der Darstellung des Untersuchungsablaufs in Kapitel 4.4.

4.3 Untersuchungsablauf

Etwa 10-14 Tage vor Beginn ihres jeweiligen Seminars wurden die Seminarteilnehmer über E-Mail erstmalig von der Diplomandin kontaktiert und über das Evaluationsprojekt sowie über die Datenerhebung im Rahmen der Diplomarbeit informiert. Die E-Mail diente gleichzeitig als Einladung zur Bearbeitung der Basisbefragung (t₀). Die Online-Basisbefragung enthielt die biographischen Daten und den sogenannten Evaluationsbogen, d.h. die Einschätzungen zu Selbstwirksamkeit, Motivation sowie die verhaltensbezogenen Einstellungen. Dieser Teil war für Kontrollgruppe und Teilnehmer identisch. Den Abschluss der Basisbefragung bildeten für die Teilnehmer die folgenden beiden offenen Fragen „*Was erwarten Sie von der anstehenden Vertriebs-Qualifizierungsmaßnahme (PSS)?*“ und „*Wie schätzen Sie Ihr Vertriebspotenzial ein?*“. Zwei bis drei Tage vor Beginn des Seminars erfolgte im Falle einer Nicht-Bearbeitung noch einmal eine Erinnerung. Die Basisbefragung wurde am Morgen vor Beginn des Seminars geschlossen.

Exkurs: Seminarablauf

Das Seminar „Professional Selling Skills“ wurde an neun verschiedenen Seminarterminen durchgeführt. Im Falle der vorliegenden Untersuchung wurde es jeweils von ein- und demselben Trainer abgehalten. Der Zeitraum, in dem die Seminare stattfanden, umfasste April bis Dezember 2005. Die ersten beiden Termine wurden „in-house“ im Unternehmen durchgeführt, alle weiteren Seminare fanden auswärts an insgesamt drei verschiedenen Tagungs-orten statt (sechs in Deutschland, eines in der Schweiz).

Die Dauer des Seminars betrug jeweils drei Tage, in deren Verlauf die Themen *Verkauf durch Bedürfnisbefriedigung*, *Gesprächstechniken*, *Gleichgültigkeit überwinden* und *Um-*

gang mit Einwänden behandelt wurden. An den ersten 1½ Tagen beschäftigten sich die Teilnehmer vorrangig mit dem *Verkauf durch Bedürfnisbefriedigung* und übten die *Gesprächstechniken*. In den letzten 1½ Tagen ging es unter Einbezug des bisher Gelernten darum, die Themen *Gleichgültigkeit überwinden* und *Umgang mit Einwänden* seitens des Kunden einzuüben. Neben dem selbstständigen Erarbeiten von Texten waren die vorherrschenden Vermittlungstechniken Videobeispiele sowie abwechselnde Rollenspiele in Kleingruppen.

Kurz vor Abschluss des dreitägigen Seminars (t1) verteilte der Trainer den Seminar-Feedbackbogen (seminarCheck[®]), durch den die erste Messung der Teilnehmerreaktionen erfolgte. Für den Fall, dass die Teilnehmer ihren persönlichen Code nicht verfügbar hatten, war der Trainer angehalten, die Teilnehmer beim Austeilen der Bögen darum zu bitten, in diesem Fall das Geburtsdatum der Mutter als „Hilfs-Code“ zu verwenden. Dies sollte Phantasie-Codes vermeiden und sicherstellen, dass der Hilfs-Code später auf Nachfrage der Diplomandin leicht erinnert werden konnte. Dieses Vorgehen ermöglichte eine nachträgliche Zuordnung zum eigentlichen individualisierten Code in der Datenbank. Die Feedbackbögen wurden nach dem Ausfüllen vom Trainer eingesammelt und zur Auswertung an das Institut für Begleitforschung geschickt.

Etwa 14 Tage nach dem Seminar (t2) erhielten Teilnehmer und Kontrollgruppe die Einladungs-E-Mail zur 1. Nachbefragung mit allen notwendigen Informationen zur Befragung, den Zugangsdaten sowie den persönlichen Code. Im Rahmen dieser ersten Online-Nachbefragung wurde für beide Gruppen der Wissenstest eingesetzt. Gleichzeitig erfolgte für die Teilnehmer die zweite Erhebung der Teilnehmerreaktionen (seminarUpdate[®]). Der Wissenstest in der 1. Nachbefragung war für Teilnehmer und Kontrollgruppe identisch, jedoch erhielten nur die Teilnehmer eine Rückmeldung über ihr Test-Ergebnis als individuelles Online-Feedback. Dies geschah auf Wunsch des Seminaranbieters, um bei der Kontrollgruppe unnötige Frustration zu vermeiden und dadurch die Bereitschaft zur Bearbeitung der letzten Befragung aufrechtzuerhalten. Bei Nicht-Bearbeitung wurde per E-Mail an die ausstehende Bearbeitung erinnert.

Die Einladung zur Bearbeitung der 2. Nachbefragung wurde ca. drei Monate nach dem Seminar (t3) an die Teilnehmer geschickt. Diese zweite Online-Nachbefragung enthielt

wieder den Evaluationsbogen, der, wie bei der Basisbefragung, für Kontrollgruppe und Teilnehmer identisch war. Darüber hinaus wurden zum dritten Mal die Teilnehmerreaktionen (transferCheck[®]) erhoben sowie als Pendant zur Basisbefragung folgende offene Fragen gestellt: „*Wie sehen Sie die Entwicklung Ihrer Vertriebsfertigkeiten in den letzten Wochen?*“ und „*Wie beurteilen Sie rückblickend die Vertriebs-Qualifizierungsmaßnahme?*“. Auch hier erfolgten Erinnerungen bei Nicht-Bearbeitung.

Einige Ergebnisse der drei Online-Befragungen (Basis-, 1. und 2. Nachbefragung) wurden am Ende der jeweiligen Befragung individuell online zurückgemeldet, mit der Möglichkeit, diese individuellen Ergebnisse auszudrucken. Ausgewählte Beispiele für solche Ergebnisrückmeldungen sind als ‚Screenshots‘ in Anhang A, S. 49ff (für die 1. Nachbefragung) und S. 61ff (für die 2. Nachbefragung) zu finden.

4.4 Untersuchungsvariablen

Im Folgenden sollen die einzelnen Variablen beschrieben werden, die in der vorliegenden Untersuchung erhoben wurden. Es wird darauf eingegangen, welche Instrumente Verwendung fanden und ggf. ihre Entwicklung beschrieben. Als Reliabilitätsangabe wird der Schätzer der internen Konsistenz, Cronbachs Alpha (α), angegeben. Nach Lance, Butts und Michels (2006) ist ab $\alpha = .80$ von einer ausreichenden Reliabilität auszugehen, grundsätzlich sollte dieser Wert jedoch über $\alpha = .90$ liegen. Reliabilitäten von $\alpha = .70$ erscheinen hingegen lediglich bei Messungen im Sinne von Voruntersuchungen akzeptabel, bei denen Zeit und Ressourcen gespart werden sollen. Anhang A, S. 93, bietet eine Übersicht aller berechneten Kennwerte.

4.4.1 Kriterien nach Kirkpatrick

4.4.1.1 Reaktionsebene

Die Teilnehmerreaktionen werden nicht nur einmalig über einen unmittelbaren Seminar-Feedbackbogen am Ende des Seminars gemessen, sondern zusätzlich an zwei nachfolgenden Zeitpunkten. Die Messung erfolgt durch ein von der Firma HR CheckSystems GmbH entwickeltes und standardmäßig eingesetztes Instrument (seminarPlus[®]). Die drei Module dieses Instruments (seminarCheck[®], seminarUpdate[®] und transferCheck[®]) messen zu verschiedenen Zeitpunkten unterschiedliche Aspekte der Zufriedenheit. So konzentriert sich

die erste Messung auf unmittelbare Zufriedenheits-Facetten bezüglich des Seminars, in den anderen beiden Messungen liegt der Schwerpunkt dagegen auf der Umsetzung und der Beschäftigung mit den Seminarinhalten. Zum besseren Verständnis sollen die Inhalte der einzelnen Module nun vorgestellt werden.

Erste Messung der Teilnehmerreaktionen (Seminar-Feedbackbogen = t1)

Anhand dieses standardisierten Papierfragebogens (seminarCheck[©]) wurde mit insgesamt 20 Items die Zufriedenheit der Teilnehmer bezüglich mehrerer Facetten abgefragt. So wurde mit jeweils einem Item eine *Gesamteinschätzung der Veranstaltung* und des *Trainers* erhoben. Mit jeweils zwei Items wurden *Rahmenbedingungen* ($r = .43, p < .01$) sowie die *Vorbereitung der Teilnehmer* ($r = .50, p < .01$) erfasst. Darüber hinaus enthält der unmittelbare Feedbackbogen die folgenden Subtests: *Durchführung des Seminars* mit fünf Items ($\alpha = .83$, z.B. „Die Übungen/Aufgaben waren für die Vermittlung der Inhalte gut geeignet“), *Anwendbarkeit der Inhalte* mit vier Items ($\alpha = .83$, z.B. „Ich kann die Inhalte an meinem Arbeitsplatz nutzen“) und *Differenzierte Bewertung des Trainers* mit sieben Items ($\alpha = .90$, z.B. „Der Trainer hat Ziele und Erwartungen erfragt“).

Bei der in der Praxis häufig eingesetzten Schulnotenskala von 1 ‚sehr gut‘ bis 6 ‚ungenügend‘ zeigt sich oftmals eine Kumulierung der Urteile in den beiden Kategorien 1 ‚sehr gut‘ und 2 ‚gut‘, die übrigen Noten werden eher selten vergeben. Um jedoch gerade im oberen Skalenbereich eine genauere Differenzierung zu erreichen, wird in Anlehnung an die Schulnotenskala folgende differenziertere Skala verwendet:

Zustimmung			...			Ablehnung			
1+	1	1-	2+	2	2-	3+	3	3-	>3-
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Zur Auswertung der Items wird diese Skala in eine 10-stufige Skala mit den Polen ‚1+‘ = 9 bis ‚>3-‘ = 0 umkodiert.

Eine weitere Besonderheit des hier eingesetzten Feedbackbogens ist die differenzierte Einschätzung jedes einzelnen Seminarthemas (Verkauf durch Bedürfnisbefriedigung, Gesprächstechniken, Kundeneinwände, Gleichgültigkeit überwinden) mittels zweier Items

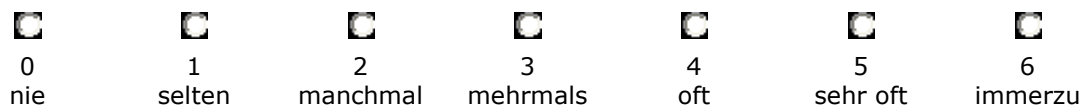
(Thema X „wurde interessant erarbeitet“; Thema X „kann ich in der Praxis umsetzen“). Das verwendete Instrument ist als Ganzes in Anhang A, S. 28 wiedergegeben.

Zweite Messung der Teilnehmerreaktionen (im Rahmen der 1. Nachbefragung = t2)

Diese Online-Befragung (seminarUpdate[®]) erfasst mit insgesamt 19 Items folgende Zufriedenheitsaspekte: *Anwendung der Inhalte* mit fünf Items ($\alpha = .64$, z.B. „Ich habe aufgrund des Seminars etwas Neues ausprobiert“), *Nutzen des Seminars* mit vier Items ($\alpha = .84$, z.B. „Meine Arbeit hat sich durch das Seminar positiv verändert“), *Erfahrungsaustausch im Seminar* mit vier Items ($\alpha = .80$, z.B. „Der Erfahrungsaustausch im Seminar war hilfreich für mich“) und *Seminarbewertung im Rückblick* mit 6 Items ($\alpha = .87$, z.B. „Das Seminar hat mir etwas gebracht“). Die Einschätzung erfolgte auf einer 7-stufigen Zustimmungsskala (0 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘).



Wie schon der unmittelbare Feedbackbogen beinhaltet auch diese Befragung neben der Einschätzung der Zufriedenheits-Facetten eine *Einschätzung der einzelnen Seminarthemen*. Diesmal wurden jedoch zwei andere Items pro Thema als im Feedbackbogen verwendet: [In den letzten sieben Tagen] (1) „...habe ich mich mit diesem Thema beschäftigt“, (2) „...habe ich gute Erfahrungen mit diesem Thema gemacht“. Zur Einschätzung der im Seminar behandelten Themen wurde eine 7-stufige Häufigkeitsskala verwendet, die sich auf den Zeitrahmen der letzten sieben Tage bezieht (0 = ‚nie‘, 6 = ‚immerzu‘).



Abschließend wurde den Teilnehmern die Möglichkeit gegeben, ihre persönlichen Ziele zu notieren, die sie sich in der Veranstaltung oder im Anschluss an das Seminar gesetzt hatten. In Anhang A, S. 42ff, sind alle Items dieses Fragebogens dargestellt.

Dritte Messung der Teilnehmerreaktionen (im Rahmen der 2. Nachbefragung = t3)

In dieser Online-Befragung (transferCheck[®]) wurden mit 14 Items folgende Zufriedenheitsaspekte eingeschätzt: *Anwendung der Inhalte* mit fünf Items ($\alpha = .80$, z.B. „Ich konnte Aspekte des Seminars umsetzen“), *Nutzen des Seminars* mit vier Items ($\alpha = .90$, z.B. „Durch das Seminar bin ich besser in der Lage, meine Arbeit zu erfüllen“) und *Umsetzung der persönlichen Ziele* mit fünf Items ($\alpha = .83$, z.B. „Meine Ziele haben sich als nützlich erwiesen“). Es wurde wie bei der ersten Nachbefragung eine 7-stufige Zustimmungsskala eingesetzt (0 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘). Anhang A, S. 52ff, bietet eine detaillierte Übersicht aller Items.

Des Weiteren wurde anhand von acht Items nach Einflussfaktoren auf den Transfer gefragt, hier in Bezug auf die Unterstützung seitens der Vorgesetzten sowie durch Kollegen und situative Faktoren am Arbeitsplatz. Für diese Einschätzungen wurde wieder eine 7-stufige Häufigkeitsskala verwendet, mit dem Referenzzeitraum der letzten vier Wochen (0 = ‚nie‘, 6 = ‚immerzu‘). So wurde die *Unterstützung durch die Führungskraft* (vier Items, $\alpha = .80$) bspw. durch das Item „[In den letzten vier Wochen] ...war meine Führungskraft daran interessiert, dass die Seminarinhalte tatsächlich umgesetzt werden“ erfragt, die *Unterstützung durch das Arbeitsumfeld* (vier Items, $\alpha = .80$) z.B. durch das Item „[In den letzten vier Wochen] ... hatte ich in meinem Berufsalltag die Möglichkeit, das Gelernte anzuwenden“.

Die Einschätzung der im Seminar behandelten Themen erfolgte pro Thema mittels drei Items: [In den letzten vier Wochen] ... (1) „...habe ich mich mit dem Thema beschäftigt“; (2) „...habe ich das Thema angewendet“; (3) „... hat mir dieses Thema etwas gebracht“.

Abschließend wurde den Teilnehmern durch offene Fragen die Möglichkeit gegeben, Anmerkungen zur Umsetzung der Seminarinhalte zu notieren: „Was haben Sie als positiv (fördernd/unterstützend) erlebt, bei der Umsetzung der Seminarinhalte?“ und „Was haben Sie als negativ (hinderlich/hemmend) erlebt, bei der Umsetzung der Seminarinhalte?“. Darüber hinaus wurde noch spezifisch nach der Meinung zum Weiterbildungsangebot und zum Weiterbildungsbedarf gefragt: „Wie ist Ihre Meinung zu den Qualifizierungsangeboten Ihres Arbeitgebers?“ und „Welche Maßnahmen (Seminare, Qualifizierungen etc.) wünschen Sie sich als nächstes für Ihre berufliche Weiterentwicklung?“.

4.4.1.2 Lernebene - Wissen

Der Wissensstand der Teilnehmer wurde einmalig im Rahmen der 1. Nachbefragung durch den PSS-Lerntest[®] erfasst, einem vom Seminaranbieter AchieveGlobal entwickelten und standardmäßig eingesetzten Wissenstest. Die 26 Items mit insgesamt 105 richtig/ falsch-Antwortalternativen decken folgende Themenbereiche ab:

- *Verkauf durch Bedürfnisbefriedigung* (zwei Items, $r = .11$)
- *Gesprächseröffnung* (drei Items, $\alpha = .53$)
- *Fragetechnik* (sechs Items, $\alpha = .65$)
- *Unterstützungstechnik* (vier Items, $\alpha = .18$)
- *Abschlussstechnik* (drei Items, $\alpha = -.48$)
- *Gleichgültigkeit überwinden* (zwei Items, $r = -.10$)
- *Umgang mit Einwänden* (sechs Items, $\alpha = .54$)

Die gefundenen niedrigen Werte der inneren Konsistenz deuten auf eine Messungenauigkeit des Wissenstests hin. Es war nicht jedoch möglich, in einem Vorversuch den Wissenstest auf seine psychometrischen Kennwerte hin zu überprüfen. Außerdem sollte durch den Wissenstest ein Erfolgsindikator für Wissen ermittelt werden. Dieses Maß wird durch die Anzahl an richtigen Antworten gewonnen, wodurch der Reliabilitätsaspekt der einzelnen Subtests auch nicht im Vordergrund steht.

Pro Frage sind eine oder mehrere Antwortalternativen richtig, ebenso können alle bzw. keine der Antworten richtig sein. Eine Rate- oder Zufallskorrektur erschien bei der Rohwertermittlung im vorliegenden Fall nicht notwendig, da es aufgrund der großen Anzahl an Aufgaben nach Lienert und Raatz (1998) kaum einen Unterschied macht, ob man mit oder ohne Zufallskorrektur arbeitet.

Abbildung 4-1 zeigt beispielhaft ein Item zum Thema *Gesprächseröffnung*, der vollständige Wissenstest ist in Anhang A, S. 31ff, zu finden.

Sie eröffnen ein Verkaufsgespräch und möchten sich mit dem Kunden darüber einigen, was während des Gesprächs erreicht werden soll. Würden Sie Folgendes sagen:

a) „Ich würde heute gerne über Ihre Systemanforderungen sprechen, so dass ich ein Angebot erstellen kann, das Ihren Bedürfnissen voll entspricht.“

ja nein

b) „Ich würde heute gerne mehr darüber erfahren, welchen Service Sie zur Zeit in Anspruch nehmen und in welchen Bereichen Sie etwas verbessern möchten. Auf diese Weise kann ich Ihnen einen Service empfehlen, der Ihre Bedürfnisse am besten befriedigen kann. Was halten Sie davon?“

ja nein

c) „Wie ich bereits am Telefon erwähnt habe, möchte ich mich heute mit Ihnen darüber unterhalten, wie Sie zur Zeit Ihre Rechnungen bearbeiten. Wäre das sinnvoll für Sie?“

ja nein

d) „Ich würde heute gerne mehr darüber erfahren, wie Sie Ihre Computer einsetzen, damit ich ein Angebot erstellen kann, das speziell auf Ihre Anforderungen zugeschnitten ist. Wäre das ein guter Ansatz für unser heutiges Gespräch?“

ja nein

Abbildung 4-1. Beispiel-Item des Wissenstests zum Thema Gesprächseröffnung. Richtige Antwortmöglichkeiten wären in vorliegendem Fall b) und d); Antworten a) und c) wären dahingegen falsch.

4.4.1.3 Lernebene - Einstellungen

Der Seminaranbieter entwickelte im Vorfeld des Evaluationsprojektes 28 Items, die sich auf Einstellungen zum Verhalten beim Kundenkontakt beziehen. Vor Aufnahme in die Datenbank wurden diese Items durch die Diplomandin nach Absprache mit dem Anbieter leicht modifiziert und in einem mehrstufigen Prozess inhaltlich in fünf Themengruppen unterteilt, die im Vertrieb bzw. bei einem Verkaufsprozess von Bedeutung sind⁸.

Mit den erhobenen Daten der Basisbefragung wurde pro Themengruppe (Subtest) eine Reliabilitätsanalyse durchgeführt. Einige Items wurden aufgrund ihrer geringen Trennschärfe herausgenommen. Die endgültige Fassung der fünf Subtests beinhaltet drei Items für den Subtest *Bedürfnisbefriedigung* ($\alpha = .69$), sechs Items für den Subtest *Vertriebsorientierung* ($\alpha = .85$), sieben Items für den Subtest *Umgang mit Einwänden* ($\alpha = .90$)

⁸ Die hier ausgewählten Themen decken sich mit den Inhalten, die allgemein in Vertriebstrainings eine Rolle spielen. Je nach Seminaranbieter werden diese Themen dabei lediglich unterschiedlich benannt.

sowie je vier Items für die Subtests *Kundenorientierung* ($\alpha = .74$) und *vertriebsbezogenes Engagement* ($\alpha = .72$). Ein Beispiel-Item des Subtests *Vertriebsorientierung* lautet: „Es gelingt mir, dem Kunden den Nutzen eines Zusatzangebots deutlich zu machen“. In Anhang A, S. 63ff, sind alle Items der verhaltensbezogenen Einstellungen dargestellt. Die Einstellungs-Items wurden auf einer 7-stufigen Zustimmungsskala (0 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘) eingeschätzt.

4.4.1.4 Verhaltensebene

Die Wahl eines geeigneten Instruments erwies sich als schwierig, da eine Verhaltensbeobachtung am Arbeitsplatz aus Zeitgründen und mangels ausreichender personeller Ressourcen als Evaluationsmöglichkeit ausgeschlossen war. Eine Lösung bot sich durch ein Item aus der 2. Nachbefragung: Die Teilnehmer mussten hier zu jedem der vier Seminarthemen anhand einer 7-stufigen Häufigkeitsskala (0 = ‚nie‘, 6 = ‚immerzu‘) angeben, wie häufig sie dieses Thema in den letzten vier Wochen angewendet hatten. Als subjektives Maß der Verhaltenseinschätzung, d.h. der Anwendungshäufigkeit der Themen, wurde der Mittelwert dieses Items über alle vier Themen gebildet.

4.4.1.5 Auswahl der Zielkriterien

Aufgrund der umfangreichen Datenerhebung der vorliegenden Untersuchung war es erforderlich, den Fokus auf diejenigen Subtests bzw. spezifischen Variablen zu legen, die für die Überprüfung der Hypothesen relevant sind. Da v.a. die Teilnehmerreaktionen durch die verwendeten Instrumente in sehr viele Subtests ausdifferenziert wurden, erscheint gerade hier eine Auswahl geeigneter Variablen zur Reduktion der Datenmenge notwendig und wird auf der Basis der unter Kapitel 2 beschriebenen theoretischen Befunde getroffen.

Mehrheitlich wird die Ausdifferenzierung der Reaktionsebene in *affective* und *utility reactions* nahegelegt (z.B. Alliger et al., 1997; Warr & Bunce, 1995). Auch in der vorliegenden Untersuchung sollten beide Zufriedenheitsaspekte berücksichtigt werden. Das zur Messung der Reaktionsebene verwendete Instrument folgte dem Konzept, dass zu verschiedenen Zeitpunkten unterschiedliche Zufriedenheitsaspekte wichtig sind. Affektive Aspekte (Zufriedenheit mit Seminar, Trainer etc.) wurden vorrangig mit dem unmittelbaren Seminar-Feedbackbogen erfasst, während bei den zwei Online-Nachbefragungen vor allem die Anwendung der Seminarinhalte in der Arbeitssituation sowie der Nutzen, den die Teilnehmer vom Seminar sehen, vordergründig waren.

Auf der Reaktionsebene wurden daher für den **unmittelbaren Seminar-Feedbackbogen** folgende Zielkriterien ausgewählt: Zur Erfassung der *affective reactions* der Subtest *Durchführung* (fünf Items, $\alpha = .83$) und zur Erfassung der *utility reactions* der Subtest *Anwendbarkeit* (vier Items, $\alpha = .83$). Als Zielkriterien der **1. Nachbefragung** wurden zur Erfassung der *utility reactions* die Subtests *Anwendung der Inhalte t2* (fünf Items, $\alpha = .64$) und *Nutzen des Seminars t2* (vier Items, $\alpha = .84$) herangezogen. Für die **2. Nachbefragung** wurden als *utility reactions* die Subtests *Anwendung der Inhalte t3* (fünf Items, $\alpha = .80$), *Nutzen des Seminars t3* (vier Items, $\alpha = .90$) sowie die *Umsetzung persönlicher Ziele t3* (fünf Items, $\alpha = .83$) ausgewählt. Der Subtest *Anwendung* beinhaltet dabei Items zur Bewertung der Umsetzung bzw. Beschäftigung mit den Inhalten, während der Subtest *Nutzen* eher die empfundenen Konsequenzen der Maßnahme darstellt (z.B. die Auswirkungen des Seminars auf die Arbeits- sowie die persönliche Situation).

