22 Erfolgsüberprüfung personalpsychologischer Arbeit

von Stefan Höft

Inhaltsübersicht

1 Einleitung und Überblick

Der Nachweis der Wirksamkeit und auch der Wirtschaftlichkeit dient nicht nur zur Gütekontrolle der geleisteten Arbeit, sondern hat auch eine Legitimationsfunktion für personalpsychologische Arbeit.

Im vorliegenden „Lehrbuch der Personalpsychologie“ wird ein großes Spektrum von Forschungsansätzen vorgestellt, in denen das Verhalten und Erleben von Individuen in Organisationen aus ganz unterschiedlichen Blickwinkeln analysiert wird. Neben konzeptionellen Verschränkungen verbindet alle Ansätze aber eine weitere Tatsache: Personalpsychologische Arbeit ohne Erfolgsüberprüfung ist auf längere Sicht zum Scheitern verurteilt. Der Nachweis der Wirksamkeit und auch der Wirtschaftlichkeit dient nicht nur zur Gütekontrolle der geleisteten Arbeit, sondern hat auch eine Legitimationsfunktion für personalpsychologische Arbeit allgemein gegenüber den meist fachlich unkundigen betrieblichen Entscheidungsträgern.

Der „Nachweis der Wirksamkeit und Wirtschaftlichkeit“ dient auch als (zugegebenermaßen grober) konzeptioneller Rahmen des vorliegenden Kapitels. In gewissem Sinne dient das Kapitel also als methodische Klammer für die zuvor berichteten Forschungsgebiete.

Als erstes wird die *Validierung eignungsdiagnostischer Verfahren* thematisiert. Hierfür wird zunächst das traditionelle Drei-Validitäten-Modell vorgestellt. Die kritische Auseinandersetzung mit den Implikationen dieses Ansatzes mündet in ein „Rahmenmodell zur Validität in der beruflichen Eignungsdiagnostik“, an dem die wichtigsten Problemfelder dieses Anwendungsgebiets veranschaulicht werden können.

Plazierungsstrategien sind gefragt, wenn neu eingestellte Personen oder wegen einer Umstrukturierung innerhalb einer Organisation freigesetzte Mitarbeiter neuen Arbeitsplätzen zugeteilt werden müssen. Im entsprechenden Abschnitt *Personalentscheidungen: Selektion und Klassifikation* werden einige grundlegende Konzepte vorgestellt.

Im Abschnitt *Evaluation von Personalentwicklungsmaßnahmen* wird zunächst ein typischer Evaluationsablauf skizziert. Nachdem einige Besonderheiten bei der Effektivitätsüberprüfung von Personalentwicklungsmaßnahmen behandelt wurden, dienen zwei Evaluationsbeispiele aus diesem Bereich zur Illustration möglicher Ansätze.

Im letzten Abschnitt über *ökonomische Nutzenanalysen* werden schließlich mehrere Modelle dargestellt, die Cascios (1991, p. vii) Aussage „the language of business is dollars, not correlation coefficients“ aufgreifen und monetäre Nutzenfunktionen personalpsychologischer Arbeit beschreiben.

2 Validierung eignungsdiagnostischer Verfahren

„Validität“ ist das wichtigste Gütekriterium der klassischen Testtheorie.

Güteprüfungen zu eignungsdiagnostischen Testverfahren (dieser Terminus soll nur in diesem Zusammenhang für alle konstrukt-, simulations- und biographieorientierten Verfahren verwendet werden; vgl. Kapitel 5-7) orientieren sich fast ausschließlich an den Gütekriterien, die im Rahmen der klassischen Testtheorie entwickelt wurden (vgl. z. B. Lienert & Raatz, 1994). Die Validität ist hiervon unzweifelhaft das wichtigste Kriterium.

2.1 Das traditionelle Drei-Validitäten-Modell

Traditionell wird zwischen drei Validitätsfacetten (konstrukt-, kontent- und kriteriumsbezogene Validität) unterschieden.