Auf der Lernebene wurde als Zielkriterium zum einen das Ergebnis im **Wissenstest** herangezogen (*Anzahl richtiger Antworten in Prozent*). Um zum anderen das Zielkriterium für die **verhaltensbezogenen Einstellungen** zu erhalten, wurde zunächst sowohl bei der Basisbefragung (t0) als auch bei der 2. Nachbefragung (t3) der Mittelwert über die fünf Einstellungsbereiche (Bedürfnisbefriedigung, Vertriebsorientierung, Umgang mit Kunden einwänden, Kundenorientierung und vertriebsbezogenes Engagement) gebildet. In einem zweiten Schritt ergab dann die *Differenz t3 - t0* das Zielkriterium der verhaltensbezogenen Einstellungen.

Auf der Verhaltensebene wurde das Zielkriterium ebenfalls durch Mittelwertsberechnungen gewonnen. Dabei ergaben die gemittelten Einschätzungen der zweiten Nachbefragung, wie oft die Teilnehmer jedes der vier Seminarthemen im Bezugszeitraum der letzten vier Wochen tatsächlich angewendet hatten, das **subjektiv eingeschätzte Verhalten** (subjektive Verhaltenseinschätzung).

4.4.2 Kriterien nach Tannenbaum

Neben den Evaluationskriterien von Kirkpatrick wurden zusätzlich im Rahmen der hier durchgeführten Befragungen einige Faktoren erhoben, die bei der Messung der Trainingseffektivität gemäß dem Modell von Tannenbaum (Cannon-Bowers et al., 1995) eine Rolle spielen. Diese werden nachfolgend näher beschrieben.

4.4.2.1 Motivation

Leistungsmotivation wird von Schuler und Prochaska (2000) neben der allgemeinen kognitiven Fähigkeit als relevantes Merkmal im Hinblick auf beruflichen Erfolg gesehen. Die Frage, wie das Konstrukt der Leistungsmotivation genau aussieht, ist nicht leicht zu beantworten: Bei Durchsicht der theoretischen Ansätze und veröffentlichten Testverfahren ergeben sich ca. 100 Teilfacetten, von denen sich allerdings einige überschneiden bzw. synonym verwendet werden (Schuler & Prochaska, 2000). Unter den am häufigsten genannten Dimensionen finden sich *Zielsetzung*, *Antriebsstärke*, *Beharrlichkeit*, *Erfolgshoffnung* und *Misserfolgsbefürchtung*. Schuler und Prochaska (2000) entwickelten ein 170 Fragen umfassendes Leistungsmotivationsinventar (LMI), von dem eine 30 Items umfassende Kurzform ebenfalls zur Verfügung steht.

Um den zeitlichen Aufwand jedoch geringer zu halten, wurden in der vorliegenden Diplomarbeit zur Messung der Leistungsmotivation in Anlehnung an die o.g. Dimensionen acht eigene Items entwickelt (z.B. „*Eine angefangene Arbeit möchte ich auch zu Ende führen*“), die auf einer 7-stufigen Zustimmungsskala (0 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘) eingeschätzt wurden. Aufgrund geringer Trennschärfen wurden drei Items herausgenommen, so dass zur Erfassung der Leistungsmotivation (im weiteren Verlauf der Arbeit nur noch Motivation genannt) fünf Items blieben. Die Reliabilitätsanalyse ergab ein $\alpha = .73$. Eine vollständige Übersicht dieser Items ist im Anhang A, S. 63ff, enthalten.

Im Hinblick auf die unter Kapitel 3.3 aufgestellten Hypothesen geht es nicht nur darum, Zusammenhänge zwischen den Einflussvariablen festzustellen, sondern vielmehr Unterschiede in der Ausprägung der Variablen herauszufinden. Daher gilt es, den Einfluss hoher im Gegensatz zu niedriger Motivation zu prüfen, weshalb die erhobenen Motivationswerte mit Hilfe eines Mediansplits dichotomisiert werden. Entsprechend wird auch für alle weiteren erhobenen Einflussgrößen die Dichotomisierung am Median durchgeführt.

In der „Median-Kategorie“ sollten die Werte idealerweise zu 50% darüber und 50% unter dem Median liegen. Hier liegt jedoch eine etwas breitere Median-Kategorie vor, die 58% der Werte darunter und entsprechend 42% darüber umfasst. Dies ergibt folgende Aufteilung: 0 = ‚*geringe bis mittlere Motivation*‘ ($n = 18$) und 1 = ‚*hohe Motivation*‘ ($n = 13$).

4.4.2.2 Selbstwirksamkeit

Zur Messung der beruflichen Selbstwirksamkeit wurde die Kurzversion der ‚Skala zur Erfassung der beruflichen Selbstwirksamkeit‘ von Schyns und v. Collani (2002) eingesetzt (z.B. *„Beruflichen Schwierigkeiten sehe ich gelassen entgegen, weil ich mich immer auf meine Fähigkeiten verlassen kann“*). Für die Parallel-Befragung der Führungskräfte wurden dieselben Items verwendet und entsprechend umformuliert. Zwei Items mussten allerdings ausgeschlossen werden, da sie kaum umzuformulieren waren und das Ziel war, die Parallelität der beiden Einschätzungen zu wahren (s. Anhang A, S. 65, für den Originalfragebogen von Schyns und v. Collani (2002) sowie Anhang A, S. 63, für die modifizierte Version).

Weitere zwei Items wurden aufgrund geringer Trennschärfe entfernt, so dass der hier verwendete Selbstwirksamkeits-Fragebogen insgesamt sechs Items beinhaltet. Diese wurden nicht wie im Original auf einer 6-stufigen Skala gemessen (1 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘), sondern in Anpassung an das Antwortformat der vorliegenden Arbeit durch die 7-stufige Zustimmungsskala (0 = ‚stimmt überhaupt nicht‘, 6 = ‚stimmt völlig‘). Die Reliabilitätsanalyse ergab ein α von .79⁹.

Nach der Dichotomisierung am Median ergaben sich folgende zwei Gruppen: 0 = ‚geringe bis mittlere Selbstwirksamkeit‘ ($n = 15$) und 1 = ‚hohe Selbstwirksamkeit‘ ($n = 16$).

4.4.2.3 Subjektiver Bedarf

Da im Vorfeld keine umfassende Bedarfsanalyse durchgeführt werden konnte, wurde der Bedarf mittels eines Items aus dem Subtest *Anwendbarkeit* des Seminar-Feedbackbogens erfasst (*„Das Seminar hat meinem Bedarf entsprochen“*, eingestuft auf der 10-stufigen Skala des Seminar-Feedbackbogens, vgl. Abschnitt 4.4.1.1). Mit Hilfe eines Mediansplits wurde dieses Item dichotomisiert und die Variable *Bedarf* mit folgenden Ausprägungen gebildet: $n = 15$ Teilnehmer mit Einschätzungen von 3⁺, 2⁻, 2 und 2⁺ wurden zur Gruppe 0 = ‚geringer bis mittlerer Bedarf‘ zusammengefasst, wohingegen $n = 15$ Teilnehmer mit Einschätzungen von 1⁻, 1 und 1⁺ die Gruppe 1 = ‚hoher Bedarf‘ bildeten.

⁹ Zum Vergleich: Die 10-Items umfassende Kurzversion der *Skala zur beruflichen Selbstwirksamkeitserwartung* weist eine Reliabilität von $\alpha = .88$ auf (Schyns, 1999).

4.4.2.4 Anwendungsmöglichkeit aus Sicht des Unternehmens

Die Häufigkeit des Kundenkontakts aus organisationaler Sicht gibt Aufschluss darüber, ob die Teilnehmer tatsächlich die Möglichkeit haben, die vertriebsbezogenen Inhalte anzuwenden. Je mehr Kundenkontakt besteht, umso größer sollte die Gelegenheit zur Anwendung sein. Diese Informationen wurden vom Personalverantwortlichen des Unternehmens eingeholt, der $n = 5$ Teilnehmer mit 0 = ‚wenig Kundenkontakt‘ klassifizierte, $n = 17$ Teilnehmer mit 1 = ‚mittlerer Kundenkontakt‘ sowie $n = 9$ Teilnehmer mit 2 = ‚viel Kundenkontakt‘.

4.4.2.5 Vorerfahrung (Expertise)

Durch die Variable *Expertise* soll erfasst werden, wie hoch die Vertriebserfahrung der Teilnehmer ist. Sie wird im biographischen Teil durch die Angaben der Teilnehmer in Bezug auf ihre Tätigkeit operationalisiert (Antwortoptionen waren *keine Angabe*, *keine Vertriebstätigkeit*, *Tätigkeit mit Vertriebsanteilen* und *reine Vertriebstätigkeit*). Da nur wenige Teilnehmer eine reine Vertriebstätigkeit angaben, erfolgte ein Zusammenschluss mit denjenigen Teilnehmern, die über Vertriebsanteile in ihrer Tätigkeit berichteten. Somit bilden $n = 15$ Teilnehmer die Gruppe 1 = ‚mit Vertriebserfahrung‘, wohingegen die Gruppe 0 = ‚ohne Vertriebserfahrung‘ $n = 16$ Teilnehmer zählt.

4.4.2.6 Transferklima

Zur Erhebung des *Transferklimas* wurden die Daten zweier Subtests zusammengefasst, die im Rahmen der Teilnehmereinschätzungen der 2. Nachbefragung erhoben wurden. Der Subtest *Unterstützung durch die Führungskraft* (z.B. ‚In den letzten 4 Wochen war meine Führungskraft daran interessiert, dass die Seminarinhalte tatsächlich umgesetzt werden‘) erfasst dabei mit vier Items ($\alpha = .80$), inwiefern die Führungskraft Interesse an einer Umsetzung der Inhalte zeigt und diesen Umsetzungsprozess aktiv fördert. Mit der *Unterstützung durch das Arbeitsumfeld* (z.B. ‚In den letzten 4 Wochen hatte ich in meinem Berufsalltag die Möglichkeit, das Gelernte anzuwenden‘) wurde anhand von vier Items ($\alpha = .80$) nicht nur die Unterstützung durch Arbeitskollegen erfragt, sondern auch inwiefern eine Anwendung der Inhalte überhaupt möglich ist und zu Veränderungen beitragen kann. Die acht Items wurden auf einer 7-stufigen Häufigkeitsskala eingeschätzt, die sich auf einen Zeitrahmen von vier Wochen bezog (0 = ‚nie‘, 6 = ‚immerzu‘).

Zur Bildung des Kennwerts *Transferklima* wurde zunächst für beide Subtests eine Dichotomisierung am Median durchgeführt, wobei jeweils für beide galt: 0 = ‚niedrige Unterstützung‘ und 1 = ‚hohe Unterstützung‘. Die dichotomisierten Werte wurden in einem zweiten Schritt folgendermaßen eingeteilt: Ein Wert von ‚0‘ in beiden Subtests wurde unter 0 = ‚*negatives Transferklima*‘ zusammengefasst ($n = 12$), wohingegen ein Wert von ‚1‘ in beiden Subtests unter 1 = ‚*positives Transferklima*‘ zusammengefasst wurde ($n = 11$). Die restlichen Teilnehmer mit den Kombinationen ‚0 und 1‘ bzw. ‚1 und 0‘ wurden der Gruppe 2 = ‚*uneinheitliches Transferklima*‘ zugeteilt ($n = 8$).

4.4.2.7 Seminarbewertung

Allgemeine Seminarbewertung (affective reactions)

Die allgemeine Seminarbewertung ergibt sich aus der Gesamteinschätzung der Veranstaltung, die im unmittelbaren Feedbackbogen durch ein Item gemessen wurde. Um herauszufinden, inwiefern sich eine unmittelbare Gesamtbewertung auf weitere Einschätzungen auswirkt, wurden die Daten durch einen Mediansplit dichotomisiert. Dabei bildeten $n = 16$ Teilnehmer mit Einschätzungen von ‚2⁻, 2, 2⁺ und 1‘ die Gruppe 0 = ‚mittel bis gut‘ und weitere $n = 15$ Teilnehmer mit Einschätzungen von ‚1 und 1⁺‘ die Gruppe 1 = ‚sehr gut‘.

Spezifische Seminarbewertung (utility reactions)

Mehreren Autoren (Alliger et al., 1997; Morgan & Casper, 2000; Warr & Bunce, 1995) empfehlen, auf der Ebene der Teilnehmerreaktionen Einschätzungen zur Anwendbarkeit zu erheben. Zur Operationalisierung dieser spezifischen Bewertung wird in der vorliegenden Untersuchung nicht der eigenständige Subtest Anwendbarkeit des Seminar-Feedbackbogens herangezogen, da dieser bereits für die verschiedenen Hypothesen als Prüfvariable (Zielkriterium) verwendet wird. Um eine Konfundierung zu vermeiden, wird stattdessen die themenbezogene Anwendbarkeit verwendet, die pro Seminarthema mit jeweils einem Item im Rahmen des Feedbackbogens einzuschätzen ist („Ich kann dieses Thema in der Praxis anwenden“, 10-stufige Skala des Seminar-Feedbackbogens). Eine Dichotomisierung am Median ergab als Gruppe 0 = ‚mittlere bis gute Bewertung‘ ($n = 15$) und als Gruppe 1 = ‚sehr gute Bewertung‘ ($n = 16$).

Im Rahmen des Evaluationsprojekts wurden als zusätzliche Instrumente zum einen die überarbeitete Fassung des Job Diagnostic Survey (JDS) von Kallus und Schmut (2004) zur

Einschätzung der allgemeinen Arbeitsbedingungen eingesetzt. Zum anderen wurde der auf dem Erholungs-Belastungs-Fragebogen (EBF) von Kallus (1995) aufbauende arbeits-spezifische EBF-78-Work (Kallus & Jiménez, 2005) eingesetzt. Diese Fragebögen wurden für die vorliegende Untersuchung zwar nicht weiter ausgewertet, sollen aber erwähnt werden, da sich die Bearbeitungszeit für die Teilnehmer v.a. durch den Einsatz des EBF-78-Work um etwa 10 Minuten erhöhte (von ca. 10 auf ca. 20 Minuten bei der Basisbefragung sowie von ca. 23 auf ca. 33 Minuten bei der 2. Nachbefragung).

4.5 Datenaufbereitung

Eines der Anliegen dieser Arbeit bestand in der Analyse der zeitlich nachgelagerten Einschätzungen, weshalb, wie bereits beschrieben, einige Datensätze ausgeschlossen wurden.

4.5.1 Rücklaufquote

Die Gesamt-Rücklaufquote der Seminarteilnehmer lag mit 47 von 72 Trainees bei ca. 65%. Im Zuge der Auswertung werden unterschiedliche Ansätze herangezogen, bei denen auf unterschiedlich große Datensätze zurückgegriffen wird.

Beispielweise können für die Berechnung der Korrelationsanalysen alle 47 Datensätze herangezogen werden, wohingegen für die Varianz- bzw. Regressionsanalysen vollständige Datensätze über alle berücksichtigten Zeitpunkte notwendig sind, so dass hierfür lediglich die Daten von 31 Seminarteilnehmern miteinbezogen werden können. Wie sich die Bearbeitungsquote der verbleibenden Seminarteilnehmer sowie der Kontrollgruppe und der Führungskräfte aufgliedert, verdeutlicht Tabelle 4-3¹⁰.

¹⁰ Eine genaue Aufschlüsselung, wie viele Seminarteilnehmer welche Befragungskombination bearbeitet haben, zeigt die entsprechende Tabelle in Anhang B, S. 67.

Tabelle 4-3. Rücklaufquoten (Seminarteilnehmer, Kontrollgruppe und Führungskräfte) nach Eliminierung der unvollständigen Datensätzen

	t0 Basisbefragung (ca. 2 Wo. vor Seminar)	t1 Seminar	t2 1. Nachbefragung (nach 2-3 Wochen)	t3 2. Nachbefragung (nach 3 Monaten)
Trainees (N = 47)	44	47	34	44
Kontrollgruppe (N = 22)	21	-	11	11
Führungskräfte ^a	30	-	-	9

Anmerkungen. ^a N = 30 bzw. N = 9 gibt die Anzahl an einzelnen Fremdeinschätzungen durch die Führungskräfte wieder. Insgesamt waren es 17 Führungskräfte, die jeweils eine unterschiedliche Anzahl von Seminarteilnehmern einschätzen sollten. Es wurden keine Mitarbeiter der Kontrollgruppe eingeschätzt, um den Aufwand für die Führungskräfte zu minimieren.

Ein Einbruch in der Rücklaufquote ist bei den Seminarteilnehmern bei der ersten Nachbefragung zu verzeichnen, was möglicherweise an der langen Bearbeitungszeit von ca. 33 Minuten lag. Insgesamt haben 31 der 47 verbleibenden Trainees alle vier Befragungen bearbeitet. Die Beteiligung der Kontrollgruppe reduzierte sich nach der Basisbefragung um fast 50%. Aufgrund der geringen Beteiligungsquote der Führungskräfte bei der Fremdeinschätzung der Trainees zu t3 wurde diese Befragung nicht weiter verfolgt.

4.5.2 Statistische Auswertung

Für die Beantwortung der in Kapitel 3.1 bis 3.3 formulierten Hypothesen wurden die inferenzstatistischen Berechnungen mit SPSS[®] für Windows[®] durchgeführt.

4.5.2.1 Prüfung der Hypothesen

Die erste Fragestellung, bei der die Zusammenhänge zwischen Reaktionsebene, Lernebene (*Wissenstest, Einstellungsänderungen*) und Verhaltensebene (*subjektive Verhaltenseinschätzung*) untersucht werden, ist mittels korrelativer Analyse zu prüfen.

Zur Überprüfung der zweiten Fragestellung (Vorhersage des Seminarerfolgs durch Teilnehmerreaktionen) werden multiple Regressionsanalysen gerechnet. hinsichtlich der Ergebnisse im *Wissenstest*, in den *Einstellungsänderungen* und in der *subjektiven Verhaltenseinschätzung* bestimmt werden. Beim hier verwendeten schrittweisen Verfahren werden die einzelnen Teilnehmerreaktionen nacheinander in die Regressionsgleichung einbezogen. Ausgewählt wird dabei jeweils diejenige Variable, die zur Maximierung des

Bestimmtheitsmaßes R^2 beiträgt (Maß der Varianzaufklärung der Kriteriumsvariablen durch den/die Prädiktor/en). Da die Anzahl an Prädiktoren dieses Maß beeinflusst, wird zusätzlich das *adjustierte* R^2 angegeben, welches um eine entsprechende Korrekturgröße vermindert ist (Backhaus, Erichson, Plinke & Weiber, 2000). Die Korrelationsmatrix der Prädiktorvariablen kann erste Hinweise auf das Vorliegen von Multikollinearität liefern, Backhaus et al. (2000) schlagen jedoch als besseres Prüfkriterium den Toleranzwert der Regressoren vor. Eine weitere Statistik, die es im Rahmen der Regressionsanalyse zu beachten gilt, ist die des Durbin-Watson-Tests zur Überprüfung der Autokorrelation der Residualwerte. Aufgrund der vielen Prämissen mag die Anwendung einer Regressionsanalyse zwar begrenzt erscheinen, sie ist jedoch gegenüber kleineren Verletzungen relativ robust und daher gut anwendbar (Backhaus et al., 2000).

Bei der dritten Fragestellung gilt es, den Effekt der unter Kapitel 4.4.2 dargestellten Variablen auf Teilnehmerreaktionen, Wissenstest-Ergebnis, verhaltensbezogene Einstellungen und Maß der subjektiven Verhaltenseinschätzung zu überprüfen. Für jede Einflussgröße wird eine einfaktorielles multivariate Varianzanalyse (MANOVA) ohne Messwiederholung gerechnet. Dabei geht die jeweilige Einflussgröße als unabhängige Variable (Faktor) mit ihren Ausprägungen (Faktorstufen) ein. Als abhängige Variablen werden die unter Kapitel 4.4.1.5 ausgewählten sieben Teilnehmerreaktionen, das Ergebnis im Wissenstest, die verhaltensbezogenen Einstellungen sowie das Maß der subjektiven Verhaltenseinschätzung herangezogen. Zunächst wird der multivariate Einfluss des jeweiligen Faktors auf die Gesamtheit der abhängigen Variablen ermittelt. Im Anschluss an diese globale Unterschiedsprüfung wird für jede einzelne abhängige Variable geprüft, ob sich für den Faktor ein Unterschied zeigt (= univariater Effekt). Bei den dreistufigen Faktoren Transferklima und Kundekontakt wird mittels nachgeschobener univariater Paarvergleiche geklärt, welche Faktorstufen bei den abhängigen Variablen signifikant voneinander abweichen. Hierfür werden der Scheffé-Test, der als konservativ gilt und selbst bei unterschiedlich großen Gruppen verwendet werden kann, sowie bei Aufdeckung ungleicher Varianzen durch den Levene-Test entsprechend der Games-Howell-Test berechnet.

4.5.2.2 Effektstärke

Bei der Durchführung von Evaluationsstudien ist die Angabe von Effektstärken von großer Bedeutung (Lind, 2005). Bortz und Döring (1995, S. 568) zufolge sollte sogar „auf die Ermittlung der in einer Untersuchung tatsächlich erzielten Effektgröße [...] niemals verzichtet werden“. Im Gegensatz zu statistischen Signifikanzmaßen hängen die Effektstärken nicht von der Stichprobengröße ab, wodurch Vergleiche von Untersuchungen mit verschiedenen großen Stichprobenumfängen möglich werden. Da sie allerdings wie die Signifikanztests von der Standardabweichung beeinflusst werden, schlägt Lind (2005) die Bezeichnung der relativen Effektstärke vor. Für die Klassifikation von Effektgrößen gibt Cohen (1988) im Hinblick auf die in der vorliegenden Untersuchung anzuwendenden Verfahren folgende Empfehlungen.

Tabelle 4-4. Übersicht über Höhe der Effektstärken bei verschiedenen Signifikanztests nach Cohen (1988)

	Produkt-Moment-Korrelation	Varianzanalyse ^a	Multiple Regression
Höhe der Effektstärke	r	$f = \sigma_u / \sigma$	$F^2 = R^2 / (1 - R^2)$
Klein	.10	.10	.02
Mittel	.30	.25	.15
Groß	.50	.40	.35

Anmerkungen. ^a Diese Effektgröße lässt sich auch durch η^2 (dem Anteil der Gesamtvarianz der auf die unabhängige Variable zurückgeht) darstellen (Bortz & Döring, 1995). Berechnet wird dieser Wert anhand folgender Formel: $\eta^2 = f^2 / (1 + f^2)$, umgekehrt erhält man f aus η^2 wie folgt: $f = \sqrt{[\eta^2 / (1 - \eta^2)]}$.

Allgemein gesehen ist die Effektstärke eines Signifikanztests nur eine von vier relevanten Größen und ist mit den anderen Größen, nämlich der Stichprobengröße, dem Signifikanzniveau sowie der Teststärke wechselseitig verbunden.

Während in Tabelle 4-4 die Konventionen für die Klassifizierung von Effektstärken dargestellt wurden, wird das Signifikanzniveau i.d.R. mit $\alpha = .01$ bzw. $\alpha = .05$ festgelegt. Für die Teststärke scheint mit $\varepsilon = 1 - \beta = .80$ ebenfalls eine Konvention festzustehen (Bortz & Döring, 1995). In Bortz und Döring (1995) lassen sich für die einzelnen Signifikanztests bei den o.g. Werten verschiedene Tabellen zur Ermittlung des optimalen Stichprobenumfangs finden, Cohen (1988) bietet weitere Tabellen für andere Kombinationen.

Das Signifikanzniveau liegt in der vorliegenden Diplomarbeit bei $\alpha = .05$, allerdings werden auch Ergebnisse bis zu einer Irrtumswahrscheinlichkeit von 10% noch als marginal signifikant angesehen und entsprechend interpretiert. Bei den hier beschriebenen Verfahren wird zweiseitig auf Signifikanz getestet.

Die im Anschluss an die MANOVA durchgeführten einfaktoriellen univariaten Varianzanalysen sowie die paarweisen Vergleiche bergen das Problem der Kumulierung des Alpha-Fehlers. Im Falle mehrerer simultaner Tests wird daher die Bonferroni-Korrektur verwendet, bei der das ursprüngliche Alpha (α) durch die Anzahl an durchzuführenden simultanen Tests (k) dividiert wird ($\alpha^* = \alpha / k$). Das berechnete α^* bildet somit das neue Signifikanzniveau.

4.5.2.3 Auswertungen im Vorfeld

Bevor im nächsten Kapitel 5 auf die Ergebnisse der Datenauswertung im Hinblick auf die Hypothesen übergegangen wird, sollen zunächst noch einige Auswertungen vorgenommen werden. Zum einen erfolgt die Überprüfung eines Selektionseffektes, d.h. ob zwischen Seminarteilnehmern, die lediglich den Seminar-Feedbackbogen bearbeitet haben und Teilnehmern, die sich zusätzlich an einer der beiden Online-Nachbefragungen oder an beiden beteiligt haben, Unterschiede in ihren Einschätzungen bestehen. Zum anderen ist es notwendig, zu untersuchen, ob zwischen den Seminarteilnehmern und der Kontrollgruppe mögliche Unterschiede in den biographischen Variablen sowie im Wissenstest bestehen, um die Vergleichbarkeit der beiden Gruppen zu gewährleisten.

Um einen Selektionseffekt auszuschließen, wurde eine einfaktorielle multivariate Varianzanalyse (MANOVA) ohne Messwiederholung durchgeführt, mit dem zweistufigen Faktor „Bearbeitung“ (*nur Feedbackbogen bearbeitet* vs. *zusätzlich eine oder beide Nachbefragungen*) und als abhängige Variablen den sieben Reaktionen des Feedbackbogens (*Trainer allgemein, Veranstaltung allgemein, Rahmenbedingungen, Vorbereitung der Teilnehmer, Durchführung, Anwendbarkeit, Trainer differenziert*). Es zeigt sich kein signifikanter Haupteffekt, d.h. für den Seminar-Feedbackbogen unterscheiden sich die Einschätzungen der Seminarteilnehmer, die nur den obligatorischen Fragebogen bearbeiteten, nicht von den Einschätzungen derjenigen Teilnehmer, die eine oder beide der Nachbefragung bearbeitet hatten ($F_{(1,67)} = .71, p = .66$). Eine weitere Prüfung, etwa auf Unterschiede in den biographischen Daten (z.B. Alter, Dauer der Zugehörigkeit im Unternehmen

bzw. im derzeitigen Job etc.) war nicht möglich, da diese Variablen in der Basisbefragung erhoben wurden und bei den Trainees, die lediglich den Feedbackbogen ausgefüllt hatten, nicht vorlagen.

Für die Unterschiedsprüfung zwischen Seminarteilnehmern und Kontrollgruppe bezüglich der biographischen Variablen, wurden diese entsprechend ihrem jeweiligen Skalenniveau auf Signifikanz getestet. Für kategoriale (nominalskalierte) Daten wie Geschlecht und Familienstand bzw. ordinalskalierte Daten wie Schulabschluss, derzeitige Tätigkeit und Teilnahme an Vertriebs Schulungen wurden nicht-parametrische Testverfahren (Chi²-Test bzw. Mann-Whitney-U-Test) verwendet. Die intervallskalierten Variablen Alter, Dauer der Organisations- und Jobzugehörigkeit wurden mit einem t-Test für unabhängige Stichproben auf Mittelwertsunterschiede getestet.

Tabelle 4-5. *Ergebnisse der Unterschiedsprüfungen der erhobenen demographischen Variablen zwischen Seminarteilnehmern und Kontrollgruppe*

	Teilnehmer			Kontrollgruppe			Teststatistiken
	M	SD	N (%)	M	SD	N (%)	
Alter	37.81	4.92	N = 43	38.05	4.85	N = 22	$t_{(63)} = .18, p = .86$
Dauer im Unternehmen (in Jahren)	7.93	4.32	N = 43	8.68	4.17	N = 20	$t_{(61)} = .65, p = .52$
Dauer im derzeitigen Job (in Jahren)	3.49	3.03	N = 43	4.40	3.01	N = 20	$t_{(61)} = 1.11, p = .27$
Geschlecht^a			N = 44			N = 22	$\chi^2_{(1, N=66)} = 1.47, p = .23$ Fisher's exakter Test: $p = .28$
Weiblich			5 (11%)			5 (23%)	
Männlich			39 (89%)			17 (77%)	
Familienstand^a			N = 44			N = 22	$\chi^2_{(4, N=66)} = 5.59, p = .23$
alleinstehend			8 (18%)			1	
Lebensgemeinschaft			5 (11%)			6 (27%)	
verheiratet			25 (57%)			14 (64%)	
geschieden			1			0	
keine Angabe			5 (11%)			1	
Schulabschluss			N = 44			N = 22	$Z = -1.19, p = .23$
Hauptschule			1			1	
Mittlere Reife			3 (7%)			0	
FH			14 (32%)			6 (27%)	
Abitur			24 (55%)			15 (68%)	
keine Angabe			2 (4.5%)			0	

Vertriebstätigkeit	N = 44	N = 22	
reine Vertriebstätigkeit	5 (12%)	4 (18%)	
Tätigkeit mit Vertriebsanteilen	16 (36%)	5 (23%)	Z = -1.01, p = .31
keine Vertriebstätigkeit	22 (50%)	10 (46%)	
keine Angabe	1	3 (13%)	
Vertriebsschulung	N = 44	N = 22	
reine Vertriebsschulung	4 (9%)	2 (9%)	
Schulung mit Vertriebsanteilen	12 (27%)	4 (18%)	Z = -.32, p = .75
keine Vertriebsschulung	28 (64%)	14 (64%)	
keine Angabe	0	2 (9%)	

Anmerkungen: Angaben in absoluten Häufigkeiten (Prozent), getrennt für Teilnehmer und Kontrollgruppe. Prozentwerte beziehen sich auf die Stichprobengröße n der jeweiligen Versuchsgruppe.

^a Bei der Variable Geschlecht wurde aufgrund zu geringer Zellenbesetzung der exakte Test nach Fischer berechnet, der jedoch ebenfalls keine Signifikanz ergab.

t = Prüfgröße des t-Test für unabhängige Stichproben, χ^2 = Prüfgröße des χ^2 -Test nach Pearson, Z = Prüfgröße für den Mann-Whitney-U-Test (bei ausreichend großer Stichproben $N_{Ges} > 30$), p = Irrtumswahrscheinlichkeit.

Wie die Werte in Tabelle 4-5 verdeutlichen, ergibt die Unterschiedsprüfung zwischen Trainees und Kontrollgruppe eine gute Vergleichbarkeit der beiden Gruppen in den job- und seminarrelevanten Variablen.

Ebenso erfolgte eine Unterschiedsprüfung im Wissenstest bezüglich der sieben Themen *Verkauf durch Bedürfnisbefriedigung, Gesprächseröffnung, Fragetechnik, Unterstützungstechnik, Abschlusstechnik, Gleichgültigkeit überwinden* und *Umgang mit Kundeninwänden*. Nicht alle Trainees und alle Mitarbeiter der Kontrollgruppe bearbeiteten den Wissenstest, daher blieben zur inferenzstatistischen Überprüfung $n = 32$ Trainees sowie $n = 11$ Mitglieder der Kontrollgruppe. Die durchgeführte einfaktorielle MANOVA ohne Messwiederholung offenbart einen signifikanten Unterschied ($F_{(1,41)} = 5.12, p < .001$). Die Seminarteilnehmer erreichen im Durchschnitt mit ca. 83% richtiger Antworten ($M = 83.04, SD = 8.66$) ein signifikant höheres Ergebnis im Wissenstest als die Kontrollgruppe mit ca. 71% richtiger Antworten ($M = 71, SD = 4.35$). Mit einem $\eta^2 = .51$ lässt sich dieses Ergebnis nach Cohen (1988) als großen Unterschied interpretieren. Die beobachtete Teststärke ($1-\beta$) liegt dabei bei .99 unter Verwendung von $\alpha = .05$. Nach Bortz und Döring (1995, S.575) beträgt der optimale Stichprobenumfang $N = 65$, um bei einem α von .05 und einer Teststärke von $1-\beta = .80$ eine mittlere Effektstärke statistisch zu sichern. Für einen großen Effekt, wie er hier vorgefunden wurde, beträgt der optimale Stichprobenumfang $N = 26$.

Ein anschließender Vergleich der einzelnen Themen mittels univariater Varianzanalysen ergab einen signifikanten Gruppenunterschied in fünf der sieben Einzelthemen. So erzielten die Trainees im Vergleich zur Kontrollgruppe etwas bessere Ergebnisse bei den Themen *Verkauf durch Bedürfnisbefriedigung* ($F_{(1,41)} = 3.59, p < .10$), *Gesprächseröffnung* ($F_{(1,41)} = 3.94, p < .10$) sowie *Abschlusstechnik* ($F_{(1,41)} = 3.69, p < .10$). Bei den Themen *Fragetechnik* ($F_{(1,41)} = 18.58, p < .001$) und *Umgang mit Kundeneinwänden* ($F_{(1,41)} = 25.56, p < .001$) beantworten die Trainees dagegen deutlich mehr Fragen richtig als die Kontrollgruppe. Kein signifikanter Unterschied in der Anzahl richtiger Antworten zeigt sich bei den Themen *Unterstützungstechnik* ($F_{(1,41)} = .17, p = .67$) sowie *Gleichgültigkeit überwinden* ($F_{(1,41)} = 2.41, p = .13$). Der fehlender Unterschied in Bezug auf die *Unterstützungstechnik* kann nach Aussagen des Unternehmens dadurch erklärt werden, dass dieser Aspekt ein fester Bestandteil im Tätigkeitsbereich beider Gruppen ist und sie dieses Thema allein durch ihre Berufspraxis gleich gut beherrschen. Für den Bereich *Gleichgültigkeit überwinden* könnte der statistisch nicht signifikante Unterschied an den sehr hohen Streuungen liegen (siehe Tabelle 4-6).

Tabelle 4-6. Deskriptive Statistiken des Wissenstests (Seminarteilnehmer und Kontrollgruppe)

	Teilnehmer (N = 32)		Kontrollgruppe (N = 11)		F	p	η^2
	M	SD	M	SD			
Verkauf durch Bedürfnisbefriedigung	76.74	12.42	66.67	21.66	3.59	.07 [†]	.08
Gesprächseröffnung	85.16	14.16	75.76	11.46	3.94	.05 [†]	.09
Fragetechnik	84.38	13.54	65.04	10.39	18.58	.00 ^{***}	.31
Unterstützungstechnik	85.55	14.59	83.52	8.96	.17	.67	.01
Abschlusstechnik	72.16	11.30	63.64	16.26	3.69	.06 [†]	.08
Gleichgültigkeit überwinden	75.39	20.94	64.77	14.60	2.41	.13	.06
Kundeneinwände	89.00	7.81	73.91	10.47	25.56	.00 ^{***}	.38

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Aufgrund der überprüften und gegebenen Vergleichbarkeit bei den soziodemographischen Daten kann der gefundene Unterschied in der Anzahl richtiger Antworten als Wirkung des Seminars interpretiert werden.

5 Ergebnisse

In diesem Kapitel werden in der Reihenfolge der formulierten Hypothesen die entsprechenden Ergebnisse beschrieben und bei Signifikanz gegebenenfalls graphisch veranschaulicht. Weitere Tabellen bzw. Graphiken befinden sich im Anhang.

5.1 Zusammenhänge zwischen den Ebenen nach Kirkpatrick

Das Prüfergebnis der bivariaten Zusammenhänge zwischen den Teilnehmerreaktionen aller drei Messzeitpunkte und den Erfolgskriterien der anderen beiden Ebenen (d.h. Ergebnis des Wissenstests, Veränderungen in den verhaltensbezogenen Einstellungen sowie Verhalten) ist in Tabelle 5-1 wiedergegeben. Es wird deutlich, dass die Teilnehmerreaktionen aller drei Zeitpunkte vorwiegend mit dem subjektiv eingeschätzten Verhalten korrelieren.

Laut Tabelle 5-1 liegt die Mehrzahl der gefundenen Zusammenhänge zwischen Reaktionen und **Wissen** im negativen Bereich ($r = -.02$ bis $r = -.37$), es gibt nur sehr wenige positive Korrelationen ($r = .03$ bis $r = .19$). Es gibt weder signifikante Zusammenhänge zwischen der ersten noch der dritten Reaktionsmessung und dem Wissenstest. Für die zweite Reaktionsmessung zeigt sich lediglich eine einzige signifikante Korrelation mit dem Wissen, allerdings ist dieser Zusammenhang zwischen der Anzahl richtiger Antworten und dem Subtest *Erfahrungsaustausch t2* negativ ($r = -.37$, $p < .05$). Teilnehmer, die weniger mit dem Erfahrungsaustausch zufrieden sind, erzielen demnach höhere Werte im Wissenstest, wohingegen zufriedene Teilnehmer hier insgesamt weniger richtige Antworten erzielten.

Für die **verhaltensbezogenen Einstellungsänderungen** zeigt sich in Tabelle 5-1 bei den Teilnehmereinschätzungen der ersten Reaktionsmessung (t1) lediglich eine marginal signifikante Korrelation mit der *Vorbereitung der Teilnehmer* ($r = .26$, $p < .10$). Alle weiteren Einschätzungen korrelieren zu diesem Zeitpunkt sehr schwach und vorwiegend negativ ($r = -.17$ bis $r = .02$). Bei der zweiten Reaktionsmessung ergeben sich ebenfalls nur schwache positive Korrelationen ($r = .12$ bis $r = .18$). Für die dritte Reaktionsmessung zeigt sich neben einem marginal signifikanten positivem Zusammenhang mit der *Anwendung zu*

$t3$ ($r = .30$, $p < .10$) auch eine signifikante Korrelation mit dem *Nutzen zu $t3$* ($r = .31$, $p < .05$) sowie mit der *Umsetzung der persönlichen Ziele* ($r = .36$, $p < .05$).

Tabelle 5-1. Korrelationen zwischen den Teilnehmereinschätzungen des unmittelbaren Feedbackbogens ($t1$), der 1. Nachbefragung ($t2$) sowie der 2. Nachbefragung ($t3$) mit den Kriterien der Lern- und Verhaltensebene

	Richtige Antworten im Wissenstest (%)	Einstellungs- änderungen ^a	Verhalten
Erste Reaktionsmessung ($t1$)	$N = 34$	$N = 43$	$N = 44$
Trainer (allgemein)	.03	-.02	-.20
Veranstaltung (allgemein)	-.18	-.02	.00
Rahmenbedingungen	.07	-.04	.29[†]
Vorbereitung der Teilnehmer	-.09	.26[†]	.36*
Allgemeine Durchführung	.07	-.03	.23
Anwendbarkeit der Inhalte	-.07	.02	.51***
Trainer (differenziert)	-.02	-.17	-.15
Zweite Reaktionsmessung ($t2$)	$N = 34$	$N = 31$	$N = 31$
Anwendung $t2$	-.21	.18	.69***
Nutzen $t2$	-.13	.18	.67***
Erfahrungsaustausch	-.37*	.11	.30[†]
Seminarbewertung im Rückblick	-.24	.12	.24
Dritte Reaktionsmessung ($t3$)	$N = 34$	$N = 43$	$N = 44$
Anwendung $t3$	-.06	.30[†]	.67***
Nutzen $t3$	-.21	.31*	.67***
Umsetzung pers. Ziele	.19	.36*	.66***

Anmerkungen. ^a Differenz $t3-t0$ der verhaltensbezogenen Einstellungen des Evaluationsbogens.

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Die meisten Korrelationen zeigen sich, anders als für Wissen und Einstellungen, zwischen den Teilnehmerreaktionen und der **subjektiven Verhaltenseinschätzung**. Bei der ersten Reaktionsmessung ($t1$) korrelieren die Subtests *Vorbereitung der Teilnehmer* ($r = .37$, $p < .05$) und *Anwendbarkeit* ($r = .51$, $p < .001$) signifikant positiv mit der subjektiven Verhaltenseinschätzung. Zum Zeitpunkt der zweiten Reaktionsmessung ($t2$) zeigt sich eine marginal signifikante positive Korrelation mit dem Subtest *Erfahrungsaustausch* ($r = .30$, $p < .10$), wohingegen die Subtests *Anwendung $t2$* und *Nutzen $t2$* jeweils signifikant positiv mit der subjektiven Verhaltenseinschätzung korrelieren ($r = .69$ bzw. $r = .67$, $p < .001$). Bei

der dritten Reaktionsmessung (t3) zeigen alle Subtests signifikante Zusammenhänge zur subjektiven Verhaltenseinschätzung ($r = .66$ bis $r = .67$, $p < .001$).

Zusammenfassend kann die Nullhypothese 1-A („*Es bestehen keine positiven Korrelationen zwischen den Teilnehmerreaktionen und der Lernebene*“) nur zum Teil abgelehnt werden: In Bezug auf das gemessene *Wissen* lassen sich tatsächlich keine systematischen Korrelationen mit den Teilnehmerreaktionen finden. Anders ist dies jedoch für die *verhaltensbezogenen Einstellungen*, bei denen sich zwar nur schwache, aber dennoch signifikant positive Zusammenhänge mit zwei Variablen der dritten Reaktionsmessung (t3) zeigen. Die Nullhypothese 1-B („*Es bestehen keine positiven Korrelationen zwischen den Teilnehmerreaktionen und der Verhaltensebene*“) wird dagegen verworfen: Bei der subjektiven Verhaltenseinschätzung zeigen sich zu allen drei Zeitpunkten der Reaktionsmessung mehrere signifikante positive Korrelationen.

5.1.1 Lernebene und Verhaltensebene

Der in der Nebenfragestellung N-1 postulierte positive Zusammenhang zwischen der Anzahl richtiger Antworten im Wissenstest und dem Mittelwert der eingeschätzten Anwendung der Inhalte auf der Verhaltensebene ($r = -.05$, $p = .78$, $N = 30$) bleibt ebenso aus wie der erwartete Zusammenhang zwischen den verhaltensbezogenen Einstellungen und der subjektiven Verhaltenseinschätzung ($r = .09$, $p = .56$, $N = 42$).

Die Nullhypothese der Nebenfragestellung N-1 („*Es bestehen keine positiven Korrelationen zwischen der Lern- und der Verhaltensebene*“) kann daher nicht abgelehnt werden.

5.2 Vorhersage des Seminarerfolgs anhand der Teilnehmerreaktionen

Zur Überprüfung, welche Teilnehmerreaktionen am besten zur Vorhersage der Erfolgskriterien (Wissenstest, verhaltensbezogene Einstellungen und Verhalten) beitragen, werden insgesamt drei multiple Regressionsanalysen berechnet. Dabei gehen als Prädiktoren folgende sieben Teilnehmerreaktionen ein: *Durchführung t1*, *Anwendbarkeit t1*, *Anwendung zu t2*, *Nutzen zu t2*, *Anwendung zu t3*, *Nutzen zu t3* sowie *Umsetzung persönlicher Ziele t3*.

Um erste Hinweise auf Multikollinearität zu erhalten, soll zunächst die Matrix der Interkorrelationen der Prädiktoren betrachtet werden.

Tabelle 5-2. *Iteminterkorrelationen der Prädiktoren*

	1	2	3	4	5	6	7
1 Durchführung (t1)	-						
2 Anwendbarkeit (t1)	.39**	-					
3 Anwendung (t2)	.06	.45**	-				
4 Nutzen (t2)	.31 [†]	.37*	.84***	-			
5 Anwendung (t3)	.23	.60***	.51***	.58***	-		
6 Nutzen (t2)	.35*	.49***	.64**	.79***	.73***	-	
7 Umsetzung pers. Ziele (t3)	.15	.42**	.37*	.46**	.77***	.53***	-

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Es wird deutlich, dass einige Prädiktoren mittlere bis hohe Korrelationen untereinander aufweisen, wobei bei empirischen Daten jedoch stets eine gewisse lineare Abhängigkeit vorliegt (Backhaus et al., 2000). Solange das Maß der Toleranz (T) den kritischen Wert von .10 nicht unterschreitet, was eine Ineffizienz der Schätzung zur Folge hätte, werden die oben gefundenen Korrelationen daher in Kauf genommen.

5.2.1 Lernebene

Wie die in der korrelativen Analyse gefundenen niedrigen bzw. fehlenden Zusammenhänge bereits vermuten lassen, zeigt sich in der Regressionsanalyse für den **Wissenstest** kein signifikantes Ergebnis: Keine einzige der ausgewählten Variablen der Reaktionsebene (*Durchführung t1*, *Anwendbarkeit t1*, *Anwendung zu t2*, *Nutzen zu t2*, *Anwendung zu t3*, *Nutzen zu t3* sowie *Umsetzung persönlicher Ziele t3*) trägt zur Vorhersage des Ergebnisses im Wissenstest bei.

Die für die **verhaltensbezogenen Einstellungen** durchgeführte Regressionsanalyse zeigt ein signifikantes Ergebnis ($F_{(1,29)} = 5.00$, $p < .05$, $R^2 = .15$, $adj. R^2 = .12$). Von allen sieben Prädiktorvariablen leistet lediglich der Subtest *Umsetzung persönlicher Ziele t3* einen signifikanten Beitrag für die Vorhersage der Veränderungen in den verhaltensbezogenen Einstellungen ($\beta = .38$, $t_{(29)} = 2.24$, $p = .03$). Alle anderen Prädiktoren fanden keinen

Eingang in die Schätzgleichung. Tabelle 5-3 gibt die Ergebnisse der Regressionsanalyse wieder.

Tabelle 5-3. *Ergebnisse der Regressionsanalyse ausgewählter Prädiktoren auf der Reaktionsebene zur Vorhersage des Kriteriums **verhaltensbezogene Einstellungen** (N = 31)*

Variable	B	SE B	Beta (β)	T	VIF
Umsetzung persönlicher Ziele (t3)	.17	.07	.38*	1.00	.00
R^2	.15				
adjustiertes R^2	.12				
$F_{(1,29)} = 5.00$	$p < .05$				
Durbin-Watson	2.18				

* $p < .05$.

Eine Multikollinearität ist entsprechend den Werten der Toleranz und der VIF auszuschließen. Da die gefundene Durbin-Watson-Statistik $d = 2.18$ beträgt und erst Werte wesentlich größer bzw. kleiner 2 auf eine Autokorrelation hinweisen (Janssen & Laatz, 2003), wird eine solche ebenfalls ausgeschlossen. Obwohl das adjustierte R^2 mit .12 niedrig erscheint, muss beachtet werden, dass diese 12% Varianzaufklärung durch eine einzige Variable zustande kommt. Nach Cohen (1988, Tabelle 4-4) lässt sich die mit dem adjustierten R^2 berechnete Effektstärke mit .14 als mittlere Effektstärke einordnen.

Zusammenfassend kann die Nullhypothese 2-A („Die ausgewählten Prädiktoren leisten keinen Beitrag zur Vorhersage des Erfolgs auf Lernebene“) nur zum Teil abgelehnt werden. So zeigt sich im Hinblick auf den Wissenstest, ähnlich wie bei der korrelativen Analyse, kein signifikantes Ergebnis. Keine einzige der ausgewählten Facetten der Teilnehmerreaktionen vermag das Ergebnis im **Wissenstest** vorherzusagen. Bei den **verhaltensbezogenen Einstellungen** findet sich hingegen ein ausgewählter Prädiktor (*Umsetzung persönlicher Ziele t3*) mit einem Vorhersagewert hinsichtlich der Änderungen in den verhaltensbezogenen Einstellungen.

5.2.2 Verhaltensebene

Die Regressionsanalyse zur Vorhersage der *subjektiven Verhaltenseinschätzung* durch die ausgewählten Teilnehmerreaktionen ergibt ein signifikantes Ergebnis ($F_{(3,27)} = 30.61$,

$p < .001$). Dabei leistet die *Anwendung zu t2* den höchsten Beitrag zur Vorhersage der subjektiven Verhaltenseinschätzung ($\beta = .40$, $t_{(27)} = 3.75$, $p < .01$), gefolgt von der *Anwendung zu t3* ($\beta = .35$, $t_{(27)} = 2.16$, $p < .05$) und der *Umsetzung persönlicher Ziele* ($\beta = .31$, $t_{(27)} = 2.07$, $p < .05$). Die übrigen Prädiktoren (*Durchführung t1*, *Anwendbarkeit t1*, *Nutzen zu t2* und *Nutzen zu t3*) besitzen für dieses Kriterium keine Vorhersagekraft. Tabelle 5-4 zeigt die gefundenen Ergebnisse.

Tabelle 5-4. *Ergebnisse der Regressionsanalyse ausgewählter Prädiktoren auf der Reaktionsebene zur Vorhersage des Kriteriums Verhalten (N = 31)*

Variable	B	SE B	Beta (β)	T	VIF
Anwendung zu t3	.41	.19	.35*	.33	3.05
Anwendung zu t2	.59	.16	.40**	.74	1.34
Umsetzung persönlicher Ziele (t3)	.28	.14	.31*	.38	2.62
R^2	.77				
adjustiertes R^2	.75				
$F_{(3,27)} = 30.61$	$p < .001$				
Durbin-Watson	1.81				

* $p < .05$, ** $p < .01$.

Gemäß den T- und VIF-Werten liegt keine starke Multikollinearität vor, des Weiteren gibt es auch keine Hinweise für eine Autokorrelation (Durbin-Watson-Statistik $d = 1.81$). Die Vorhersagekraft des Modells ist dabei mit ca. 75% an erklärter Varianz sehr stark ($R^2 = .77$, $adj. R^2 = .75$). Nach Cohen (1988) steht dieser adjustierte R^2 -Wert für einen großen Effekt.