„Die Validität oder Gültigkeit eines Tests gibt den Grad der Genauigkeit an, mit dem dieser Test dasjenige Persönlichkeitsmerkmal oder diejenige Verhaltensweise, das (die) er messen oder vorhersagen soll, tatsächlich mißt oder vorhersagt“ (Lienert & Raatz, 1994, S. 10).

Die Standards für psychologische Testverfahren und Diagnoseinstrumente der „American Psychological Association“ von 1954 unterscheiden hierbei erstmals zwischen drei Validitätsaspekten (vgl. auch Cronbach & Meehl, 1955):

Charakterisierung der unterschiedlichen Validitätsfacetten

- Die *Kontent- oder Inhaltsvalidität* bezeichnet die Güte, mit der ein Test das interessierende Personmerkmal und seine Verhaltensäußerungen abbildet. Ein kontentvalider Test ist demnach repräsentativ für die möglichen Äußerungen des definierten Merkmals.
  In der Literatur sind zwar viele systematische Verfahren zur Konstruktion und Überprüfung kontentvalider Tests zu finden (z. B. Edmundson, Koch & Silverman, 1993), in der diagnostischen Praxis kommt die Kontentvaliditätsprüfung aber meist nicht über eine einfache Beurteilung durch Experten hinaus.
- Die *Kriteriumsbezogene Validität* eines Tests wird geprüft, indem die Testergebnisse mit Variablen außerhalb des Tests verglichen werden. Von diesen Variablen (den Kriterien) wird angenommen, daß sie das fragliche Merkmal in gleicher Weise erfassen. Je nach dem Zeitpunkt der Kriteriumsmessung werden zwei Formen unterschieden: Bei der konkurrenten Form erfolgt eine zeitgleiche Messung, während bei der prognostischen Form der kriteriumsbezogenen Validität die Kriteriumsmessung erst wesentlich später erfolgt.
- *Konstruktvalidität* behandelt die Frage, was hinter den Testergebnissen als verursachendes System (= Konstrukt) steht. Im Mittelpunkt steht also die Güte, mit der die als geltend angesehenen (psychologischen) Theorien und Konzepte die konkreten Ergebnisse erklären können. Cronbach (1990) unterscheidet beispielsweise drei konzeptionelle Annäherungsweisen an die Konstruktvalidität: logisch-inhaltliche Überlegungen, experimentelle Prüfung und korrelationsstatistische Analysen. In der Eignungsdiagnostik ist der letztere Ansatz am häufigsten anzutreffen. Hierbei kann zwischen Multitrait-Multimethod- und den schwächeren nomologischen Netzwerkanalysen unterschieden werden kann.
  Beim *Multitrait-Multimethod-Ansatz* (ursprünglich von Campbell & Fiske, 1959, eingeführt) werden mehrere Konstrukte simultan mit mehreren Verfahren erhoben. Es wird gefordert, daß Messungen eines Konstrukts über die unterschiedlichen Methoden hinweg hoch miteinander zusammenhängen (Konvergenz), während Messungen unterschiedlicher Konstrukte auch beim Einsatz desselben Meßinstruments einen möglichst niedrigen Zusammenhang aufweisen (Diskriminanz).
  Beim *nomologischen Netzwerkansatz* (ursprünglich von Cronbach & Meehl, 1955, eingeführt) dienen die Interkorrelationsmuster unterschiedlicher Konstruktmessungen als tentativer Hinweis auf Konstruktzusammenhänge. Da diesen Analysen meist nur sehr einfache Zusammenhangshypothesen zugrunde liegen und ein großer Freiraum für Alternativhypothesen verbleibt, stellt dieser Ansatz eine schwächere Form von Konstruktvalidierung dar (vgl. Cronbach, 1990).

In Kasten 1 soll die Validitätsüberprüfung im Sinne dieser Kriterien anhand einer hypothetischen Extraversionsskala beispielhaft und stark verkürzt veranschaulicht werden.