Die Nullhypothese 2-B („Die ausgewählten Prädiktoren leisten keinen Beitrag zur Vorhersage des Erfolgs auf Verhaltensebene“) wird abgelehnt. Für die subjektive Verhaltenseinschätzung zeigen sich signifikante Ergebnisse, d.h. einige der ausgewählten Prädiktoren tragen sehr stark zur Vorhersage der subjektiven Verhaltenseinschätzung bei.

5.3 Auswirkung möglicher Einflussgrößen auf die Evaluationsebenen

Für jede einzelne der unter Kapitel 4.4.2 dargestellten acht Einflussgrößen wurde eine einfaktorielte multivariate Varianzanalyse (MANOVA) ohne Messwiederholung berechnet. Dabei ging pro MANOVA jeweils eine Einflussgröße als Faktor und ihre jeweiligen Aus-

prägungen als Faktorstufen ein, während die Ergebnisse der drei gemessenen Ebenen Reaktionen, Lernen und Verhalten die abhängigen Variablen (Zielkriterien) darstellten. Diese bestehen aus den sieben ausgewählten Teilnehmereinschätzungen (*Durchführung t1, Anwendbarkeit t1, Anwendung zu t2, Nutzen zu t2, Anwendung zu t3, Nutzen zu t3 und Umsetzung persönlicher Ziele t3*), aus den beiden Ergebnissen der Lernebene (*Wissen und Einstellungen*) sowie der subjektiven Verhaltenseinschätzung. Aufgrund der unterschiedlichen Antwortformate wurden die Rohwerte zuvor z-transformiert. Vor der Darstellung der MANOVA-Ergebnisse werden zunächst die Korrelationen zwischen den oben genannten zehn Zielkriterien und den Faktoren betrachtet.

Tabelle 5-5. *Korrelationen zwischen Zielkriterien und Einflussgrößen (Faktoren der einzelnen einfaktoriellen MANOVAs)*

N = 31	Ausgangslage		Subj. Bedarf	Anw.-Mögl.	Expertise	Unterstützung		Bewertung	
	Moti.	SW				FK	AU	allg.	spez.
1 Durchführung (t1)	.15	.14	.33	.11	-.01	.31 [†]	.40*	.33 [†]	.38*
2 Anwendbarkeit (t1)	.42*	.28	.72***	.30 [†]	.19	.42*	.56**	.36*	.79***
3 Anwendung (t2)	.26	.16	.03	.30 [†]	-.17	.58**	.48**	-.14	.16
4 Nutzen (t2)	.17	.25	-.07	.28	-.23	.62***	.53**	-.15	.19
5 Anwendung (t3)	.43*	.29	.13	.38*	.16	.65***	.70***	.08	.60***
6 Nutzen (t3)	.24	.30	.18	.33 [†]	.04	.65***	.68***	-.05	.40*
7 Umsetzung pers. Ziele (t3)	.31 [†]	.17	-.11	.35 [†]	.36*	.53**	.65***	.02	.48**
8 Wissen	-.01	.03	-.12	-.17	.39*	.11	.00	-.20	-.06
9 verhaltensbezogene Einstellungen	-.16	-.52**	-.07	.11	.31 [†]	.14	.25	-.07	.05
10 subj. Verhaltenseinschätzung	.38*	.28	-.05	.39*	.11	.63**	.75***	-.14	.37*

Anmerkungen. Moti. = Motivation, SW = Selbstwirksamkeit, Anw.-Mögl. = Anwendungsmöglichkeit.

[†] $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

Wie Tabelle 5-5 zeigt, bestehen zwischen den zehn ausgewählten Zielkriterien viele mittlere und hohe Korrelationen. Diese sind als Hinweise dafür anzusehen, für welche Faktoren sich ein multivariater Effekt ergeben könnte. Ob und wie stark sich die Einschätzungen in den einzelnen AVs in Abhängigkeit des jeweiligen Faktors und seinen Stufen tatsächlich unterscheiden, wird nachfolgend geprüft.

5.3.1 Motivation

So zeigte der globale Signifikanztest durch die MANOVA für den zweistufigen Faktor Motivation (*hoch* vs. *niedrig*) keinen signifikanten Haupteffekt in den Einschätzungen der zehn Zielkriterien ($F_{(1,28)} = 1.63, p = .17$).

Die Nullhypothese 3-A („*Der Faktor Motivation hat keinen Einfluss auf die Zielkriterien*“) wird nicht verworfen.

Anschließende einfaktorielle univariate Varianzanalysen decken jedoch signifikante Unterschiede für die *Anwendbarkeit t1* ($F_{(1,28)} = 9.48, p < .01$)¹¹ und die *Anwendung zu t3* ($F_{(1,28)} = 3.13, p < .10$)¹² auf. Hierbei bewerten höher motivierte Teilnehmer die unmittelbare *Anwendbarkeit* signifikant höher ($M = .47, SD = .64$) als Teilnehmer mit einer geringeren Motivation ($M = -.54, SD = 1.01$). Drei Monate später geben die Teilnehmer mit der höheren Motivation an, die Inhalte tendenziell öfter angewendet zu haben ($M = .36, SD = 1.04$) als ihre geringer motivierten Kollegen ($M = -.28, SD = .92$). In Anhang B, S. 84, finden sich alle weiteren Ergebnisse dieser Auswertung.

5.3.2 Selbstwirksamkeit

Auch hier ergab die globale Prüfung durch die MANOVA keinen signifikanten Haupteffekt für den zweistufigen Faktor Selbstwirksamkeit ($F_{(1,28)} = 1.64, p = .17$). Die Ausgangslage der Selbstwirksamkeit (*hoch* vs. *niedrig*) hat demnach keinen Einfluss auf die Einschätzungen in den zehn Zielkriterien.

Auch diese Nullhypothese 3-B („*Der Faktor Selbstwirksamkeit hat keinen Einfluss auf die Zielkriterien*“) kann nicht verworfen werden.

Die im Anschluss durchgeführten einfaktoriellen univariaten Varianzanalysen ergeben jedoch einen signifikanten Unterschied zwischen den Gruppen bei den *verhaltensbezogenen Einstellungen* ($F_{(1,28)} = 6.62, p < .05$)¹³. Die Teilnehmer mit einer geringeren Selbstwirksamkeit weisen positive Veränderungen in den Einstellungen auf ($M = .46, SD = 1.15$), wohingegen sich Teilnehmer mit einer hohen Selbstwirksamkeit verschlechtern

¹¹ Nach Bonferroni-Korrektur ($\alpha^* = 0.01/2 = 0.005$) bleibt dieser Unterschied mit $p = .005$ signifikant (5%-Niveau).

¹² Nach Bonferroni-Korrektur ($\alpha^* = 0.05$) wird dieser Unterschied mit $p = .088$ nicht mehr signifikant.

¹³ Nach Bonferroni-Korrektur ($\alpha^* = 0.025$) bleibt dieser Unterschied mit $p = .016$ signifikant.

($M = -.51$, $SD = .89$). Alle weiteren Ergebnisse dieser Berechnung sind in Anhang B, S. 85, zu finden.

5.3.3 Subjektiver Bedarf

Für den zweistufigen Faktor Bedarf (*hoher vs. mittlerer bis geringer Bedarf*) zeigte sich bei der globalen Prüfung durch die MANOVA ein signifikanter Haupteffekt ($F_{(1,28)} = 4.01$, $p < .01$). Demnach unterscheiden sich die beiden Gruppen in ihren Einschätzungen bzw. Ergebnissen der zehn Zielkriterien deutlich. Die gefundene Effektstärke steht mit einem $\eta^2 = .68$ für einen großen Effekt. Um diesen Effekt bei $\alpha = .05$ und einer Teststärke von $1-\beta = .80$ statistisch zu sichern, sollte der Stichprobenumfang für einen großen Effekt mindestens $N = 21$ betragen (Bortz & Döring, 1995), was mit einem $N = 30$ gegeben ist.

Für diesen Einflussfaktor wird die Nullhypothese 3-C („*Der Faktor Bedarf hat keinen Einfluss auf die Zielkriterien*“) verworfen.

Wie die einfaktoriellen univariaten Varianzanalysen im Anschluss zeigen, besteht dieser Gruppenunterschied allerdings nur für den Subtest *Anwendbarkeit t1* ($F_{(1,28)} = 17.31$, $p < .001$)¹⁴. Während Teilnehmer mit einem mittleren bis geringem Bedarf im unmittelbaren Feedbackbogen keine spätere Anwendbarkeit der Inhalte sehen ($M = -.70$, $SD = .79$), gehen Teilnehmer, die für sich einen hohen Bedarf sehen, davon aus, dass sie die Seminarinhalte zukünftig gut anwenden können ($M = .52$, $SD = .82$). Darüber hinaus finden sich zwischen den beiden Gruppen keine weiteren signifikanten Unterschiede auf univariater Ebene. Auf eine graphische Darstellung der Ergebnisse wird an dieser Stelle verzichtet und auf Anhang B, S. 86, verwiesen.

5.3.4 Anwendungsmöglichkeit

Die MANOVA ergab keinen signifikanten Haupteffekt für den dreistufigen Faktor Kundenkontakt ($F_{(2,27)} = .60$, $p = .89$), d.h. die Einschätzungen in den zehn Zielkriterien werden nicht nur das Ausmaß des Kundenkontakts und dadurch der Anwendungsmöglichkeit beeinflusst.

Demnach kann die Nullhypothese 3-D („*Der Faktor Anwendungsmöglichkeit hat keinen Einfluss auf die Zielkriterien*“) nicht verworfen werden.

In den anschließenden einfaktoriellen univariaten Varianzanalysen ergeben sich bezüglich der drei Faktorstufen jedoch marginal signifikante Unterschiede. Beim Subtest *Nutzen zu t2* ist der Effekt so gering ($F_{(2,27)} = 2.59, p < .10$)¹⁵, dass sich die drei Gruppen zwar insgesamt voneinander unterscheiden, der nachgeschobene Scheffé-Test allerdings nicht signifikant wird. Dadurch wird zwar nicht deutlich, ob der Unterschied zwischen den Stufen *wenig* und *mittlerer Kundenkontakt* besteht oder vielmehr zwischen *wenig* und *viel Kundenkontakt* besteht. Den mittleren Einschätzungen zufolge beurteilen jedoch Teilnehmer mit wenig Kundenkontakt und somit wenig Anwendungsmöglichkeit den Nutzen des Seminars in der ersten Nachbefragung schlechter ($M = -.93, SD = .88$) als Teilnehmer mit mittlerem ($M = .04, SD = .85$) und mit viel Kundenkontakt ($M = .15, SD = 1.05$). Beim *Nutzen zu t3* ist der Unterschied ebenfalls marginal signifikant ($F_{(2,27)} = 2.76, p < .10$)¹⁶. Hier verdeutlicht der nachgeschobene Scheffé-Test, dass Teilnehmer mit wenig Anwendungsmöglichkeit den *Nutzen zu t3* tendenziell schlechter einschätzen ($M = -.95, SD = .66$) als Teilnehmer, die mittleren Kundenkontakt und somit mehr Anwendungsmöglichkeit haben ($M = .21, SD = .83$). Eine ausführliche Darstellung aller Ergebnisse ist Anhang B, S. 87, zu entnehmen.

5.3.5 Vorerfahrung

Die für den zweistufigen Faktor Expertise (Vorerfahrung mit Vertriebstätigkeiten) durchgeführte MANOVA ergab keinen signifikanten Haupteffekt ($F_{(1,28)} = 1.64, p = .17$). Bezüglich der zehn Zielkriterien unterscheiden sich Teilnehmer mit Vorerfahrung in Vertriebstätigkeiten nicht von unerfahrenen Kollegen.

Die Nullhypothese 3-E („Der Faktor Vorerfahrung hat keinen Einfluss auf die Zielkriterien“) kann nicht verworfen werden.

Bei den anschließenden einfaktoriellen univariaten Varianzanalysen ergeben sich jedoch signifikante Unterschiede. Abbildung 5-1 stellt die Unterschiede in den verschiedenen Einschätzungen graphisch dar.

¹⁴ Dieser Unterschied bleibt selbst bei Anwendung der Bonferroni-Korrektur ($\alpha^* = 0.0005$) signifikant.

¹⁵ Nach Bonferroni-Korrektur ($\alpha^* = 0.05$) wird dieser Unterschied mit $p = .094$ nicht mehr signifikant.

¹⁶ Nach Bonferroni-Korrektur ($\alpha^* = 0.05$) wird dieser Unterschied mit $p = .081$ nicht mehr signifikant.

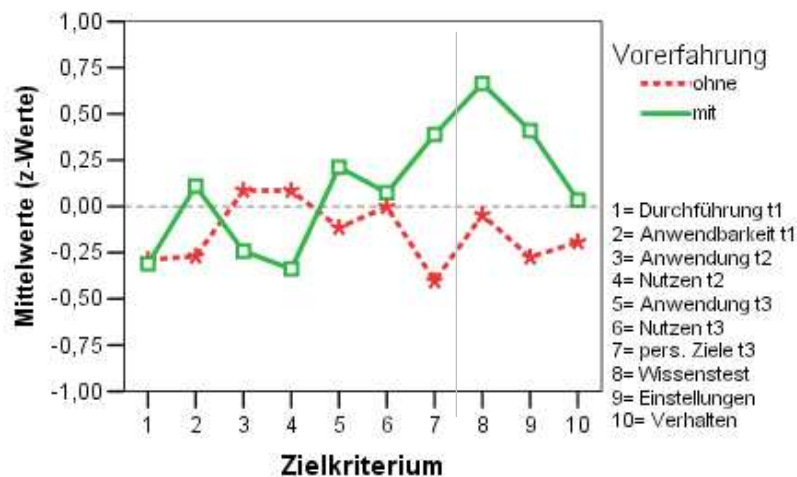


Abbildung 5-1. Z-standardisierte Mittelwerte der sieben Teilnehmereinschätzungen (1-7), des Wissenstests (8), der verhaltensbezogenen Einstellungen (9) und der subjektiven Verhaltenseinschätzung (10). Gruppierungsfaktor = Ausprägungen des Faktors *Vorerfahrung* („ohne“ bzw. „mit Vertriebsvorerfahrung“). Die vertikale Linie trennt zur besseren Übersicht die sieben Einschätzungen der Reaktionsebene von den drei Kriterien der Lern- bzw. Verhaltensebene.

Die ersten sechs Teilnehmereinschätzungen in Abbildung 5-1 verlaufen sehr uneinheitlich, die Einschätzungen beider Gruppen scheren jedoch ab der *Umsetzung persönlicher Ziele t3* deutlich auseinander. Obwohl der Verlauf der Teilnehmerreaktionen 1-6 augenscheinlich verschieden ist, ergeben sich hier keine statistischen Signifikanzen (siehe Tabelle 5-6). Erst bei der *Umsetzung persönlicher Ziele t3* ($F_{(2,27)} = 3.30, p < .10$)¹⁷, beim *Wissenstest* ($F_{(2,27)} = 4.91, p < .05$)¹⁸ sowie bei den *verhaltensbezogenen Einstellungen* ($F_{(2,27)} = 3.13, p < .10$)¹⁹ wird der Unterschied zwischen den Gruppen signifikant bzw. marginal signifikant. Bei der subjektiven Verhaltenseinschätzung nähern sich die Werte beider Gruppen wieder an, der Unterschied wird nicht mehr signifikant.

Teilnehmer ohne Vorerfahrung verfolgen demzufolge tendenziell ihre persönlichen Ziele weniger stark ($M = -.40, SD = 1.29$) als erfahrene Kollegen ($M = .30, SD = .70$) und schneiden im Wissenstest nicht so gut ab ($M = -.05, SD = 1.06$) wie Teilnehmer mit Vertriebsvorerfahrung ($M = .66, SD = .61$). In den verhaltensbezogenen Einstellungen zeigen sich bei Teilnehmern ohne Vorerfahrung keine Verbesserungen ($M = -.28, SD = .88$)

¹⁷ Bei Bonferroni-Korrektur ($\alpha^* = 0.05$) wird dieser Unterschied mit $p = .08$ nicht mehr signifikant.

¹⁸ Bei Bonferroni-Korrektur ($\alpha^* = 0.025$) wird der Unterschied mit $p = .04$ lediglich marginal signifikant.

¹⁹ Bei Bonferroni-Korrektur ($\alpha^* = 0.05$) wird der Unterschied mit $p = .09$ nicht mehr signifikant.

im Vergleich zu den Teilnehmern mit Vorerfahrung ($M = .45$, $SD = 1.35$). Alle weiteren Ergebnisse sind nachfolgend in Tabelle 5-6 dargestellt.

Tabelle 5-6. Deskriptive Statistiken (z-Werte) und Kennwerte der im Rahmen der MANOVA berechneten univariaten Varianzanalysen für den Einflussfaktor Expertise (Vorerfahrung)

Zielkriterium	Vorerfahrung				$F_{(1,28)}$	p	η^2
	ohne (n = 16)		mit (n = 14)				
	M	SD	M	SD			
1 Durchführung (t1)	-.29	.80	-.40	1.04	.11	.74	.00
2 Anwendbarkeit (t1)	-.27	1.08	.02	.93	.62	.44	.02
3 Anwendung (t2)	.09	1.01	-.21	1.03	.63	.43	.02
4 Nutzen (t2)	.09	.98	-.30	.93	1.19	.29	.04
5 Anwendung (t3)	-.12	1.17	.09	.79	.31	.59	.01
6 Nutzen (t3)	-.00	1.10	-.03	1.00	.01	.94	.00
7 Umsetzung pers. Ziele (t3)	-.40	1.29	.30	.70	3.30	.08[†]	.11
8 Wissen	-.05	1.06	.66	.61	4.91	.04*	.15
9 Einstellungen	-.28	.88	.45	1.35	3.13	.09[†]	.10
10 subj. Verhaltenseinschätzung	-.19	1.21	-.06	.84	.13	.73	.00

[†] $p < .10$, * $p < .05$.

5.3.6 Transferklima

Bei der einfaktoriellen MANOVA für den dreistufigen Faktor Transferklima (*negativ* vs. *positiv* vs. *uneinheitlich*) ergab sich ein signifikanter Haupteffekt ($F_{(2,27)} = 3.21$, $p < .01$). In Abhängigkeit davon, ob Teilnehmer über ein negatives, ein positives oder ein uneinheitliches Transferklima berichten, unterscheiden sie sich in ihren Einschätzungen und ihren Ergebnissen der zehn Zielkriterien. Mit $\eta^2 = .63$ handelt es sich um einen großen Effekt. Um diesen Effekt bei $\alpha = .05$ und einer Teststärke von $1-\beta = .80$ statistisch zu sichern, sollte der Stichprobenumfang mindestens $N = 21$ betragen (Bortz & Döring, 1995), was mit einem $N = 30$ gegeben ist.

Die Nullhypothese 3-F („Der Faktor Transferklima hat keinen Einfluss auf die Zielkriterien“) wird demzufolge verworfen.

Dieser signifikante Unterschied zeigt sich in den anschließenden einfaktoriellen univariaten Varianzanalysen für insgesamt acht der zehn Zielkriterien (sechs der sieben Teilnehmer-

reaktionen, Einstellungen und subjektive Verhalteneinschätzung) – im Wissenstest zeigt sich dahingegen kein signifikanter Unterschied. In Abbildung 5-2 sind die Verläufe der Einschätzungen aller drei Gruppen dargestellt.

Es wird deutlich, dass Teilnehmer mit einem positiven (günstigen) Transferklima bis auf den Wissenstest in allen Zielkriterien die höchsten Werte aufweisen, wohingegen Teilnehmer mit einem negativen (ungünstigen) Transferklima durchgehend im negativen Bereich liegen. Teilnehmer, die über ein uneinheitliches Transferklima berichten, liegen mit ihren Werten zwischen den beiden anderen Gruppen.

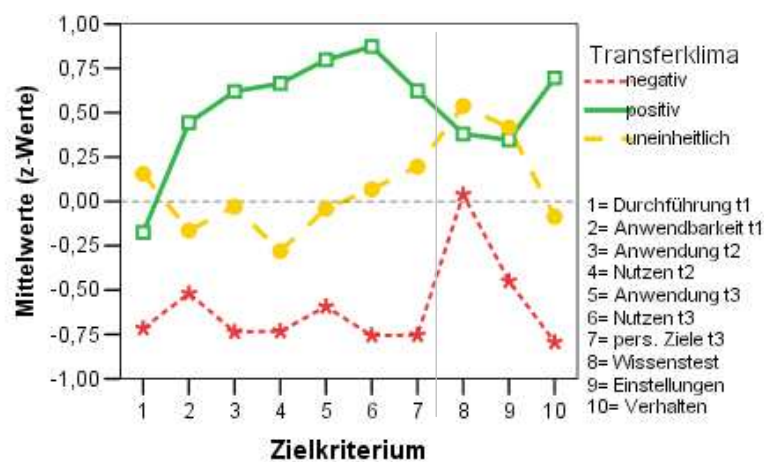


Abbildung 5-2. Z-standardisierte Mittelwerte der sieben Teilnehmereinschätzungen (1-7), des Wissenstests (8), der verhaltensbezogenen Einstellungen (9) und der subjektiven Verhaltenseinschätzung (10). Gruppierungsfaktor = Ausprägungen des Faktors *Transferklima* („negativ“, „uneinheitlich“ bzw. „positiv“). Die vertikale Linie trennt zur besseren Übersicht die sieben Einschätzungen der Reaktionsebene von den drei Kriterien der Lern- bzw. Verhaltensebene.

Für die Einschätzungen der *Anwendung zu t2* zeigt sich ein signifikanter Unterschied ($F_{(2,27)} = 9.21, p < .01$)²⁰. Wie der nachgeschobene Scheffé-Test verdeutlicht, wenden Teilnehmer mit einem negativen Transferklima die Seminarinhalte deutlich weniger an ($M = -.74, SD = .87$) als Teilnehmer mit einem positiven Transferklima ($M = .75, SD = .78$). Ein weiterer signifikanter Unterschied ergibt sich bei der Einschätzung des *Nutzens zu t2* ($F_{(2,27)} = 14.08, p < .001$)²¹: Dem Scheffé-Test zufolge sehen Teilnehmer mit

²⁰ Nach Bonferroni-Korrektur bleibt der Unterschied signifikant.

²¹ Nach Bonferroni-Korrektur bleibt der Unterschied höchst signifikant.

ungünstigem Transferklima in der Maßnahme keinen Nutzen ($M = -.73$, $SD = .83$), während Kollegen mit einem günstigen Transferklima sehr wohl einen Nutzen angeben ($M = .83$, $SD = .67$). In der *Anwendung zu t3* zeigt der nachgeschobene Scheffé-Test einen signifikanten Unterschied ($F_{(2,27)} = 5.98$, $p < .01$)²². Teilnehmer mit einem negativen Transferklima berichten auch drei Monate nach dem Seminar eine geringere Anwendungshäufigkeit der Inhalte ($M = -.59$, $SD = 1.09$) als ihre Kollegen mit positiven Transferklima ($M = .69$, $SD = .73$). Gleiches gilt für den Subtest *Nutzen zu t3* ($F_{(2,27)} = 10.16$, $p < .01$)²³, bei dem Teilnehmer, die über ein ungünstiges Transferklima berichten, weniger Nutzen in der Maßnahme sehen ($M = -.76$, $SD = 1.01$) als Teilnehmer mit einem günstigen Transferklima ($M = .81$, $SD = .64$). Schließlich unterscheiden sich die drei Transferklima-Gruppen in der *Umsetzung persönlicher Ziele t3* signifikant ($F_{(2,27)} = 5.15$, $p < .05$)²⁴. Der nachgeschobene Games-Howell-Test zeigt, dass in einem negativen Transferklima Teilnehmer ihre persönlichen Ziele deutlich weniger umsetzen ($M = -.75$, $SD = 1.29$) als Kollegen mit einem positivem Transferklima ($M = .52$, $SD = .72$). Während sich die drei Teilgruppen weder im Ergebnis des *Wissenstests* ($F_{(2,27)} = .75$, $p = .482$) noch in den verhaltensbezogenen Einstellungen ($F_{(2,27)} = 2.10$, $p = .142$) signifikant unterscheiden, offenbart sich zwischen den drei Gruppen ein signifikanter Unterschied in der subjektiven Verhaltenseinschätzung ($F_{(2,27)} = 7.52$, $p < .01$)²⁵. Im Gegensatz zu Teilnehmern, die ein ungünstiges Transferklima angeben ($M = -.79$, $SD = 1.05$), berichten Teilnehmer mit einem günstigen Transferklima, die Seminarinhalte öfter angewendet zu haben ($M = .64$, $SD = .75$).

Die nachgeschobenen Tests ergaben bisher vor allem Unterschiede zwischen den Gruppen mit *negativem* und *positivem* Transferklima. Für die Teilnehmergruppe mit *uneinheitlichem* Transferklima zeigt der Scheffé-Test einen signifikanten Unterschied beim *Nutzen zu t2*. Im Vergleich zu Teilnehmern mit durchweg positivem Transferklima ($M = .83$, $SD = .67$) schätzen Teilnehmer mit uneinheitlichem Transferklima den Nutzen des Seminars nach etwa 14 Tagen geringer ein ($M = -.28$, $SD = .46$). Bei der *Umsetzung persönlicher Ziele t3* zeigt der Scheffé-Test, dass die Teilnehmer mit einem uneinheitlichen Transferklima ihre

²² Nach Bonferroni-Korrektur wird der Unterschied auf 5%-Niveau signifikant.

²³ Nach Bonferroni-Korrektur bleibt der Unterschied hoch signifikant.

²⁴ Nach Bonferroni-Korrektur bleibt der Unterschied signifikant.

Ziele zwar tendenziell besser verfolgen können ($M = .20$, $SD = .58$) als Teilnehmer mit negativem Transferklima ($M = -.75$, $SD = 1.29$), allerdings können die Teilnehmer mit einem positiven Klima ($M = .52$, $SD = .72$) dies noch besser. Alle Detailergebnisse sind in Tabelle 5-7 dargestellt (siehe Anhang B, S. 89ff, für eine Übersicht der Rohwerte).

Tabelle 5-7. Deskriptive Statistiken (z-Werte) und Kennwerte der im Rahmen der MANOVA berechneten univariaten Varianzanalysen für den Einflussfaktor Transferklima

Zielkriterium	Transferklima						F _(2,27)	p	η^2
	negativ (n = 12)		positiv (n = 10)		uneinheitlich (n = 8)				
	M	SD	M	SD	M	SD			
1 Durchführung (t1)	-.72	.70	-.29	.99	.16	.90	2.51	.10	.16
2 Anwendbarkeit (t1)	-.52	1.14	.35	.71	-.16	.95	2.24	.13	.14
3 Anwendung (t2)	-.74	.87	.75	.78	-.03	.75	9.21	.001**	.41
4 Nutzen (t2)	-.73	.83	.83	.67	-.28	.46	14.08	.000***	.51
5 Anwendung (t3)	-.59	1.09	.69	.73	-.04	.58	5.98	.007**	.31
6 Nutzen (t3)	-.76	1.01	.81	.64	.07	.65	10.16	.001**	.43
7 Umsetzung pers. Ziele (t3)	-.75	1.29	.52	.72	.20	.58	5.15	.013*	.28
8 Wissen	.04	1.21	.38	.67	.54	.73	.75	.48	.05
9 Einstellungen	-.45	.93	.40	1.46	.42	.85	2.10	.14	.14
10 subj. Verhaltenseinschätzung	-.79	1.05	.64	.75	-.09	.64	7.52	.003**	.36

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

5.3.7 Allgemeine Seminarbewertung

Für den zweistufigen Faktor Allgemeine Seminarbewertung (*sehr gut vs. mittel bis gut*) zeigte die durchgeführte MANOVA in Bezug auf die zehn Zielkriterien keinen signifikanten Haupteffekt ($F_{(1,28)} = 1.83$, $p = .12$).

Die Nullhypothese 3-G („Der Faktor allgemeine Seminarbewertung hat keinen Einfluss auf die Zielkriterien“) kann aufgrund ausbleibender Signifikanzen nicht verworfen werden.

Ein interessanter signifikanter Unterschied ergibt sich bei den anschließenden einfaktoriellen univariaten Varianzanalysen für die subjektive Verhaltenseinschätzung ($F_{(1,28)} = 4.68$,

²⁵ Nach Bonferroni-Korrektur wird der Unterschied noch auf 5%-Niveau signifikant.

$p < .05$)²⁶. Diejenigen Teilnehmer, die das Seminar in der unmittelbaren Einschätzung lediglich mittel bis gut einschätzten, wenden die Themen nach ca. drei Monaten signifikant häufiger an ($M = .23$, $SD = 1.12$) als Teilnehmer mit einer sehr guten Seminarbewertung ($M = -.54$, $SD = .78$). Aus Anhang B, S. 88, sind alle weiteren Ergebnisse zu entnehmen.

5.3.8 Spezifische Seminarbewertung

Beim zweistufigen Faktor Spezifische Seminarbewertung (*sehr gut* vs. *mittel bis gut*) ergab die MANOVA einen signifikanten Haupteffekt ($F_{(1,28)} = 5.64$, $p < .01$). Die zehn Zielkriterien unterscheiden sich demnach deutlich in Abhängigkeit der Anwendbarkeit der Inhalte. Es liegt mit $\eta^2 = .75$ ein großer Effekt vor (vgl. Bortz & Döring, 1995).

Die Nullhypothese 3-H („Der Faktor spezifische Seminarbewertung hat keinen Einfluss auf die Zielkriterien“) wird verworfen.

Außer beim Subtest *Durchführung* t1 unterscheiden sich die beiden Gruppen bei der anschließend durchgeführten einfaktoriellen univariaten Varianzanalyse in allen anderen ausgewählten Teilnehmerreaktionen signifikant. In der subjektiven Verhaltenseinschätzung zeigen sich ebenfalls signifikante Unterschiede, im Wissenstest und in den Einstellungen dagegen nicht. Den Verlauf der Einschätzungen zeigt Abbildung 5-3.

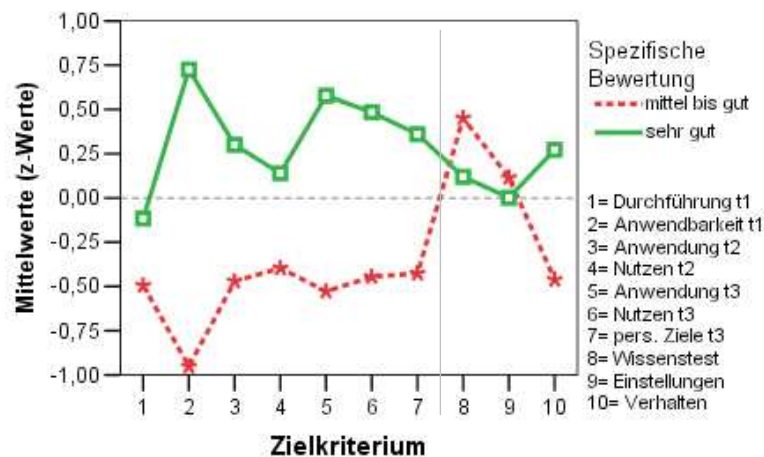


Abbildung 5-3. Z-standardisierte Mittelwerte der sieben Teilnehmereinschätzungen (1-7), des Wissenstests (8), der verhaltensbezogenen Einstellungen (9) und der subjektiven Verhaltenseinschätzung (10). Gruppierungsfaktor = Ausprägungen des Faktors *spezifische Seminarbewertung* („mittel bis gut“ bzw. „sehr gut“). Die vertikale Linie trennt zur besseren Übersicht die sieben Einschätzungen der Reaktionsebene von den drei Kriterien der Lern- bzw. Verhaltensebene.

²⁶ Nach Bonferroni-Korrektur wird der Unterschied nicht mehr signifikant.

Laut Abbildung schneiden Teilnehmer, die bereits unmittelbar nach dem Seminar für jedes Thema eine hohe Anwendbarkeit der Inhalte vorhersehen, d.h. eine sehr gute spezifische Seminarbewertung abgeben, allgemein besser ab als diejenigen Teilnehmer mit einer mittleren bis guten Bewertung. Lediglich für den Wissenstest sowie die verhaltensbezogenen Einstellungen scheint der Verlauf zu „kippen“: Die Teilnehmer mit einer mittleren bis guten Bewertung erzielen laut Abbildung 5-3 zwar augenscheinlich bessere Werte, allerdings zeigt sich hier keine statistische Signifikanz (vgl. Tabelle 5-8).

Für den Subtest *Anwendbarkeit* t1 besteht ein signifikanter Unterschied ($F_{(1,28)} = 61.26$, $p < .001$)²⁷. Sind sich die Teilnehmer sicher, die spezifischen Seminarthemen in der Praxis umsetzen zu können, ist konsequenterweise auch die Einschätzung der Anwendbarkeit höher ($M = .68$, $SD = .53$) als bei Teilnehmern, die die Umsetzung der Inhalte in die Praxis nicht so deutlich sehen ($M = -.95$, $SD = .61$). Ein weiterer signifikanter Unterschied zeigt sich für die *Anwendung* zu t2 ($F_{(1,28)} = 6.01$, $p < .05$)²⁸. Teilnehmer, die bereits unmittelbar nach dem Seminar die Wahrscheinlichkeit zur Umsetzung der Themen als geringer einschätzen, haben zum Zeitpunkt der ersten Nachbefragung die Inhalte auch weniger oft angewendet ($M = -.47$, $SD = .99$) als Teilnehmer mit einer höheren Einschätzung der Umsetzungswahrscheinlichkeit ($M = .37$, $SD = .88$). Der *Nutzen* zu t2 ergibt eine marginale Signifikanz zwischen beiden Gruppen ($F_{(1,28)} = 3.24$, $p < .10$)²⁹. Hierbei beurteilen Teilnehmer mit einer geringeren Einschätzung der Umsetzungswahrscheinlichkeit den Nutzen der Maßnahme etwas schlechter ($M = -.40$, $SD = .74$) als Teilnehmer, die nach dem Seminar sehr gute Umsetzungsmöglichkeiten für die einzelnen Themen sehen ($M = .21$, $SD = 1.08$). Für die *Anwendung* zu t3 wird der Unterschied signifikant ($F_{(1,28)} = 10.14$, $p < .01$)³⁰. Nach etwa drei Monaten geben diejenigen Teilnehmer, die schon nach dem Seminar keine guten Chancen zur Umsetzung der Inhalte vorhersahen, die Anwendung zu diesem späteren Zeitpunkt ebenfalls als seltener an ($M = -.53$, $SD = .89$) – im Vergleich zu den Teilnehmern, die die Umsetzung der Inhalte in die Praxis von Beginn an als sehr gut einstufen ($M = .49$, $SD = .86$). Im Hinblick auf den *Nutzen* zu t3 unterscheiden sich die

²⁷ Nach Bonferroni-Korrektur bleibt der Unterschied signifikant (1%-Niveau).

²⁸ Nach Bonferroni-Korrektur wird der Unterschied nur noch marginal signifikant (10%-Niveau).

²⁹ Nach Bonferroni-Korrektur wird der Unterschied nicht mehr signifikant.

³⁰ Nach Bonferroni-Korrektur bleibt der Unterschied signifikant (5%-Niveau).

beiden Gruppen ebenfalls signifikant ($F_{(1,28)} = 6.04, p < .05$)³¹: Teilnehmer, welche die Umsetzungswahrscheinlichkeit nach dem Seminar nicht so gut einstufen, sahen den Nutzen zum Zeitpunkt der zweiten Nachbefragung – wie bereits bei der ersten Nachbefragung – als geringer an ($M = -.45, SD = .83$) als ihre Kollegen, die gute Umsetzungsmöglichkeiten sahen ($M = .41, SD = 1.07$).

Des Weiteren unterscheiden sich die Teilnehmer in Abhängigkeit der spezifischen Bewertung marginal in der *Umsetzung persönlicher Ziele* t3 ($F_{(1,28)} = 3.30, p < .10$)³². Wer von den Teilnehmern eine gute Umsetzungsmöglichkeit voraussah, verfolgte im Gegensatz zu den Teilnehmer mit weniger guten Umsetzungschancen ($M = -.43, SD = 1.24$) seine Ziele stärker ($M = .28, SD = .85$). Ebenso gaben diese Teilnehmer nach drei Monaten an, die Inhalte tendenziell häufiger angewendet zu haben ($M = .20, SD = 1.06$) als Kollegen, die nach der Maßnahme keine so guten Umsetzungsperspektiven sahen ($M = -.46, SD = .93$), $F_{(1,28)} = 3.33, p < .10$ ³³.

Tabelle 5-8. Deskriptive Statistiken (z-Werte) und Kennwerte der im Rahmen der MANOVA berechneten univariaten Varianzanalysen für den Einflussfaktor spezifische Seminarbewertung

Zielkriterium	Spezifische Seminarbewertung				$F_{(1,28)}$	p	η^2
	mittel bis gut (n = 15)		sehr gut (n = 15)				
	M	SD	M	SD			
1 Durchführung (t1)	-.49	.89	-.19	.92	.86	.36	.03
2 Anwendbarkeit (t1)	-.95	.61	.68	.53	61.26	.000***	.69
3 Anwendung (t2)	-.47	.99	.37	.88	6.01	.02*	.18
4 Nutzen (t2)	-.40	.74	.21	1.08	3.24	.08†	.10
5 Anwendung (t3)	-.53	.89	.49	.86	10.14	.004**	.27
6 Nutzen (t3)	-.45	.83	.41	1.07	6.04	.02*	.18
7 Umsetzung pers. Ziele (t3)	-.43	1.24	.28	.85	3.30	.08†	.11
8 Wissen	.45	.66	.12	1.15	.95	.34	.03
9 Einstellungen	.12	1.45	.01	.83	.06	.81	.00
10 subj. Verhaltenseinschätzung	-.46	.93	.20	1.06	3.33	.08†	.11

† $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

³¹ Nach Bonferroni-Korrektur wird der Unterschied nur noch marginal signifikant (10%-Niveau).

³² Nach Bonferroni-Korrektur wird der Unterschied nicht mehr signifikant.

Aufgrund des geringen Stichprobenumfangs standen in dieser Untersuchung einfaktorielle Analysen zur Überprüfung möglicher Einflüsse der Faktoren auf die abhängigen Variablen im Vordergrund. Obwohl die in Tabelle 5-1 dargestellten Korrelationen mögliche Wechselwirkungen zwischen den Einflussvariablen erwarten lassen, wurden keine zweifaktoriellen Analysen berechnet, denn es sollten gerade die einzelnen Einflüsse der aus dem Tannenbaum-Modell (Cannon-Bowers et al., 1995) entnommenen Einflussgrößen auf die Evaluationsebenen nach Kirkpatrick (1996) geprüft werden. Es ist zwar zu vermuten, dass tatsächlich Interaktionen bestehen, angesichts der Komplexität der Daten sollte jedoch der Fokus auf den Hauptfragestellungen verweilen und auf weitere Auswertungen verzichtet werden.

5.4 Folgeanalysen

Im folgenden Abschnitt werden einerseits die Ergebnisse nachgeschobener Analysen in Bezug auf die formulierten Hypothesen dargestellt. Andererseits finden hier weitere Auswertungen Eingang, deren zugrundeliegende Fragestellungen zwar nicht explizit als Hypothesen formuliert worden waren, sich aber im Verlauf der Auswertung ergaben. Da diese Fragestellungen zumindest im theoretischen Teil Eingang gefunden hatten, hat auch ihre Beantwortung für die Hauptfragestellungen dieser Arbeit einen entsprechenden Stellenwert und werden daher mittels Folgeanalysen ausgewertet.

5.4.1 Konsequenzen der Trennung in *affective* und *utility reactions*

Verschiedentlich ist in der fehlenden Ausdifferenzierung von *affective* und *utility reactions* die Ursache geringer Korrelationen der Reaktionen mit den weiteren Ebenen gesehen worden (Morgan & Casper, 2000; Warr & Bunce, 1995). Da in dieser Untersuchung diese beiden Facetten ausdifferenziert wurden, sollten in einer nachgeschobenen Analyse die Auswirkungen dieser Trennung auf die Höhe der Korrelationen überprüft werden.

Die einzelnen Korrelationen wurden zunächst anhand der entsprechenden Tabelle (z.B. Bortz, 1999, S. 786) in Fishers Z-Werte transformiert. Zur Berechnung der

³³ Nach Bonferroni-Korrektur wird der Unterschied nicht mehr signifikant.

durchschnittlichen Korrelationen ist, da bis auf eine Ausnahme unterschiedlich große Stichprobenumfang vorliegen, nach Bortz (1999, S. 210) folgende Formel anzuwenden:

$$\bar{Z} = \frac{\sum_{j=1}^k (n_j - 3) * Z_j}{\sum_{j=1}^k (n_j - 3)}$$

Hierbei sind Z_j die Fishers Z-Werte der zu mittelnden Korrelationen und n_j die jeweiligen Stichprobenumfänge. Dann ist der \bar{Z} -Wert anhand der Tabelle (z.B. Bortz, 1999, S. 786) in den dazugehörigen durchschnittlichen Korrelationswert zurück zu transformieren.

Neben der Normalverteilung der Werte ergibt diese Transformation den Vorteil, dass die Werte die Eigenschaften einer Kardinalskala³⁴ annehmen, wodurch reelle Vergleiche zwischen den Werten möglich werden.

Tabelle 5-9. *Korrelationen zwischen den drei Ebenen bei Ausdifferenzierung der Teilnehmerreaktionen in affektive Reaktionen (affective reactions) und Einschätzungen zu Anwendung/Nutzen (utility reactions)*

	Wissen	Einstellungen	Verhalten
Reaktionen t1	<i>N = 32</i>	<i>N = 43</i>	<i>N = 43</i>
Affective			
Trainer (allg.)	(.03) .03	(-.02) -.02	(-.20) -.20
Veranstaltung (allg.)	(-.18) -.18	(-.02) -.02	(.00) .00
Rahmenbedingungen	(.07) .07	(-.04) -.04	(.29) .30
Vorbereitung der TN	(-.09) -.09	(.26) .27	(.36) .39
Allg. Durchführung	(.07) .07	(-.03) -.03	(.23) .23
Trainer (diff.)	(-.02) -.02	(-.17) -.17	(-.15) -.15
Utility			
Anwendbarkeit t1	(-.07) -.07	(.02) .02	(.51) .56
Reaktionen t2	<i>N = 32</i>	<i>N = 31</i>	<i>N = 31</i>
Affective			
Erfahrungsaustausch	(-.37) -.39	(.12) .12	(.30) .31
Rückblickende Seminarbewertung	(-.24) -.25	(.12) .12	(.24) .25
Utility			
Anwendung t2	(-.21) -.21	(.18) .18	(.69) .85
Nutzen t2	(-.13) -.13	(.18) .18	(.67) .81

³⁴ Kardinalskala ist der Oberbegriff für Intervall- und Verhältnisskala (vgl. Bortz, 1995, S. 617).

	Wissen	Einstellungen	Verhalten
Reaktionen t3	N = 30	N = 42	N = 43
Affective	(-)	(-)	(-)
Utility			
Anwendung t3	(-.06) -.06	(.30) .31	(.67) .81
Nutzen t3	(-.21) -.21	(.31) .32	(.67) .81
Umsetzung pers. Ziele	(.19) .19	(.36) .38	(.66) .79

Durchschnittliche Korrelationen^a			
affective (t1 bis t3)	-0.09	.09	.51
utility (t1 bis t3)	-.17	.48	.92

Anmerkungen. Werte in Klammern sind die Rohwerte aus Tabelle 5-1 (S. 76), fett gedruckte Werte stellen die Korrelationswerte als Fishers Z-Werte dar.

^a Durchschnittliche Korrelationen wurden nach der oben genannten Formel berechnet und anhand einer Tabelle für Fishers Z-Werte in Korrelationen rücktransformiert. Lediglich die durchschnittliche Korrelation zwischen *affective reactions* (t1 bis t3) und Wissenstest konnte anders berechnet werden, da hier die Stichprobengröße gleich war. Hierfür wurde der Median der Werte herangezogen.

Für beide Facetten bleiben signifikante Korrelationen mit dem Wissenstest aus, tendenziell zeigen sich hier vorwiegend schwache negative Korrelationen ($r_{affective} = -.09$ vs. $r_{utility} = -.17$). Im Vergleich zu den *affective reactions* findet sich eine etwas stärkere Korrelation der *utility reactions* mit den Einstellungen ($r_{affective} = .09$ vs. $r_{utility} = .48$). Aufgrund der Vergleichbarkeit durch die Fishers Z-Wert-Transformation kann festgehalten werden, dass die Korrelation zwischen *utility reactions* ($Z = 0.52$) und den Einstellungen deutlich höher ist als die der *affective reactions* ($Z = 0.09$), nämlich knapp sechs mal so hoch. Am auffälligsten scheren die Korrelationen in Bezug auf die subjektive Verhaltenseinschätzung auseinander: Hier zeigen die Z-Werte, dass der Zusammenhang zwischen *utility reactions* und der subjektiven Verhaltenseinschätzung ($Z = 1.53$) knapp drei mal höher ist als der der *affective reactions* ($Z = 0.56$). Bei Rücktransformation der Z-Werte in Korrelationen erhält man $r_{affective} = .51$ und $r_{utility} = .92$.

5.4.2 Nachgeschobene Regressionsanalysen

Da keine der sieben ausgewählten Teilnehmerreaktionen einen Beitrag zur Vorhersage des Ergebnisses im **Wissenstest** leisten konnte (vgl. Abschnitt 5.2), wurden anschließend weitere Regressionsanalysen durchgeführt. Auf diese Weise sollte nachträglich pro

Reaktionsmessung untersucht werden, ob sich unter den nicht-ausgewählten Reaktionen eventuell Prädiktoren für das Wissen identifizieren lassen und wie stark ihr Beitrag zur Vorhersage gewesen wäre. In diese drei Analysen gingen pro Messzeitpunkt alle Reaktionen des jeweiligen Zeitpunkts als Prädiktoren sowie als Kriterium das Ergebnis im Wissenstest ein. Keine der Reaktionen des Feedbackbogens (t1) noch der zweiten Nachbefragung (t3) konnten das Ergebnis des Wissenstests vorhersagen. Für die erste Nachbefragung (t2) diente lediglich der *Erfahrungsaustausch* t2 mit 14 % Varianzaufklärung als einziger Prädiktor (eine ausführliche Übersicht dieser nachträglichen Auswertung findet sich in Anhang B, S. 77).

In Bezug auf die **Einstellungsänderungen** wurden ebenfalls drei Regressionsanalysen nachgeschoben. Auch hier gingen pro Messzeitpunkt alle Reaktionen des jeweiligen Zeitpunkts als Prädiktoren ein sowie als Kriterium die Differenz der Einstellungen (t0-t3). Es fanden sich weder Prädiktoren unter den Reaktionen des Feedbackbogens (t1) noch der ersten Nachbefragung (t2). Analog zur ursprünglichen Analyse (Kapitel 5.2) wurde lediglich der Subtest *Umsetzung persönlicher Ziele* t3 als Prädiktor in die Schätzgleichung einbezogen (die Detail-Ergebnisse hierzu finden sich im Anhang B, S. 78).

Bei der Überprüfung, welche nicht-ausgewählten Reaktionen möglicherweise zur Vorhersage der **Verhaltensebene** herangezogen werden können, identifizieren die drei Folgeanalysen (pro Messzeitpunkt mit allen Reaktionen des jeweiligen Zeitpunkts als Prädiktoren sowie die subjektive Verhaltenseinschätzung als Kriterium) erstmals auch Teilnehmerreaktionen des Feedbackbogens (t1) als Prädiktoren. Die Subtests *Anwendbarkeit*, *differenzierte Trainereinschätzung* sowie *Durchführung des Seminars* tragen ca. 47% zur Aufklärung der Gesamtvarianz der subjektiven Verhaltenseinschätzung bei. Aus den Reaktionen der ersten Nachbefragung (t2) findet sich lediglich die *Anwendung zu t2* als Prädiktor (46% Varianzaufklärung), wohingegen bei der zweiten Nachbefragung (t3) alle drei Teilnehmerreaktionen mit 56% Vorhersagekraft das Ergebnis der subjektiven Verhaltenseinschätzung voraussagen (alle Ergebnisse dieser zusätzlichen Auswertung finden sich in Anhang B, S. 79).

5.4.3 Höhe der Zufriedenheit im unmittelbaren Feedbackbogen und in der ersten Nachbefragung

Ein von Konradt et al. (2002) und Nork (1991) angesprochener Aspekt sind die zu positiven Beurteilungen bei Erhebung der Zufriedenheit unmittelbar am Ende der Maßnahme. Daher wurden explorativ die Häufigkeiten aufgelistet, inwiefern Teilnehmer mit einer sehr positiven unmittelbaren Beurteilung des Trainings dieses zwei bis drei Wochen später weiterhin positiv oder weniger positiv beurteilen.