Ein Beispiel für die Untersuchung der unterschiedlichen Validitätsfacetten

Kasten 1:
Die Untersuchung der drei Validitätsfacetten am Beispiel einer Extraversionsskala

Personen mit einer positiven Ausprägung im Persönlichkeitsmerkmal „Extraversion" können als gesellig, aktiv, gesprächig, personen-orientiert, herzlich, optimistisch und heiter beschrieben werden (vgl. Borkenau & Ostendorf, 1993). Welche Kriterien muß eine „valide" Extraversionsskala gemäß den traditionellen Maßstäben erfüllen?

- Zur Wahrung der *Kontentvalidität* muß überprüft werden, ob alle wesentlichen Facetten des Extraversionskonstrukts in den Items abgebildet sind. Dies kann über einen Abgleich mit der relevanten Literatur und den bestehenden Theorien erfolgen, durch Konsultation persönlichkeitspsychologischer Experten usw.

- Für eine *kriteriumsbezogene Validitätsüberprüfung* kann das Ergebnis der betreffenden Skala in Beziehung gesetzt werden zu etablierten Extraversionsskalen. Möglich ist aber auch der Vergleich mit Bekanntenurteilen zur Extraversion. Als Kriterium kann auch eine Verhaltensstichprobe dienen, z. B. kann das Verhalten der Person unter kontrollierten Bedingungen auf Anzeichen für die Extraversionsausprägung analysiert werden.
- Hinweise zur *Konstruktvalidität* können über eine weitergehende Zusammenhangsanalyse unserer Skala mit anderen Verfahren erfolgen, die nicht unbedingt ein identisches Konstrukt messen. So ist z. B. eine hohe Korrelation des betreffenden Instruments mit einem konkurrierenden Extraversionsverfahren zunächst ein positiver (kriteriumsbezogener) Validitätshinweis. Wenn die Skala allerdings gleich hoch mit einem Verfahren korreliert, das Leistungsmotivation erheben soll, stellt diese mangelnde Diskriminanz den Wert der vorher gefundenen Konvergenz wieder in Frage. Möglicherweise gibt auch eine Binnenanalyse zur Skala (z. B. über eine explorative oder konfirmatorische Faktorenanalyse) weitere Hinweise zur Aufklärung des Konstruktzusammenhangs.

Weitere Möglichkeiten zur Konstruktvalidierung nennt Schuler (1996, S. 54). Das Beispiel zeigt aber schon so recht deutlich, daß innerhalb des vorgegebenen Rahmens der verschiedenen Validitätsansätze genügend Spielraum für unterschiedliche Überprüfungsstrategien verbleibt.

Bereits innerhalb des vorgegebenen Rahmens der verschiedenen Validitätsansätze bleibt großer Raum für unterschiedliche Überprüfungsstrategien.

Während diese Begriffe heute zum Standardrepertoire testorientiert arbeitender Wissenschaftler gehören, ist über ihre Exklusivität (*„Müssen noch andere Validitätsaspekte berücksichtigt werden?“*) und ihre Wertigkeit (*„Welcher Aspekt ist der wichtigste?“*) viel diskutiert worden.

So kritisiert Landy (1986) z. B. sehr heftig die unreflektierte Übernahme der drei Validitätsfacetten in die Praxis, bei der Testverfahren im Sinne eines Checklistenverfahrens ausschließlich hinsichtlich dieser Aspekte im Sinne einer „Dreifaltigkeitslehre“ untersucht werden.

Während die vereinfachte Validitätsdreiteilung weiterhin in der Praxis rege Anwendung findet, werden in neueren Validitätsmodellen die Kontent- und die kriteriumsbezogene Validität der Konstruktvalidität untergeordnet.

Die besondere Anziehungskraft der vereinfachten Validitätsdreiteilung zeigt sich auch in dem Umstand, daß die Praxisverwendung schon seit Jahren nicht mehr konform geht mit den Fortschritten in der Validitätstheorie: So wurde die Dominanz der drei Validitätsaspekte in den APA-Standards von 1954 bis 1985 (ein Vergleich findet sich bei Messick, 1989, pp. 28-30) schrittweise zurückgenommen. In der 1985er Ausgabe (deutsche Übersetzung: Häcker, Leutner & Amelang, 1998) werden die Validitäten nicht mehr als Typen eingeführt, sondern in abgeschwächter Form nur noch als Validitätskategorien („content related“, „criterion related“ und „construct-related evidence of validity“) bezeichnet.