Von den 47 Seminarteilnehmern hatten 34 Teilnehmer neben dem Seminar-Feedbackbogen auch die erste Nachbefragung bearbeitet. Die Mediandichotomisierung der unmittelbaren Seminarbewertung (t1) ergab 17 Teilnehmer mit einer ‚guten‘ (2⁻, 2, 2⁺, 1⁻) und weitere 17 Teilnehmer mit einer ‚sehr guten‘ Bewertung (1, 1⁺). Die Dichotomisierung für die rückblickende Seminarbewertung (t2) ergab insgesamt 16 Teilnehmer mit ‚schlechter bis mittlerer‘ Bewertung (<4.5) und 18 Teilnehmer mit einer ‚guten bis sehr guten‘ Bewertung (>=4.5). Es ergab sich folgende Häufigkeitstabelle:

Tabelle 5-10. Seminarbewertung der Teilnehmer im unmittelbaren Feedback (t1) sowie in der ersten Nachbefragung (t2)

		Rückblickende Seminarbewertung t2		Gesamt
		schlecht bis mittel < 4.5	gut bis sehr gut >=4.5	
Unmittelbare allgemeine Seminarbewertung t1	Gut 2-/ 2/ 2+/ 1-	10	7	17
	sehr gut 1/ 1-	6	11	17
Gesamt		16	18	34

Von den 17 Teilnehmern, die das Seminar unmittelbar als ‚gut‘ eingeschätzt hatten, blieben 10 Teilnehmer in der ersten Nachbefragung eher kritisch und bewerteten es nur noch als ‚schlecht bis mittel‘, wohingegen immerhin sieben Teilnehmer die Maßnahme besser bewerteten als zuvor. Umgekehrt fiel das Urteil von 6 der 17 Teilnehmer, die das Seminar unmittelbar als ‚sehr gut‘ eingeschätzt hatten, in der ersten Nachbefragung auf ‚schlecht bis mittel‘ ab. Die Einschätzungen der übrigen 11 Teilnehmer blieb dagegen bei ‚sehr gut‘.

Um explorativ zu prüfen, inwiefern sich die Einschätzungen der unmittelbaren Seminarbewertung und der rückblickenden Seminarbewertung unterscheiden, wurde der exakte McNemar-Test berechnet. Die Zellenbesetzung unterschied sich jedoch nicht signifikant, weshalb – auch aufgrund der kleinen Stichprobe – nicht von einer zeitlichen Stabilität ausgegangen werden kann, wie sie beispielweise Latham und Frayne (1989) berichten.

5.4.4 Zusammenhang der eingeschätzten Anwendbarkeit der Seminarinhalte und der tatsächlichen Anwendung der Inhalte

Nach den explorativen Ergebnissen der Analyse zur Zufriedenheitseinschätzung erschien es interessant, den Verlauf der anwendungsbezogenen Einschätzungen näher zu untersuchen, d.h. den Zusammenhang zwischen der unmittelbaren Anwendbarkeits-Einschätzung und der tatsächlichen Anwendung zu t2 und t3 zu prüfen.

Der antizipierte Anwendbarkeit korrelierte signifikant mit der Anwendung zu t2 ($r = .45$, $p < .01$) und mit der Anwendung zu t3 ($r = .60$, $p < .001$). Um zu überprüfen, ob sich die Einschätzungen der Anwendung zu t2 und der Anwendung zu t3 im Zeitverlauf verändern, wurde ein t-Test für abhängige Stichproben gerechnet. Es fand sich kein signifikantes Ergebnis ($t_{(30)} = -1.14$, $p = .26$), d.h. es gibt keinen Unterschied im Verlauf der beiden Einschätzungen über Zeit.

Es ist zu erwarten, dass Seminarteilnehmer, die eine hohe Anwendbarkeit der Seminarinhalte sehen, diese auch tatsächlich eher anwenden als solche mit einer geringen antizipierten Anwendbarkeit. Zur Prüfung dieses Unterschieds wurde eine Mediandichotomisierung für den Subtest *Anwendbarkeit t1* vorgenommen, woraus sich $n = 15$ Teilnehmer mit ‚*niedriger antizipierten Anwendbarkeit*‘ der Inhalte sowie $n = 16$ Teilnehmer mit ‚*hoher antizipierten Anwendbarkeit*‘ ergaben. Es wurde eine einfaktorielle ANOVA im Messwiederholungsdesign durchgeführt (zweistufiger Between-Faktor Gruppe: antizipierte Anwendbarkeit ‚*hoch*‘ vs. ‚*niedrig*‘, zweistufiger Within-Faktor Zeit: ‚*Anwendung nach 2 Wochen*‘ vs. ‚*nach 3 Monaten*‘)³⁵, die einen signifikanten Haupteffekt für den Gruppen-Faktor ergab ($F_{(1,29)} = 9.41$, $p < .01$).

³⁵ Die Subtests *Anwendung t2* und *t3* wurden mit denselben Items und derselben 7-stufigen Skala gemessen.

Weder der Zeit-Faktor noch die Wechselwirkung Gruppe x Zeit wurden signifikant. Während sich die Angaben über die tatsächliche Anwendung der Inhalte im betrachteten Zeitrahmen nicht verändern, hängt das Ausmaß der tatsächlichen Anwendung von der antizipierten Anwendbarkeit der Inhalte ab (vgl. Abbildung 5-4). Wie dieser Gruppenunterschied ausfällt, wurde in nachgeschalteten t-Tests für unabhängige Stichproben spezifiziert. Es ergab sich für beide Zeitpunkte ein signifikanter Unterschied (t_2 ($t_{(29)} = -2.91$, $p < .01$; t_3 ($t_{(29)} = -2.18$, $p < .05$), der so gerichtet war, dass die Teilnehmergruppe mit einer niedrigen antizipierten Anwendbarkeit angab, die Inhalte weniger angewendet zu haben als die Teilnehmergruppe mit der hohen antizipierten Anwendbarkeit (Befragung nach zwei Wochen: $M_{\text{delta}} = -.62$, $SD_{\text{delta}} = .28$; Befragung nach 3 Monaten: $M_{\text{delta}} = -1.00$, $SD_{\text{delta}} = .34$).

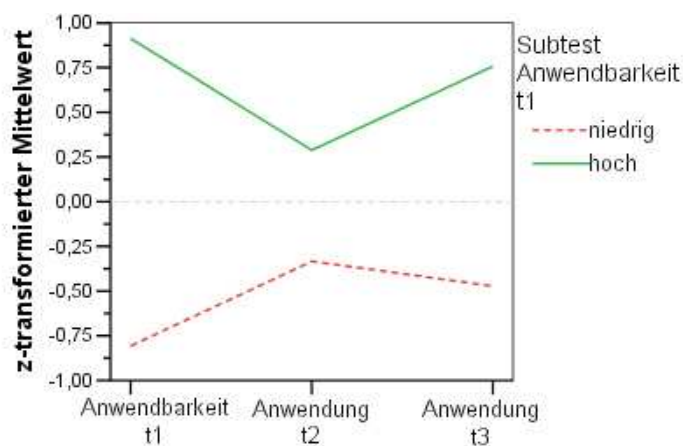


Abbildung 5-4. Verlauf der Teilnehmereinschätzungen (z-Werte): Anwendbarkeit t1, Anwendung zu t2 und t3

6 Diskussion

Dieses Kapitel enthält die Zusammenfassung und Integration der in Kapitel 5 dargestellten Ergebnisse. Bei der Interpretation der Befunde sollen mögliche Implikationen für die Praxis herausgearbeitet werden und auf methodische Limitationen eingegangen werden.

6.1 Zusammenfassung der Ergebnisse

In der vorliegenden Arbeit wurde das in der Evaluationspraxis von PE-Maßnahmen vorherrschende Vier-Ebenen-Modell von Kirkpatrick (1996) näher betrachtet. Der Fokus lag dabei im Besonderen auf den Zusammenhängen der Teilnehmerreaktionen mit den anderen Ebenen und in der Ermittlung ihrer Rolle bei der Vorhersage dieser Ebenen.

6.1.1 Zusammenhänge zwischen Reaktionen, Lernen und Verhalten

Allgemein findet sich bei Durchsicht der Literatur ein inkonsistentes Bild im Hinblick auf die Zusammenhänge, die zwischen den einzelnen Kirkpatrick-Ebenen bestehen. In der Metaanalyse von Alliger und Janak (1989) fanden sich zwischen Reaktionen und den weiteren Ebenen geringe Korrelationen, weshalb die Autoren bspw. eine Modifikation des Vier-Ebenen-Modells vorschlugen, welches die Reaktionen unverbunden zu den anderen Ebenen darstellt. Morgan und Casper (2000) erklärten diese niedrigen oder ausbleibenden Korrelationen durch eine fehlende Ausdifferenzierung der Teilnehmerreaktionen, und auch in der Studie von Warr und Bunce (1995) ergaben sich durch die Aufteilung der Reaktionen in Zufriedenheitseinschätzungen sowie nutzenbezogenen Einschätzungen höhere Korrelationen mit den anderen Ebenen.

In der vorliegenden Untersuchung wurden die Teilnehmerreaktionen in verschiedene Facetten (*affective* und *utility reactions*) differenziert und mittels unterschiedlicher Fragebogen an drei Zeitpunkten erhoben. Die Betrachtung der Zusammenhänge zwischen den einzelnen hier gemessenen Ebenen (*Reaktionen*, *Lernen* und *Verhalten*) ergab ein sehr unterschiedliches Bild. Der *Erfahrungsaustausch t2* korrelierte als einzige von 14 Reaktionsfacetten signifikant, aber negativ, mit dem Wissenstest. Dagegen korrelierten 4 der 14 Reaktionsfacetten positiv mit den verhaltensbezogenen Einstellungen (*Vorbereitung*

der Teilnehmer *t1*, Anwendung zu *t3*, Nutzen zu *t3* sowie Umsetzung persönlicher Ziele *t3*). Mit 9 von 14 Facetten korrelierte schließlich die Mehrzahl der Reaktionen mit der subjektiven Verhaltenseinschätzung, wobei sich die Korrelationen vorwiegend mit den Reaktionsfacetten der beiden Nachbefragungen (*t2* und *t3*) ergaben. Diese Befunde zeigen die Bedeutung von Teilnehmerreaktionen, die – entgegen der Ansicht von Alliger und Janak (1989) – durchaus mit den anderen Ebenen korrelieren, allerdings je nach Ebene in unterschiedlichem Ausmaß. Die hier gefundenen Zusammenhänge mit der Einstellungs- und der Verhaltensebene sprechen in diesem Fall gegen eine Loslösung der Reaktionen von den anderen Ebenen und befürworten gleichzeitig eine ausdifferenzierte Messung (durch mehrere Items bzw. Subtests) sowie eine Messung der Reaktionen zu mehreren Zeitpunkten.

Für eine Trennung der Reaktionen in *affective* und *utility reactions* und den Stellenwert von *utility reactions* spricht die Höhe der Korrelationen dieser beiden Zufriedenheitsfacetten. Auffallend sind die hohen (Einzel- und Durchschnitts-) Korrelationen der *utility reactions* speziell mit der subjektiven Verhaltenseinschätzung. Betrachtet man die **einzelnen Korrelationen**³⁶, so blieben die Zusammenhänge der *affective reactions* mit den gemessenen Ebenen einerseits entweder aus oder waren negativ. Es fanden sich jedoch vor allem zur subjektiven Verhaltenseinschätzung einige signifikant positive Korrelationen, die z.T. nach Cohen (1988) einem mittleren Effekt entsprechen ($r = .30$). Andererseits korrelierten die *utility reactions* vorwiegend signifikant positiv (außer im Wissenstest) und entsprechen kleinen ($r = .10$), mittleren oder sogar großen Effekten ($r = .50$, vgl. Cohen, 1988). Die Bildung **mittlerer Korrelationen** (anhand Fishers *Z*-Werten) ergab allgemein höhere Korrelationen der *utility reactions* und den Ebenen als die *affective reactions*.

Fehlende Zusammenhänge zwischen Lern- und Verhaltensebene, wie in dieser Untersuchung vorgefunden, werden in der Literatur ebenfalls berichtet (z.B. Tracey et al., 1995). Tracey et al. (1995) erklären dies damit, dass es, trotz des inhaltlichen Bezugs beider Maße auf die Trainingsinhalte, keine Überlappung der Items gibt. Bei entsprechend verschiedenen Messinhalten verwundere es nicht, wenn Zusammenhänge ausblieben, während umgekehrt eine inhaltliche Abstimmung der Messungen hohe Korrelationen wahr-

³⁶ Vgl. zur Übersicht Tabelle 5-8 (S. 92): Dort stehen in Klammern die Korrelationen, daneben fett gedruckt die Fisher *Z*-Werte.

scheinlicher machen müssten. In der vorliegenden Untersuchung besteht für den Wissenstest die Problematik der Messungenauigkeit, wodurch sich die ausbleibenden Zusammenhänge erklären lassen. Diese methodische Schwäche wird in Kapitel 6.4 noch ausführlicher dargestellt.

6.1.2 Die Rolle der Teilnehmerreaktionen zur Vorhersage des Seminarerfolgs

Neben der Frage nach den Zusammenhängen zwischen den Ebenen wurde die Vorhersagekraft der Teilnehmerreaktionen für die Erfolgskriterien der weiteren Ebenen geprüft. Laut Kirkpatrick (1996) wird der Erfolg von PE-Maßnahmen durch die Evaluationskriterien Reaktionen, Lernen, Verhalten und nicht zuletzt durch die Ergebnisse abgebildet. Mit der Erweiterung des Vier-Ebenen-Modells um die Ebene des ROI (Return on investment) durch Phillips (2005) steigt unter Managern und Weiterbildungsverantwortlichen das Interesse an der Ergebnis- und der ROI-Ebene (Phillips, 2005), da diese leichter mit ökonomischen Aspekten in Verbindung zu bringen sind. Wie Abbildung 6-1 zeigt, findet sich hier allerdings eine Diskrepanz zwischen denjenigen Informationen, die den Managern interessanter erscheinen (Ebenen 4 und 5) und denen, die sie in der Regel nach Ende einer Maßnahme zurückgemeldet bekommen (Ebenen 1 bis 3). Dies entspricht der bereits erörterten Diskrepanz in der Anwendungshäufigkeit der einzelnen Kirkpatrick-Ebenen (Borchert & Rutschke, 2005).

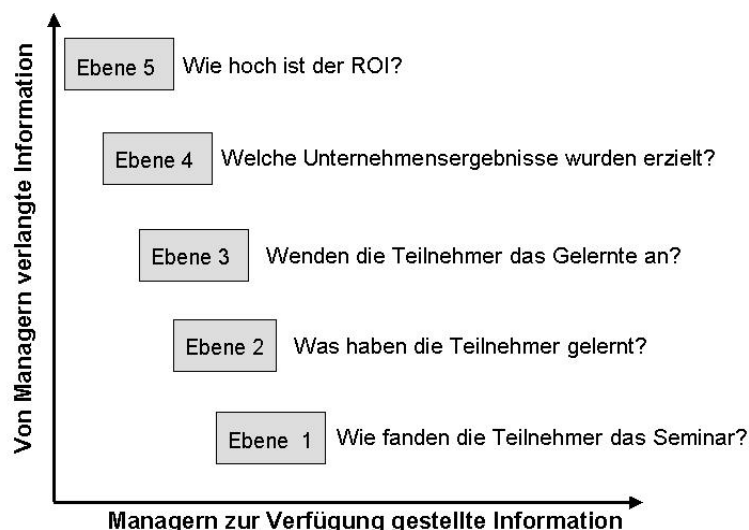


Abbildung 6-1. Diskrepanz zwischen der Information, die sich Manager von einer Evaluation erhoffen und der Information, die ihnen zur Verfügung gestellt wird (nach Phillips, 2005, S. 2).

Angesichts der vorherrschenden Evaluierung auf Reaktionsebene war ein zweites Anliegen dieser Arbeit, die Vorhersagekraft der Teilnehmerreaktionen im Hinblick auf die Erfolgskriterien der höheren Ebenen zu identifizieren. Als Prädiktoren dienten dabei die nutzenbezogenen Einschätzungen zu *Anwendbarkeit t1* bzw. *Anwendung t2* und *t3*, zum *Nutzen* zu *t2* und *t3*, zur *Umsetzung persönlicher Ziele t3* und die affektiven Einschätzungen zur *Durchführung des Seminars t1* gehandelt.

Dabei wurde aus den genannten Teilnehmerreaktionen kein einziger Prädiktor für das Ergebnis im Wissenstest in die Schätzgleichung aufgenommen, und auch in Bezug auf die verhaltensbezogenen Einstellungen stellte sich mit dem Subtest *Umsetzung persönlicher Ziele* lediglich ein einziger Prädiktor heraus (12% Varianzaufklärung). Im Gegensatz dazu leisten die *Anwendung zu t2* und *zu t3* sowie die *Umsetzung persönlicher Ziele t3* einen großen Beitrag zur Vorhersage der subjektiven Verhaltenseinschätzung (75% Varianzaufklärung).

Im Hinblick auf die in der Evaluationspraxis von PE-Maßnahmen vorherrschende Messung auf Reaktionsebene machen die vorliegenden Ergebnisse Mut. Die Annahme einer prädiktiven Funktion von Teilnehmerreaktionen (und hier vorwiegend der *utility reactions*) von Morgan und Casper (2000) ließ sich bestätigen. Darüber hinaus wurde der Nutzen mehrmaliger Messungen deutlich: Obwohl hier selbst eine einzelne Reaktionsmessung (z.B. mittels Feedbackbogen) die tatsächliche Anwendung der Inhalte nach mehreren Monaten fast zur Hälfte vorhersagen konnte, verbesserte sich die Vorhersage der Erfolgskriterien mit weiteren Reaktionsmessungen nach zwei Wochen und nach drei Monaten, da diese nicht mehr nur die antizipierte Anwendbarkeit erfassen, sondern bereits die tatsächliche Anwendung.

6.1.3 Faktoren mit Einfluss auf Reaktionen, Lernen und Verhalten

Neben den Ebenen des Kirkpatrick-Modells (1996) wurden aus dem Trainingseffektivitäts-Modell von Tannenbaum (Cannon-Bowers et al, 1995) zusätzlich Effektivitätsfaktoren erhoben und ihr Einfluss auf die Evaluationsebenen untersucht. Denn obwohl die Teilnehmerreaktionen ein Bestandteil des Modells zur Trainingseffektivität von Tannenbaum (Cannon-Bowers et al., 1995; Höft, 2001) sind, stehen sie in keiner

Verbindung mit den anderen Ebenen oder den Einflussvariablen³⁷. Von ihnen wird keinerlei Einfluss angenommen, es bestehen auch keine Verbindungen zwischen Training und Reaktionen, obwohl sich die Reaktionen doch erst aus dem Training ergeben. Seitens der organisationalen bzw. situationalen Merkmale wird ebenfalls kein Einfluss angenommen.

Im Gegensatz dazu deuten die vorliegenden Ergebnisse auf eine solche Beeinflussung hin (vgl. Tabelle 6-1). Die erhobenen Einflussgrößen sollten Aufschluss darüber geben, inwiefern sich Teilnehmer in ihren Reaktionen und Ergebnissen unterscheiden, und zwar in Abhängigkeit des Ausmaßes der individuellen und organisationalen Variablen. Während die einzelnen einfaktoriellen, multivariaten Varianzanalysen (MANOVAs) (mit den Einflussgrößen als Faktor) einen Einfluss vom *subjektiven Bedarf*, *Transferklima* und *spezifischer Seminarbewertung* offenbarten, fand sich für keine der weiteren Einflussvariablen (*Anwendungsmöglichkeit*, *Vorerfahrung*, *Motivation*, *Selbstwirksamkeit* sowie *allgemeine Seminarbewertung*) ein signifikanter Haupteffekt. Mittels anschließender einfaktorieller Varianzanalysen (ANOVAs) wurden hier jedoch Einflüsse auf univariater Ebene aufgedeckt (Tabelle 6-1).

Tabelle 6-1. Überblick über die Wirkung der Einflussgrößen auf die zehn Zielkriterien

	Ausgangslage		Subj. Bedarf	Anwendungs- möglichkeit	Expertise	Transfer- klima	Bewertung	
	Motivation	SW ^a					allg.	spez.
1 Durchführung (t1)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
2 Anwendbarkeit (t1)	**	(-)	***	(-)	(-)	(-)	(-)	***
3 Anwendung (t2)	(-)	(-)	(-)	(-)	(-)	**	(-)	*
4 Nutzen (t2)	(-)	(-)	(-)	†	(-)	***	(-)	†
5 Anwendung (t3)	†	(-)	(-)	(-)	(-)	**	(-)	**
6 Nutzen (t3)	(-)	(-)	(-)	†	(-)	**	(-)	*
7 Pers. Ziele (t3)	(-)	(-)	(-)	(-)	†	*	(-)	†
8 Wissen	(-)	(-)	(-)	(-)	*	(-)	(-)	(-)
9 Einstellungen	(-)	*	(-)	(-)	†	(-)	(-)	(-)
10 Verhalten	(-)	(-)	(-)	(-)	(-)	**	*	†

Anmerkungen. In fetter Schrift sind in der Kopfzeile der Tabelle diejenigen Einflussfaktoren markiert, bei denen sich ein signifikanter Haupteffekt zeigte. In den Zellen ist eingetragen, ob univariat ein Unterschied zwischen den Faktoren besteht.

^a SW = Selbstwirksamkeit.

(-) kein Einfluss, † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$.

³⁷ Für das gesamte Modell zur Trainingseffektivität siehe Abbildung 2-3 auf S. 24.

Tabelle 6-1 zeigt den signifikanten Unterschied der Einflussgröße *subjektiver Bedarf*. Trainees, die für sich einen hohen Bedarf sahen, schätzten die *Anwendbarkeit t1* im unmittelbaren Feedbackbogen deutlich besser ein als Kollegen mit geringerem subjektivem Bedarf. Dieses Ergebnis spricht für die Durchführung einer Bedarfsanalyse, deren Bedeutung Alvarez et al. (2004) oder Salas und Cannon-Bowers (2001) hervorgehoben haben. Außerdem impliziert dies neben einer zielgerichteten Auswahl der Trainees im Vorfeld eine bessere Information über den Sinn der Maßnahme. Diese Aufklärung der Trainees ist notwendig, um ein Bewusstsein zu schaffen, welche Bedeutung eine Teilnahme an der Maßnahme hat und welche Unternehmensziele damit verfolgt werden. Denn ist den Trainees die Unternehmensstrategie hinsichtlich dieser Maßnahme nicht bewusst bzw. ist sie nicht ausreichend bewusst gemacht worden, schreibt das Unternehmen den Teilnehmern zwar Bedarf zu, die Teilnehmer selbst sehen diesen Bedarf jedoch (noch) nicht. Dadurch könnte die Maßnahme aus Sicht der Teilnehmer sprichwörtlich an ihrem subjektiven Bedarf „vorbeigegangen sein“, wohingegen aus Sicht des Unternehmens durchaus Bedarf besteht. Werden PE-Maßnahmen auf eine Linie mit den durch eine Bedarfsanalyse ermittelten Unternehmenszielen gebracht, ist eine Steigerung der Unternehmenseffektivität zu erwarten (Alvarez et al., 2004).

Ein Einfluss des *Transferklimas* auf die Einschätzungen der Teilnehmer wird aus Tabelle 6-1 ebenfalls deutlich. Neben den Reaktionen unterscheidet sich auch die subjektive Verhaltenseinschätzung bei entsprechender Unterstützung durch Vorgesetzte, Kollegen sowie förderlicher situationaler Gegebenheiten. Dies spricht für die Bedeutung des Transferklimas auf den Praxistransfer, dessen Bedeutung z.B. Rouiller und Goldstein (1993) ebenso wie Tracey et al. (1995) betonen. Vor allem ein positives Transferklima führte zu besseren Teilnehmereinschätzungen in Bezug auf die *Anwendung* und den *Nutzen* nach zwei Wochen bzw. drei Monaten sowie auf die *Umsetzung persönlicher Ziele*. Diese Teilnehmer gaben im Gegensatz zu Teilnehmern mit negativem Transferklima ferner an, die Seminarthemen tatsächlich häufiger anzuwenden. Wenn daher an Transfersicherungsmaßnahmen gedacht wird, erscheint es in diesem Zusammenhang sinnvoll, noch vor Beginn einer Maßnahme auf das Umfeld zu schauen, in welchem diese Maßnahme später seine Wirkung entfalten soll. Eine optimale Wirkung kann nur erzielt werden, wenn – neben erfolgtem Lernzuwachs und Veränderungen im Verhalten – auch die

Randbedingungen stimmen: die Unterstützung und Förderung am Arbeitsplatz und im Arbeitsumfeld.

Erfolgt eine Differenzierung in eine allgemeine Seminarbewertung (d.h. die Zufriedenheit mit dem gesamten Training) und in eine spezifische Bewertung (im Sinne der Anwendbarkeitsbeurteilung), stellt sich folgendes heraus: Die ***seminarspezifische Bewertung***, also die Höhe der Anwendbarkeitsbeurteilung des unmittelbaren Feedbacks, hat nicht nur einen Einfluss auf die nachfolgenden Einschätzungen der *Anwendung zu t2* und *t3*, sondern auch auf die subjektive Verhaltenseinschätzung. Demnach konnten Trainees mit einer hohen antizipierten Anwendbarkeit der Inhalte diese nach zwei Wochen sowie nach drei Monaten tatsächlich häufiger anwenden als diejenigen, die eine schlechtere Anwendbarkeit voraussahen. Zusammen mit den Ergebnissen der korrelativen und der Regressionsanalyse hebt dieses Ergebnis die Bedeutung der *utility reactions* noch einmal hervor. Durch diese Art der Einschätzungen könnten also bereits durch einen entsprechend konstruierten Seminar-Feedbackbogen Hinweise auf die spätere Anwendung gegeben werden. Bei einer einmaligen Messung der Reaktionsebene ist daher auf eine Ausdifferenzierung der Reaktionen zu achten, um Informationen zu erhalten, die über das Seminar hinaus gehen. Eine Messung auf weiteren Ebenen ist jedoch weitaus empfehlenswerter, und bei entsprechenden Ressourcen sollte auch die mehrfache Messung der Reaktionsebene in Betracht gezogen werden.

Alle weiteren untersuchten Einflussgrößen ergaben keine signifikanten Haupteffekte in den Einschätzungen (siehe Tabelle 6-1), allerdings zeigen sich für diese Faktoren in den anschließenden univariaten Varianzanalysen gewisse Einflüsse auf die Zielkriterien.

Obwohl für die ***allgemeine Seminarbewertung*** ein signifikanter Haupteffekt ausblieb, zeigte sich in der anschließenden univariaten Varianzanalyse in Abhängigkeit von der Höhe dieser unmittelbaren Zufriedenheitseinschätzung mit dem Seminar (*affective reaction*) ein unterschiedliches Ausmaß darin, wie häufig die einzelnen Themen angewendet wurden (subjektive Verhaltenseinschätzung).

Während Mathieu et al. (1993) einen Einfluss von ***Selbstwirksamkeit*** sowohl auf die Teilnehmerreaktionen als auch auf das Verhalten fanden, lassen sich die Ergebnisse der vorliegenden Untersuchung nicht in diesen Befund einreihen. Es zeigte sich für die Selbstwirksamkeit lediglich ein univariater Einfluss auf die verhaltensbezogenen Einstellungs-

änderungen: Teilnehmer mit einer höheren Selbstwirksamkeit *vor* dem Training verbesserten sich stärker in ihren Einstellungen zum Vertriebsverhalten als Teilnehmer mit geringen Werten.

In Bezug auf die *Motivation* der Teilnehmer zeigt sich ein Einfluss auf die Einschätzung der antizipierten *Anwendbarkeit t1*. Gehen Teilnehmer mit einer höheren Motivation ins Seminar, sehen sie nicht nur unmittelbar eine höhere Wahrscheinlichkeit, die Seminarinhalte anwenden zu können, sondern scheinen auch nach drei Monaten noch motivierter, diese umzusetzen. Dies zeigt sich in der höheren Anwendungseinschätzung der Themen zum Zeitpunkt der 2. Nachbefragung.

Mit der Häufigkeitsangabe des *Kundenkontakts* wurde von Seiten des Unternehmens ein objektives Maß für die *Anwendungsmöglichkeit* geliefert. Dieser Faktor wurde lediglich auf univariater Ebene jeweils beim *Nutzen zu t2* und *t3* marginal signifikant. Erstaunlicherweise blieben Unterschiede in der *Anwendbarkeit zu t1* sowie in der späteren *Anwendung zu t2* und *zu t3* ebenso aus wie Unterschiede in der subjektiven Verhaltenseinschätzung. Es wäre durchaus denkbar gewesen, dass sich die Möglichkeit zur Anwendung, die sich im Ausmaß des Kundenkontakts widerspiegelt, auch in der Häufigkeit der Anwendung der Themen (= subjektive Verhaltenseinschätzung) zeigt.

Allerdings besteht hier die Schwierigkeit, dass Kundenkontakt nicht gleich Kundenkontakt ist. Wie die Teilnehmer im Seminar bzw. teilweise in ihren Antworten auf die offenen Fragen berichteten (siehe Anhang C, S. 94ff) können bei langjährigen Kunden oder sehr spezifischen Anfragen manche Techniken (um etwa Gleichgültigkeit zu überwinden) nicht greifen und somit nicht angewendet werden. Teilnehmer sind dann in der schwierigen Situation, zwar viel Kundenkontakt zu haben, innerhalb dieser Kundenbeziehungen aber die erlernten Themen nicht anwenden zu können. Wie bereits beim Faktor *subjektiver Bedarf* andiskutiert, deutet dies darauf hin, dass nicht nur der Bedarf ermittelt werden muss, sondern dass das subjektive Bedarfsempfinden der Teilnehmer mit dem Bedarf aus Sicht des Unternehmens im Einklang gebracht werden muss. Analog dazu ist der hier gemessene Faktor *Anwendungsmöglichkeit* zu sehen – es reicht nicht aus, von Unternehmensseite die Möglichkeit zur Anwendung zu sehen, vielmehr sollte feiner abgestimmt werden, ob dies tatsächlich in der Arbeitspraxis aus Sicht des einzelnen Mitarbeiters der Fall ist. Ist dem nicht so, kann unter Umständen eine bessere Information über die Unternehmensziele Abhilfe schaffen – oder aber die geplante Maßnahme ist tatsächlich nicht für den

betreffenden Mitarbeiter geeignet. Weitere zu berücksichtigende Faktoren sind, im Hinblick auf die Anwendungsmöglichkeit, die Unterstützung durch das soziale Arbeitsumfeld und förderliche situationale Bedingungen im Arbeitskontext, also das *Transferklima*. Wie bereits angesprochen können sich hier Hindernisse ergeben: Das Seminar mag zwar als gut eingeschätzt werden und möglicherweise erfolgt ein Lernzuwachs oder eine Einstellungsveränderung, eine tatsächliche Anwendung bleibt jedoch aufgrund ungünstiger Arbeitsbedingungen aus – z.B. aufgrund eines Umfeldes, in dem es nicht gewünscht ist, neues Verhalten oder neue Strategien beim Kunden anzuwenden³⁸.

Obwohl für den Faktor *Vorerfahrung* bedeutsame Unterschiede erwartet wurden, blieb ein signifikanter Haupteffekt aus. Univariat zeigte sich hier als einzige aller erhobenen abhängigen Variablen ein signifikanter Unterschied im Ergebnis des Wissenstests. Die Signifikanz an sich überrascht insofern nicht, als dass Teilnehmer, die mehr Vertriebserfahrung haben, entsprechend tiefer in die gesamte Thematik eingearbeitet sind und insofern über mehr Wissen verfügen bzw. dieses leichter abrufen können. Was allerdings nicht zufriedenstellend zu beantworten ist, ist die Frage, ob sich die Teilnehmer mit Vorerfahrung von denen ohne Vorerfahrung im Ausmaß des (durch das Training) erworbenen Wissens unterscheiden. Um diese Frage zu klären ist eine Prä-Post-Messung des Wissenstests notwendig, die in dieser Untersuchung aufgrund des hohen Zeitaufwandes nicht realisiert werden konnte.

Für alle hier untersuchten Faktoren wäre im Zuge zukünftiger Forschung zu prüfen, ob sich die gefundenen bzw. ausbleibenden Zusammenhänge verifizieren lassen oder sich durch andere Operationalisierungen andere Ergebnisse finden. Ebenso ist eine Vergrößerung der Stichprobe wünschenswert, um mögliche Wechselwirkungen zwischen den einzelnen Einflussgrößen aufzudecken.

Zusammenfassend erscheint es nicht ausreichend, nur Messungen auf Reaktions-, Lern-, Verhaltens- und Ergebnisebene durchzuführen: Die im Kontext der dritten Fragestellung gefundenen Ergebnisse weisen auf den Nutzen hin, neben diesen Evaluationskriterien mögliche Einflussgrößen zu berücksichtigen: Beispielsweise wird durch das beschriebene Zusammenspiel von Transferklima und Anwendungsmöglichkeit die enge Verknüpfung

³⁸ Eine Überprüfung der Interaktion zwischen Anwendungsmöglichkeit und Transferklima wurde aufgrund der Komplexität der Daten und der Stichprobengröße nicht mehr vorgenommen.

organisationaler und situationaler Einflussfaktoren deutlich, was die Bedeutung einer Erfassung dieser Faktoren unterstreicht. Nur wenn eine Trainingsevaluation mit Aspekten der Trainingseffektivität ergänzt wird (vgl. Alvarez et al., 2004), geht der Evaluationsprozess über die Evaluation einer einzelnen Weiterbildung hinaus und mündet für alle an dieser Maßnahme beteiligten Personen in einem Optimierungsprozess. Schwächen und Hindernisse im Transfer (z.B. auf Unternehmensseite) können aufgedeckt und es kann gezielt entgegengewirkt werden.

6.2 Verlauf der Einschätzungen zur Anwendung der Seminarinhalte (Transferverlauf)

Im Rahmen der zusätzlichen Auswertung fand sich ein interessanter Verlauf der Anwendung der Seminarinhalte in Abhängigkeit von der antizipierten Anwendbarkeit (vgl. Abbildung 5-4, S. 93). Es zeigte sich einerseits, dass Teilnehmer, die bereits am Seminarende davon ausgingen, die Inhalte zukünftig anwenden zu können, entsprechend über eine höhere Anwendungshäufigkeit nach zwei Wochen bzw. drei Monaten berichten als Teilnehmer, die eine solche Anwendbarkeit nicht annahmen. Andererseits zeigt sich gerade bei diesen Teilnehmern mit hoher antizipierter Anwendbarkeit ein signifikanter Abfall der beiden nachfolgenden Einschätzungen im Vergleich zur ursprünglichen Einschätzung. Dies bedeutet, dass sie die Inhalte zwar tatsächlich besser anwenden können, allerdings nicht in dem Ausmaß, wie unmittelbar nach dem Seminar gedacht.

Neben anderen Kurven beschreiben Baldwin und Ford (1988) in ihrer Arbeit eine Kurve, die einen ähnlichen Verlauf aufweist wie den der Gruppe der ‚*hohen antizipierten Anwendbarkeit*‘ (Abbildung 5-4). Die Autoren identifizieren verschiedene Verläufe von „transfer maintenance curves“. Ursprünglich finden sich solche Verläufe in der Lernforschung, die den zeitlichen Verlauf sowie den Umfang des Lernens beschreiben. In Anlehnung daran beschreiben die „transfer maintenance curves“ den Verlauf der Anwendung und Aufrechterhaltung neu erlernter Inhalte nach einer Trainingsmaßnahme (Baldwin & Ford, 1988). In der Trainingspraxis wird dieser Verlauf häufig beobachtet und ist in der Literatur als „Transferlücke“ bekannt (Buchester, 2003): Ein durch das Seminar angestiegenes Wissens- oder Leistungsniveau fällt mit der Rückkehr in die gewohnte Arbeitsumgebung wieder ab (evtl. auf Vor-Trainings-Niveau). Neben schlichtem Vergessen der Inhalte kann dieser

Umstand in intraorganisationalen Widerständen begründet sein (Buchester, 2003). Eine weitere Ursache könnte daran liegen, dass Trainees durch ihre Teilnahme an einer Maßnahme ihrem Tagesgeschäft nicht nachgehen können. Die aufgestaute Arbeit muss erst erledigt werden, bevor eine erneute bzw. zielgerichtete Beschäftigung mit dem neu Erlernten erfolgt.

Diesem Effekt ist nach Rank und Thiemann (1998) durch Nachfolgemeasures vorzubeugen – sog. Follow-up-Veranstaltungen, wie sie im Modell von Tannenbaum (Cannon-Bowers et al., 1995) als *post training interventions* zu finden sind. Laut Neuberger (1994) kann „allein die Ankündigung solcher Treffen ... die Bereitschaft steigern, sich um die Verwirklichung des Gelernten zu bemühen“ (S.189). Dies ist konsistent mit den Befunden von Baldwin und Magjuka (1991), die bei MBA-Studenten eine höhere Transfermotivation nachweisen konnten, wenn diese das Gefühl hatten, die Maßnahme würde wahrhaftig unterstützt, d.h. wenn ihre Manager nicht nur grünes Licht gaben (*permission*), sondern sich darüber hinaus zeitlich verpflichteten, das Programm zu begleiten bzw. Follow-up-Maßnahmen mit den Trainees zu gestalten (*more meaningful support*).

Trotz der Bedeutung von Follow-up-Maßnahmen werden diese bislang nur selten durchgeführt (Rank & Thiemann, 1998). Dies könnte wieder an den damit verbundenen Kosten und dem zusätzlichen personellen, zeitlichen und organisatorischem Aufwand liegen. Wenn bereits aus diesen Gründen in der Praxis vorwiegend auf der Ebene 1 evaluiert wird, liegt der Vorschlag von Clement (1982) als einfachste Lösung auf der Hand: die mehrmalige Messung der Reaktionen.

6.3 Bedeutung von Mehrfachmessungen

Die Messung der Teilnehmerreaktionen zu mehreren Zeitpunkten ist, wie die Ergebnisse der vorliegenden Untersuchung zeigen, aus verschiedenen Gründen sinnvoll.

Die hier eingesetzten Instrumente zur Erfassung der Teilnehmerreaktionen haben den Vorteil, eine mehrmalige Messung der Reaktionsebene zu ermöglichen, wie sie z.B. von Nork (1991) empfohlen wird. Entsprechend dem Vorschlag verschiedener Autoren (Clement, 1982; Rank & Thiemann, 1998; Cannon-Bowers et al., 1995), unterschiedliche Reaktionen zu erheben statt eine schlichten Wiederholungsmessung durchzuführen, wurden auch hier unterschiedliche Zufriedenheitsaspekte abgefragt. Somit kann den Reaktions-

messungen beispielsweise entnommen werden, ob die Teilnehmer unmittelbar nach dem Seminar *glauben*, die Inhalte anwenden zu können. Weitere Messungen offenbaren nach wenigen Wochen bzw. Monaten, wie die Teilnehmer aus einer gewissen Distanz zum Seminar und mehr Nähe zum Arbeitsplatz die *tatsächliche* Anwendung der Inhalte sehen. Genau dieser Abstand erlaubt einen kritischen Blick auf förderliche und hinderliche Faktoren bei der aktuellen Anwendung. Diese Informationen sind für den Veränderungsprozess innerhalb der Evaluation essentiell, da sie, bei Rückmeldung der Ergebnisse an Teilnehmer, Vorgesetzte und Trainingsbeauftragte eine gezielte Stärkung der positiven Aspekte bzw. eine Beseitigung von Transferbarrieren ermöglichen.

Die Mehrfachmessung erlaubt allerdings nicht nur die Identifikation solcher Problematiken, sie stellt in gewisser Weise auch eine Nachfolgemaßnahme dar, wie sie Rank und Thiemann (1998) fordern. Zwar beziehen sich Rank und Thiemann (1998) auf eine Post-Trainings-Maßnahme im Sinne einer weiteren Veranstaltung, möglicherweise könnte jedoch auch ein Fragebogen (im Sinne einer Nachmessung) eine solche Post-Trainings-Intervention darstellen. Denn durch aufgestaute Arbeit nach einem Seminar können die neu erlernten Inhalte zunächst in Vergessenheit geraten. Eine Nachbefragung mit entsprechendem Ergebnisfeedback kann als Motivationsschub dienen und anschließend – im Sinne der Aktionsforschung von Lewin (1947) – die erstmalige oder erneuten Beschäftigung mit den Seminarinhalten oder Themen anregen. In dieser Studie wurde durch die eingesetzten Instrumente versucht, dem Thema Feedback aufgrund seiner grundlegenden Bedeutung für individuelles Lernen (Jöns, 1997) Rechnung zu tragen und entsprechend Feedback über die persönlichen Ergebnisse vergeben. Ob allerdings der Einsatz der beiden Online-Nachbefragungen an sich, die Tatsache der Bearbeitung dieser Befragungen oder aber die Konfrontation mit den Ergebnissen (z.B. nach dem Wissenstest) einen Effekt hat, ist mit vorliegendem Stichprobenumfang und Studiendesign nicht zu beantworten und erfordert weitere Untersuchungen.

Von mehrfachen Reaktionsmessungen können weiterhin alle am Evaluationsprozess beteiligten Personen einen spezifischen Nutzen ziehen. Oft wird nicht deutlich, dass neben Teilnehmern (und Unternehmen) der Trainer sowie das Weiterbildungsinstitut ebenfalls Akteure innerhalb dieses Prozesses sind. Tabelle 6-2 soll darstellen, welche Zielgruppe wann von welcher Ergebnisrückmeldung am meisten profitieren kann.

Tabelle 6-2. *Bedeutung der einzelnen Messungen für die verschiedenen Personengruppen, die am Evaluationsprozess beteiligt sind (in Anlehnung an Wottawa & Thierau, 2003)*

	Unmittelbare Einschätzung	nach ca. 2 Wochen	nach ca. 3 Monaten
Teilnehmer	+	+++	+
Trainer	+++	+	+
Trainingsanbieter	++	+	+++
Unternehmen	++	+	+++

Die Messungen zu den einzelnen Zeitpunkten haben dabei neben ihrer Rolle innerhalb des Evaluationsprozesses auch eine weitere besondere Bedeutung inne. Ein **unmittelbarer Feedbackbogen** am Seminarende signalisiert den Teilnehmern allgemeines Interesse der Trainingsverantwortlichen an ihrer Meinung, sei es zur Überprüfung der Qualität oder zur Verbesserung der Maßnahme. Dieser letzte Aspekt ist auch für den Trainer interessant. Für Management oder Weiterbildungsverantwortliche ist der Feedbackbogen häufig Entscheidungsgrundlage für die Buchung weiterer Seminare. Die Relevanz für den Trainingsanbieter liegt daher ebenfalls auf der Hand. Eine **Messung nach ca. zwei Wochen** ist vermutlich vor allem für die Teilnehmer relevant – sie kann die oben beschriebenen motivationalen Auswirkungen haben oder eine Auffrischung der Themen darstellen (im Sinne eines „updates“). Für Trainingsanbieter und vor allem für Unternehmen erscheint dahingegen die **Messung nach mehreren Monaten** am wichtigsten. Hier zeichnet sich ab, inwiefern die Maßnahme Leistungssteigerungen zur Folge hat oder andere, für das Team, die Abteilung oder das Unternehmen relevante Veränderungen bringt. Außerdem sollte sich nach dieser Zeit langsam der Erfolg der Maßnahme auf härtere Kennzahlen (z.B. ROI nach Philipps, 2003) niederschlagen.

Oftmals besteht bei den Personalverantwortlichen jedoch eine gewisse Ungeduld im Hinblick auf die erfolgreiche Umsetzung der Trainingsinhalte in die Praxis. Konkrete Ergebnisse werden schneller erwartet, als sie in der Arbeitsrealität eintreten. Vor allem für Maßnahmen, in denen neue Verhaltensweisen erlernt werden, sollte klar gestellt werden, dass Ergebnisse nicht notwendigerweise sofort oder nach wenigen Wochen zu erwarten sind. So ist das Verhalten von Mitarbeitern, wenn es sich bereits über Jahre hinweg aufgebaut und gefestigt hat, nicht oder nur geringfügig durch eine punktuelle Maßnahme zu verändern (Rank & Thiemann, 1998). Dementsprechend sollte sich eine Evaluation an

diesem Zeitverlauf orientieren. Wie in der vorliegenden Untersuchung gezeigt werden konnte, kann die mehrfache und ausdifferenzierte Messung von Reaktionen bereits Anhaltspunkte zur Anwendung der Seminarinhalte nach einigen Wochen liefern, ohne dass explizit die Transferleistung gemessen werden muss. Personalverantwortliche erhalten dadurch zumindest eine Tendenz, in welche Richtung die Ergebnisse dieser Maßnahme gehen.

6.4 Methodische Limitationen

Neben den einzelnen Schwachstellen, die bereits im Verlauf der Ergebnisinterpretation diskutiert wurden, gibt es noch weitere Besonderheiten bzw. problematische Aspekte, die anzusprechen sind.

Eine der größten Limitationen dieser Untersuchung bestand im reduzierten Stichprobenumfang. Da für die Auswertung vollständige Datensätze erforderlich waren und die Teilnehmer in dieser Untersuchung zu vier Messzeitpunkten Fragebogen bearbeiten mussten, fielen zu den einzelnen Zeitpunkten mehrere Teilnehmer aus. So gingen zwar 47 Teilnehmer in die endgültige Stichprobe der Seminarteilnehmer ein, zur die Überprüfung der zweiten und dritten Fragestellung konnten jedoch nur die Datensätze von 30 Teilnehmern herangezogen werden. Für die Analyse der Vorhersagekriterien (Regressionsanalyse) und der Einflussfaktoren (MANOVA) wäre eine größere Stichprobe im Hinblick auf die Relevanz dieser Thematik wünschenswert gewesen. Die kleinen Stichprobenumfänge der berechneten MANOVAs schränken daher die Interpretierbarkeit der Ergebnisse zwar ein, sie können jedoch zumindest als Tendenz in die Richtung dieser Ergebnisse und als Anstoß für weitere Untersuchungen angesehen werden.

Obwohl für die vorliegende Untersuchung parallel zu den Teilnehmern eine vergleichbare Kontrollgruppe gewonnen werden konnte, ist kritisch anzumerken, dass auch diese Stichprobe relativ klein war. Die Kontrollgruppe sollte gerade im Hinblick auf die Ergebnisse des Wissenstests herangezogen werden, da dieser bei den Teilnehmern einmalig als Post-Messung nach dem Training erhoben wurde. Dadurch ist kein Vergleich zwischen dem Wissensstand der Teilnehmer vor dem Seminar (als Ausgangslage) und dem Wissensstand nach dem Seminar möglich, d.h. es lassen sich keine Aussagen über einen Wissenszuwachs bzw. einen Lernprozess treffen. In diesem Kontext sollte der Einsatz einer Kontrollgruppe

Klarheit schaffen, inwiefern Mitarbeiter mit dem im Training erlernten Wissen besser abschneiden als solche, die diese Themen nie gezielt gelernt haben. Nach der Basisbefragung reduzierte sich die Kontrollstichprobe um 50% auf 11 Teilnehmer, wodurch die Ergebnisse entsprechend vor dem Hintergrund dieses kleinen Stichprobenumfangs zu sehen sind. Allgemein sind die Ergebnisse des Wissenstests demnach nur unter Vorbehalt interpretierbar – ideal wäre ein Prä-Post-Kontrollgruppen-Design, um einen tatsächlichen Lernerfolg nachweisen zu können.

Ein weiterer kritischer Punkt im Hinblick auf den Wissenstest sind die teilweise unzureichenden Werte für Cronbachs Alpha. Wichtig ist jedoch anzumerken, dass hier die Anzahl richtiger Antworten interessierte – dennoch erscheint eine Überarbeitung der verwendeten Fragen notwendig, um die innere Konsistenz zu erhöhen.

Die geringe Vorhersagekraft der gefundenen Prädiktoren in den für den Wissenstest durchgeführten Regressionsanalysen ist wohl zum Teil in der Messungenauigkeit dieses Tests zu sehen, andererseits kann es auf das Vorliegen weiterer Variablen hinweisen, welche die Ergebnisse auf der Lernebene (Wissen und Einstellungen) besser vorherzusagen vermögen. Möglicherweise kommen gerade hier andere individuelle, situationale oder trainingsbezogene Aspekte (z.B. kognitive Fähigkeiten, Lernstil) zum Tragen, während die Teilnehmerreaktionen als Prädiktoren nur eine untergeordnete Rolle spielen.

Als subjektive Verhaltenseinschätzung wurde in der vorliegenden Untersuchung die Häufigkeit der Anwendung der verschiedenen Seminarthemen (*Bedürfnisbefriedigung, Gesprächstechniken, Umgang mit Einwänden und Umgang mit Gleichgültigkeit*) gemittelt. Dies geschah über ein Item aus der zweiten Nachbefragung („*In den letzten vier Wochen habe ich dieses Thema angewendet*“). Da die Operationalisierung stark am Seminar selbst ausgerichtet war, liegt möglicherweise eine Überschätzung der Zusammenhänge der Reaktionen mit dem so gemessenen Verhalten vor (Common source bias). Dies zeigt sich in den hohen Korrelationen zwischen den Anwendungseinschätzungen und der subjektiven Verhaltenseinschätzung, wobei hier vor allem die Korrelation zwischen *utility reactions* und Verhalten mit $r = .92$ auffällt und auf eine solche Überschätzung hinweist. Um diesem Umstand zu begegnen, sollte auf die Verwendung unabhängiger und v.a. psychometrisch konstruierter Instrumente zur Messung der jeweiligen Variablen geachtet werden.