In dem neueren Validitätsmodell von Messick (1995) werden kriteriums- und kontentbezogene Analysen nur noch als abgeleitete, hierarchisch untergeordnete Aspekte der Konstruktvalidität eingeordnet (vgl. auch Anastasi, 1986). Zusätzlich wird der Geltungsbereich des Valditätsbegriffs erweitert, indem Aspekte der Testanwendung (z. B. Relevanz und Nutzen) sowie Konsequenzen des Testeinsatzes (z. B. Wertimplikationen und soziale Konsequenzen) explizit berücksichtigt werden.

2.2 Ein Rahmenmodell zur Validität in der beruflichen Eignungsdiagnostik

Eine einfache Übertragung des Validitätszugangs der allgemeinen Diagnostik auf den Bereich der beruflichen Eignungsdiagnostik würde die Besonderheiten dieses Forschungsbereichs übergehen. In Abbildung 1 ist deshalb ein einfaches Rahmenmodell dargestellt, das in Anlehnung an Binning und Barrett (1989) einige für die Eignungsdiagnostik relevante Validitätsaspekte

vereinigt. Eine Validitätsprüfung kann auf dieser Grundlage als erweitertes Hypothesenprüfverfahren verstanden werden.

Ein einfaches Rahmenmodell soll die im Kontext der beruflichen Eignungsdiagnostik relevanten Validitätsaspekte verdeutlichen.

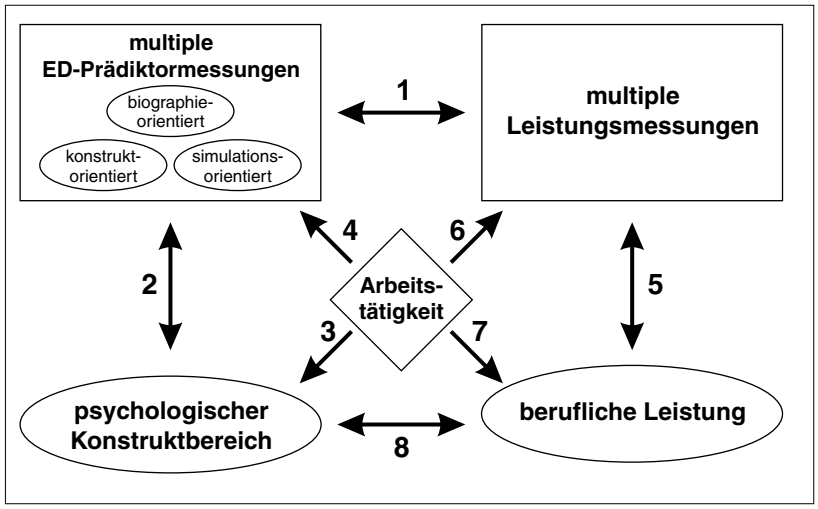


Abbildung 1:
Ein einfaches Rahmenmodell zur Validität in der beruflichen Eignungsdiagnostik

Das Prädiktor-Kriteriums-Modell und die Beschaffenheit von Prädiktoren

Konstrukt-, simulations- und biographieorientierte eignungsdiagnostische Verfahren haben immer die Funktion von Prädiktoren, die Berufserfolg als vorrangiges Zielkriterium vorhersagen sollen.

Die Vorhersage von berufserfolgsrelevanten Variablen mit Hilfe unterschiedlicher Verfahren steht im Mittelpunkt der eignungsdiagnostischen Praxis. Der Verfahrenseinsatz erfolgt also nicht als Selbstzweck, sondern hat immer die Funktion eines Prädiktors, mit dem Berufserfolg als vorrangiges Zielkriterium vorhergesagt werden soll. Diese spezifische Form von kriteriumsbezogener Validität (Schluß von der Testleistung einer Person auf deren spätere berufliche Leistung) ist also die intentionale Basis des eignungsdiagnostischen Testeinsatzes.