Ein Ausbleiben des Zusammenhangs zwischen Wissen und subjektiver Verhaltenseinschätzung könnte daran liegen, dass beide Instrumente nicht reliabel genug sind, um jeweils das Wissen und das Verhalten zu messen. Andererseits ist es als problematisch anzusehen, dass der Wissenstest die behandelten Themen inhaltlich sehr spezifisch misst, wohingegen die hier verwendete Selbsteinschätzung der Themenanwendung ein einziges Item umfasst. Das Ausbleiben eines Zusammenhangs zwischen Einstellungen und Verhalten könnte möglicherweise ebenfalls daran liegen, dass die Items zu den Einstellungen im Vertriebsverhalten fünf Aspekte berücksichtigen (*Bedürfnisbefriedigung, Vertriebsorientierung, Umgang mit Einwänden, Kundenorientierung* und *vertriebsbezogenes Engagement*), während die subjektive Verhaltenseinschätzung die Anwendungshäufigkeit der vier Seminarthemen mittelt. Auch hier fehlt teilweise die von Tracey et al. (1995) erwähnte inhaltliche Überlappung, obwohl auf beiden Messebenen thematisch auf das Seminar Bezug genommen wurde. Eine schärfere inhaltliche Abstimmung zwischen subjektiver Verhaltenseinschätzung und Evaluationsbogen, mit dem die Einstellungen erhoben worden sind, könnte dem entgegenwirken.

Es ist aber zu überlegen, inwiefern die verwendete subjektive Verhaltenseinschätzung zur Messung von tatsächlichem Verhalten ausreicht oder ob der Evaluationsbogen von einer Zustimmungsskala in eine Häufigkeitsskala überführt werden sollte. Dadurch würden nicht mehr die Einstellungen zu einem bestimmten Verhalten in Vertriebssituationen gemessen, sondern vielmehr könnte die gezielte Nachfrage, ob und vor allem wie oft das gewünschte Verhalten innerhalb eines Zeitrahmens gezeigt wurde, ein valideres Maß für tatsächliches *Verhalten* darstellen. Zwar bleibt dies eine subjektive Messung, damit wird aber immerhin die Häufigkeit gezeigten Verhaltens erfasst. Es steht außer Frage, dass diese subjektiven Messungen nach Möglichkeit durch ein objektives Maß, bspw. durch externe Verhaltensbeobachtung, Fremdeinschätzungen durch Kollegen/Vorgesetzte oder „harte“ Vertriebskennzahlen zu ergänzen sind.

In Bezug auf den Einfluss der untersuchten Einflussgrößen nach Tannenbaum (Cannon-Bowers et al., 1995) auf die einzelnen Evaluationskriterien wären mehr signifikante Unterschiede auf der Verhaltensebene zu erwarten gewesen.

Betrachtet man die Korrelationen zwischen Einflussfaktoren und abhängigen Variablen und vergleicht sie mit den Ergebnissen der MANOVA, wären mehr signifikante multivariate Effekte denkbar gewesen. Da für die Berechnung der MANOVAs jedoch die überwiegend

intervallskalierten Messwerte der Einflussgrößen dichotomisiert wurden, um entsprechend die Stufen des jeweiligen Faktors der MANOVA zu erhalten, ging ein gewisser Anteil an Varianz verloren. Dies hat zur Folge, dass sich trotz hoher Zusammenhänge der AVs mit einem bestimmten Faktor in der MANOVA kein signifikanter multivariater Effekt zeigt.

Eine Einschätzung des subjektiven Trainingsbedarfs wurde in der vorliegenden Untersuchung über die Dichotomisierung des Items „*Die Inhalte haben meinem Bedarf entsprochen*“ aus dem Seminar-Feedbackbogen realisiert. Dies geschah in Ermangelung einer vorgeschalteten Bedarfsanalyse. Es wurde davon ausgegangen, dass die Teilnehmer über dieses Item nachträglich für sich selbst prüfen, inwiefern ihr subjektiver Trainingsbedarf durch die Inhalte des Seminars gedeckt wurde. Hier stellt sich wiederum die bereits andiskutierte Frage, inwiefern der Trainingsbedarf, den ein Unternehmen sieht und aufgrund dessen Mitarbeiter zu PE-Maßnahmen geschickt werden, mit dem subjektiven Trainingsbedarf des einzelnen Mitarbeiters übereinstimmt. Eine Bedarfsanalyse im Vorfeld von PE-Maßnahmen sollte demnach zum Ziel haben, den Bedarf des Unternehmens und des einzelnen Mitarbeiters soweit in Einklang zu bringen, dass von einem beidseitigen Bedarf ausgegangen werden kann.

6.5 Fazit und Ausblick

Bei der Evaluation von PE-Maßnahmen geht es letztendlich um eines: Qualitätssicherung. Als führender Theoretiker der Qualitätssicherung im Gesundheitsbereich definiert Donabedian (1980) Qualität als Übereinstimmung zwischen dem Ergebnis und den zuvor formulierten Zielen und unterteilt Qualität in drei Dimensionen: *Strukturqualität* als Gesamtheit an organisatorischen, finanziellen, räumlichen und personellen Ressourcen (Qualifizierung des Trainers, adäquate Seminarumgebung, geeignete Materialien und Trainingsstil, geplante Follow-up-Maßnahmen), *Prozessqualität* als Inbegriff aller Maßnahmen, die im Laufe des Trainings ergriffen oder nicht ergriffen werden (ausführliches Trainingskonzept) sowie *Ergebnisqualität* als eindeutigste Bezugsbasis für eine Qualitätsbeurteilung (Ausmaß an Kongruenz zwischen den zuvor formulierten Zielen und dem Resultat). Insgesamt spiegelt die Ergebnisqualität das Zusammenspiel von Struktur- und Prozessqualität wider, weshalb daher bei Untersuchungen zu Ergebnisqualität der Einfluss moderierender Variablen sorgfältig berücksichtigt werden muss. Jede PE-

Maßnahme sollte sich daran messen lassen, ob und wie stark sie zu einer Ergebnisverbesserung beigetragen hat. Diese Messungen werden, wie eingangs beschrieben, in der Praxis vorrangig durch die erste Ebene des Vier-Ebenen-Modell Kirkpatrick's (1996) durchgeführt, nämlich die Reaktionsebene.

Der gefundene Zusammenhang zwischen *utility reactions* und der subjektiven Verhaltenseinschätzung sowie das Ausmaß, in dem diese Reaktionsfacette zur Vorhersage der subjektiven Verhaltenseinschätzung beizutragen vermag, deutet trotz methodischer Mängel auf einen wichtigeren Stellenwert der ersten Kirkpatrick-Ebene hin als von einigen Autoren angenommen (z.B. Holton, 1996). Gerade hinsichtlich ihrer starken Vorhersagekraft für das Verhalten (und im Hinblick auf die vorherrschende Evaluation auf der Reaktionsebene) erweist sich diese Art von Reaktionen als sehr wertvoll, was die Forderung nach einer ausdifferenzierten Messung unterstützt (Warr & Bunce, 1995; Morgan & Casper, 2000). Diese Ausdifferenzierung lässt nicht nur eine Aufteilung in affektive und nutzenbezogene Reaktionen zu, sondern lädt dazu ein, im Sinne von Clement (1982) andere Reaktionen abzufragen, mit denen mögliche Umsetzungsbarrieren im Transferprozess identifiziert werden können.

Als eine solche Barriere ist z.B. bei negativer Ausprägung das Transferklima anzusehen, welches seinerseits als wichtiger Einflussfaktor im Hinblick auf die Effektivität eines Trainings gilt. Bei der Untersuchung weiterer individueller und organisationaler Einflussfaktoren aus dem Tannenbaum-Modell (Cannon-Bowers et al., 1995) zeigte sich ein Einfluss dieser Variablen nicht nur auf die „höheren“ Evaluationsergebnisse wie Lernen und Verhalten, sondern auch auf die Reaktionen.

Die durch die vorliegende Untersuchung gewonnenen Hinweise auf die Bedeutung der Teilnehmerreaktionen, ihre mehrfachen Messung sowie die Rolle von individuellen und organisationalen Einflussvariablen legen eine umfassende und systematische Evaluation nahe. Dabei gilt jedoch zu bedenken, dass es zwar aus wissenschaftlicher Sicht zur Klärung von Zusammenhängen und Kausalitäten unabdingbar ist, möglichst viele der als relevant gesehenen Facetten der Evaluationsebenen und der Rahmenbedingungen zu operationalisieren. Gerade im Hinblick auf die Verknüpfung zwischen wissenschaftlicher Forschung und Unternehmenspraxis ist jedoch die Ökonomie der Durchführung für Teilnehmer und Unternehmen zu berücksichtigen. Eine zu umfangreiche Befragung im Rahmen einer Evaluation führt durch die zeitliche Belastung zu einer geringen Beteiligungsquote bzw.

einer Reduzierung der Datenqualität, wenn die Befragungen nur oberflächlich bearbeitet werden. Aus rein pragmatischen Gründen wird daher bei Evaluationen „im Feld“ oftmals auf viele dieser Zusatz-Aspekte verzichtet – wobei eine Balance zwischen wissenschaftlichen Standards und Praktikabilität angestrebt werden sollte.

Im Hinblick auf die durch eine PE-Maßnahme angestrebten Ergebnisse scheint jedoch die Berücksichtigung hinderlicher und förderlicher Faktoren für die Umsetzung der Trainingsinhalte unabdingbar. Sie enthalten – über die Messung eines Lernerfolgs oder einer Verhaltensänderung hinaus – wertvolle Informationen bezüglich Wirksamkeit bzw. Unwirksamkeit einer Maßnahme. Bei Einbeziehung von Einflussfaktoren in die Messung von Erfolgskriterien entsteht somit aus einer als punktuell empfundenen Maßnahme ein nachhaltiger Umsetzungs- und Veränderungsprozess.

„Future research that pursues this line of inquiry is necessary if we are to go beyond the question of **whether** training works to the more important question of **why** training works.“³⁹

³⁹ Tracey et al. (1995), S. 250.

7 Literaturverzeichnis

§ 41 AFG

- Ajzen, I. & Fishbein, M. (1977). Attitude-behavior relations: A theoretical analysis and review of empirical research. *Psychological Bulletin*, 84, 888-918.
- Alliger, G. M. & Janak, E. A. (1989). Kirkpatrick's levels of training criteria: Thirty years later. *Personnel Psychology*, 42, 331-342.
- Alliger, G. M., Tannenbaum, S. I., Bennett, W. Jr., Traver, H. & Shotland, A. (1997). A meta-analysis of the relations among training criteria. *Personnel Psychology*, 50, 341-358.
- Alvarez, K., Salas, E. & Garofano, C. M. (2004). An integrated model of training evaluation and effectiveness. *Human Resource Development Review*, 3 (4), 385-416.
- Arthur, W., Bennett, W., Edens, P. S. & Bell, S. T. (2003). Effectiveness of training in organizations: A meta-analysis of design and evaluation features. *Journal of Applied Psychology*, 88 (2), 234-245.
- Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2000). *Multivariate Analysemethoden* (9. Aufl.). Berlin: Springer.
- Baldwin, T. T. & Ford, J. K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology*, 41, 63-105.
- Baldwin, T.T. & Magjuka, R. J. (1991). Organizational training and signals of importance: Linking pretraining perceptions to intentions to transfer. *Human Resource Development Quarterly*, 2, 25-36.
- Baldwin, T.T., Magjuka, R. J. & Loher, B. T. (1991). The perils of participation: Effects of trainee choice on motivation and learning. *Personnel Psychology*, 44, 51-66.
- Bates, R. (2004). A critical analysis of evaluation practice: the Kirkpatrick Model and the principle of beneficence. *Evaluation and Programm Planning*, 27, 341-347.
- Borchert, M. & Rutschke, K. (2005). Performance Improvement in deutschen Unternehmen. In K. H. Schwuchow & J. Gutman (Hrsg.), *Jahrbuch für Personalentwicklung* (S. 5-13). Neuwied: Luchterhand.
- Bortz, J. (1999). *Statistik für Sozialwissenschaftler* (5. Aufl.). Berlin: Springer.
- Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (2. Aufl.). Berlin: Springer.

- Brown, K. G. (2005). An examination of the structure and nomological network of trainee reactions: A closer look at "smile sheets". *Journal of Applied Psychology*, 90 (5), 991-1001.
- Buchester, S. (2003). Bildungscontrolling. *Unveröffentlichte Dissertation, Universität Hamburg*.
- Cannon-Bowers, J. A., Salas, E., Tannenbaum, S. I. & Mathieu, J. E. (1995). Toward theoretically-based principles of training effectiveness: A model and initial empirical investigation. *Military Psychology*, 7 (3), 141-164.
- Clement, R. W. (1982). Testing the hierarchy theory of training evaluation: An expanded role for trainee reactions. *Public Personnel Management*, 11, 176-184.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.
- Donabedian, A. (1980). *The definition of quality and approaches to its assessment. Explorations in quality assessment and monitoring*. Ann Arbor, MI: Health Administration Press.
- Eichenberger, P. C. (1990). Millionen für Bildung, Pfennige für Evaluierung. *Personalwirtschaft*, 3, 35-43.
- Facteau, J. D., Dobbins, G. H., Russell, J. E. A., Ladd, R. T. & Kudisch, J. D. (1995). The influence of general perceptions for the training environment on pretraining motivation and perceived training transfer. *Journal of Management*, 21, 1-25.
- Ford, J. K. & Kraiger, K. (1995). The application of cognitive constructs and principles to the instructional systems model of training: Implications for needs assessment, design, and transfer. *International Review of Industrial and Organizational Psychology*, 10, 3-48.
- Gebert, D. & Rosenstiel, L. v., (2002). *Organisationspsychologie: Person und Organisation* (5. Aufl.). Stuttgart: Kohlhammer.
- Goodman, J. S. & Wood, R. E. (2004). Feedback specificity, learning opportunities, and learning. *Journal of Applied Psychology*, 89 (5), 809-821.
- Gülpen, B. (1996). Evaluation betrieblicher Verhaltenstrainings. *Unveröffentlichte Dissertation, Universität München*.
- Hamblin, A. C. (1974). *Evaluation and control of training*. London: McGraw-Hill.
- Hertel, G., Orlikowski, B., Jokisch, W., Schöckel, D. & Haardt, C. (2004). Entwicklung, Durchführung und Evaluation eines Basistrainings für virtuelle Teams bei der Siemens AG. In G. Hertel & U. Konradt (Hrsg.), *Human Resource Management im Inter- und Intranet* (S. 313-325). Göttingen: Hogrefe.

- Höft, S. (2001). Erfolgsüberprüfung personalpsychologischer Aufgabenfelder. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 617-651). Göttingen: Hogrefe.
- Holling, H. & Liepmann, D. (1995). Personalentwicklung. In H. Schuler (Hrsg.), *Lehrbuch Organisationspsychologie* (S. 285-316). Bern: Huber.
- Holton, E. F. III. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7, 5-21.
- Industry Report 2000. (2000). *Training*, 37 (10), 45-48.
- Janssen, J. & Laatz, W. (2003). *Statistische Datenanalyse mit SPSS für Windows* (4. Aufl.). Berlin: Springer.
- Jöns, I. (1997). Rückmeldung von Befragungsergebnissen: Konzepte und Erfahrungen am Beispiel von Vorgesetztenbeurteilungen. *ABO-Original*, 4 (1), 2-9.
- Kallus, K. W. (1995). Der Erholungs-Belastungs-Fragebogen (EBF). Handanweisung. Frankfurt: Swets und Zeitlinger.
- Kallus, K. W. & Jiménez, P. (2005). Der Erholungs-Belastungs-Fragebogen (EBF-78-Work). Universität Graz.
- Kallus, K. W. & Schmut, B. (2004). Überarbeitete Fassung der deutschen Version des Job Diagnostic Survey (JDS) von Schmidt, Kleinbeck, Ottmann und Seidel (1985). Universität Graz.
- Kirkpatrick, D. L. (1959a). Techniques for evaluating training programs. *Journal of the American Society of Training Directors*, 13 (11), 3-9.
- Kirkpatrick, D. L. (1959b). Techniques for evaluating training programs: Part 2 - learning. *Journal of the American Society of Training Directors*, 13 (12), 21-26.
- Kirkpatrick, D. L. (1960a). Techniques for evaluating training programs: Part 3 - behavior. *Journal of the American Society of Training Directors*, 14 (1), 13-18.
- Kirkpatrick, D. L. (1960b). Techniques for evaluating training programs: Part 4 - results. *Journal of the American Society of Training Directors*, 14 (2), 28-32.
- Kirkpatrick, D. L. (1996). Revisiting Kirkpatrick's four-level model. *Training & Development*, 54-59.
- Kirkpatrick, D. L. (1998). *Evaluating training programs. The four levels* (2nd ed.). San Francisco: Berrett-Koehler.
- Konradt, U., Hertel, G. & Behr, B. (2002). Interkulturelle Managementtrainings: Eine Bestandsaufnahme von Konzepten, Methoden und Modalitäten in Deutschland. *Zeitschrift für Sozialpsychologie*, 33 (4), 197-207.

- Kraiger, K., Ford, J. K. & Salas, E. (1993). Application of cognitive, skill-based, and affective theories of learning outcomes to new methods of training evaluation. *Journal of Applied Psychology*, 78, 311-328.
- Kühnlein, G. (1997). "Vertrieblichung" von Weiterbildung als Zukunftstrend? *Arbeit*, 6 (3), 267-281.
- Latham, G. P. & Frayne, C. A. (1989). Self-management training for increasing job attendance: A follow-up and a replication. *Journal of Applied Psychology*, 74, 411-416.
- Lewin, K. (1947). Frontiers in group dynamics: II. Channels of group life: Social planning and action research. *Human Relations*, 1, 143-153.
- Lienert, G. A. & Raatz, U. (1998). *Testaufbau und Testanalyse* (6. Aufl.). Weinheim: Beltz-PVU.
- Lind, G. (2005). *Effektstärken: Praktische und theoretische Bedeutsamkeit* [On-line]. Verfügbar unter: http://www.uni-konstanz.de/ag-moral/pdf/Lind-2005_Effekstaerke-Vortrag.pdf [Zugriff am 08.06.2006].
- Locke, E. A. & Latham, G. P. (1990). A theory of goal setting and task performance. Englewood Cliffs, NJ: Prentice Hall.
- Mathieu, J. E., Martineau, J. W. & Tannenbaum, S. I. (1993). Individual and situational influences on the development of self-efficacy: Implications for training effectiveness. *Personnel Psychology*, 46, 125-147.
- Mathieu, J. E., Tannenbaum, S. I. & Salas, E. (1992). Influences of individual and situational characteristics on measures of training effectiveness. *Academy of Management Journal*, 35 (4), 828-847.
- Miller, G. A., Galanter, E. & Pribram, K. H. (1973). *Strategien des Handelns*. Stuttgart: Klett.
- Morgan, R. B. & Casper, W. J. (2000). Examining the factor structure of participant reactions to training: A multidimensional approach. *Human Resource Development Quarterly*, 11 (3), 301-317.
- Nadler, D. A. (1979). The effects of feedback on task group behaviour: A review of the experimental research. *Organizational Behavior and Human Performance*, 23, 309-338.
- Neuberger, O. (1994). *Personalentwicklung*. Stuttgart: Enke.
- Noe, R. A. (1986). Trainees' attributes and attitudes: Neglected influences on training effectiveness. *Academy of Management Review*, 11 (4), 736-749.
- Noe, R. A. & Schmitt, N. (1986). The influence of trainee attitudes on training effectiveness: Test of a model. *Personnel Psychology*, 39, 497-523.

- Nork, M. E. (1991). Management-Training: Evaluation - Probleme - Lösungsansätze. In T. R. Hummel, D. Wagner & E. Zander (Hrsg.), *Hochschulschriften zum Personalwesen* (Bd. 9). München: Hampp.
- Phillips, J. J. (2005). *Return on investment in der Personalentwicklung*. Berlin: Springer.
- Phillips, P. P. & Phillips, J. J. (2001). Symposium on the evaluation of training. *International Journal of Training and Development*, 5 (4), 240-247.
- Piezzi, D. (2002). Die Transferförderung in der betrieblichen Weiterbildung. Theoretische Modellbildung und empirische Untersuchung der Bedeutung der Arbeitsumgebung sowie der Integration der Weiterbildung in die Unternehmensführung. *Unveröffentlichte Dissertation, Universität St. Gallen*.
- Rank, B. & Thiemann, T. (1998). Maßnahmen zur Sicherung des Praxistransfers. In B. Rank & R. Wakenhut (Hrsg.), *Sicherung des Praxistransfers im Führungskräfte-Training* (S. 31-77). München: Mering.
- Reiter, H. (2005, März). *Bildungseffizienz-Umfrage 2004 des GABAL e.V.* Vortrag auf der didactica, Stuttgart.
- Rosenstiel, L. v., Molt, W. & Rüttinger, B. (2005). *Organisationspsychologie*. Stuttgart: Kohlhammer.
- Rouiller, J. Z. & Goldstein, I. L. (1993). The relationship between organizational transfer climate and positive transfer of training. *Human Resource Development Quarterly*, 4, 377-390.
- Salas, E. & Cannon-Bowers, J. A. (2001). The science of training: A decade of progress. *Annual Review of Psychology*, 52, 471-499.
- Sauter, E. (1995). Bildungspolitische Aspekte der Qualitätssicherung in der Weiterbildung. In J. E. Feuchthofen & E. Severing (Hrsg.), *Qualitätsmanagement und Qualitätssicherung in der Weiterbildung* (S. 22-39). Neuwied: Luchterhand.
- Schuler, H. & Prochaska, M. (2000). Entwicklung und Konstruktvalidierung eines berufsbezogenen Leistungsmotivationstests. *Diagnostica*, 46 (2), 61-72.
- Schyns, B. (1999, September). Entwicklung einer Skala zur beruflichen Selbstwirksamkeitserwartung. *Poster präsentiert auf der 1. Tagung der Fachgruppe Arbeits- und Organisationspsychologie der Deutschen Gesellschaft für Psychologie, Marburg*.
- Schyns, B. & Collani, G. v. (2002). Berufliche Selbstwirksamkeitserwartung (Occupational Self-Efficacy). In A. Glöckner-Rist (Hrsg.), *ZUMA-Informationssystem. Ein elektronisches Handbuch sozialwissenschaftlicher Erhebungsinstrumente, Version 6.00*. Mannheim: Zentrum für Umfragen, Methoden und Analysen.
- Scriven, M. (1991). *Evaluation thesaurus*. Newbury Park, CA: Sage.

- Sieber Bethke, F. (2003). *Kompendium Controlling, Evaluation und Reporting von Weiterbildung und Personalentwicklung*. Bremen: Medien-Institut.
- Stiefel, R. T. (1997). Evaluierung als Chance. *MAO, 1*.
- Tannenbaum, S. I., Mathieu, J. E., Salas, E. & Cannon-Bowers, J. A. (1991). Meeting trainees' expectations: the influence of training fulfillment on the development of commitment, self efficacy, and motivation. *Journal of Applied Psychology, 76*, 759-769.
- Thierau, H. (1991). Analyse und empirische Überprüfung wissenschaftlicher Evaluationskonzepte in der betrieblichen Weiterbildung – dargestellt am Beispiel der Schulung von Führungskräften in Personalbeurteilung. *Unveröffentlichte Dissertation, Universität Bochum*.
- Tracey, T. B., Tannenbaum, S. I. & Kavanagh, M. J. (1995). Applying trained skills on the job: the importance of the work environment. *Journal of Applied Psychology, 80*, 239-252.
- Transferre. (1955). In *Langenscheidts Enzyklopädisches Wörterbuch der lateinischen und deutschen Sprache. 1. Teil: Lateinisch-Deutsch unter Berücksichtigung der Etymologie* (9.Aufl.). Berlin: Langenscheidt.
- Valere. (1955). In *Langenscheidts Enzyklopädisches Wörterbuch der lateinischen und deutschen Sprache. 1. Teil: Lateinisch-Deutsch unter Berücksichtigung der Etymologie* (9.Aufl.). Berlin: Langenscheidt.
- Van Buren, M. E. & Erskine, W. (2002). *The 2002 state of the industry report – ASTD's annual review of trends in employer-provided training in the United States*. Alexandria, VA: American Society of Training and Development.
- Vroom, V. H. (1964). *Work and motivation*. New York: Wiley.
- Werner, D. (2006). *Trends und Kosten der betrieblichen Weiterbildung – Ergebnisse der IW-Weiterbildungserhebung 2005* [On-line]. Verfügbar unter: http://www.iwkoeln.de/data/pdf/content/trends01_06_2.pdf [Zugriff am 1.09.2006].
- Warr, P., Bird, M. & Rackham, N. (1970). *Evaluation of management training*. Epping, UK: Gower.
- Warr, P., Allan, C. & Birdi, K. (1999). Predicting three levels of training outcome. *Journal of Occupational and Organizational Psychology, 72*, 351-375.
- Warr, P. B. & Bunce, D. (1995). Trainee characteristics and the outcomes of open learning. *Personnel Psychology, 48*, 347-375.
- Wottawa, H. & Thierau, H. (2003). *Lehrbuch Evaluation* (3. Aufl.). Bern: Huber.

8 Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorgelegte Diplomarbeit mit dem Titel „*Die Bedeutung von Teilnehmereinschätzungen zu verschiedenen Zeitpunkten für die Vorhersage des Erfolgs von Personalentwicklungsmaßnahmen*“ selbständig verfasst und keine anderen als die angegebenen Hilfsmittel und Quellen verwendet habe.

Würzburg, 05.10.2006

Diana Beck