Im Rahmenmodell ist diese Beziehung als *Zusammenhang 1* gekennzeichnet. Ausgehend von dieser Prämisse haben sich in der Eignungsdiagnostik verschiedene Testkonstruktionsansätze entwickelt, die jeweils im Hinblick auf eine möglichst gute Vorhersage von Variablen des Berufserfolgs verschiedene Strategien verfolgen.

Die verschiedenen eignungsdiagnostischen Ansätze legen unterschiedliche Gewichte auf den Bezug zum psychologischen Konstruktbereich, zur Arbeitstätigkeit und zum Leistungskomplex.

Konstruktorientierte Verfahren (vgl. Kapitel 5) stützen sich auf „wohldefinierte“ Personeigenschaften (*Zusammenhang 2*), die sich in einer Arbeitsanalyse (vgl. Kapitel 3) als berufsrelevant herausgestellt haben (*Zusammenhang 3*).

Simulationsorientierte Verfahren (vgl. Kapitel 6) streben eine direkte Abbildung der erfolgskritischen Aspekte der Arbeitstätigkeit an (*Zusammenhang 4*). Der Bezug zu psychologischen Konstrukten erfolgt post hoc und ist schwierig, wie die Untersuchungen zur Konstruktvalidität des Assessment Centers gezeigt haben.

Die dritte Gruppe der biographiebezogenen Verfahren (vgl. Kapitel 7) ist nicht so eindeutig charakterisiert, da sie eher eine Restkategorie von Verfahren darstellt, die nicht eindeutig der konstrukt- oder simulationsorientierten Gruppe zugeordnet werden können. Bei vielen Verfahren (prototypisch: biographische Fragebogen) steht die Maximierung des Zusammenhangs mit Berufserfolgsvariablen, also wiederum *Zusammenhang 1,* im Vordergrund. Erst seit einigen Jahren wird verstärkt versucht, die erhobenen biographischen Informationen auch mit psychologischen Konstrukten zu verbinden und damit die Prognosegüte von Items zu erklären (Mumford, Costanza, Connelly & Johnson, 1996).

Beispiel für unterschiedliche eignungsdiagnostische Ansätze bei der Auswahl von Verkaufspersonal

Kasten 2:
Beispiel für eigenschafts-, simulations- und biographieorientierte eignungsdiagnostische Verfahren

Welche eignungsdiagnostischen Verfahren können für die Auswahl von Verkaufspersonal im Einzelhandel eingesetzt werden?

Die Eigenschaft „Extraversion" umschreibt unter anderem Aspekte wie „aktiv auf jemanden zugehen", „gesprächig" oder „personenorientiert". Diese Facetten sind sicherlich hilfreich, um einen guten Kontakt zu einem Kunden aufzunehmen, und man könnte die Hypothese aufstellen, daß extravertiertes Verkaufspersonal unter Konstanthaltung aller anderen Bedingungen größeren Berufserfolg aufweist als introvertiertes Personal. Zur Konstruktmessung kann ein *eigenschaftsorientiertes Verfahren*, z. B. ein allgemeines Persönlichkeitsinventar (vgl. Kapitel 5), eingesetzt werden.

Alternativ könnte man allerdings auch einen erfolgsrelevanten Tätigkeitsausschnitt des Berufs simulieren. Hierfür kann z. B. ein Rollenspiel konzipiert werden, in dem die Prüflinge einem noch unschlüssigen Kunden ein Produkt offerieren müssen. Für das Abschneiden in diesem *simulationsorientierten Verfahren* wird die Ausprägung von „Extraversion" sicherlich eine nicht unbedeutende Rolle spielen, zusätzlich sind aber z. B. auch verbale Fähigkeiten, ein stringenter Argumentationsaufbau sowie sicheres Auftreten erfolgsrelevant. Schwierig ist die ersatzweise Definition eines globalen Konstrukts „verkäuferische Fähigkeiten". Dieses Etikett umschreibt ein sehr heterogenes Fähigkeitsbündel. Messungen der „verkäuferischen Fähigkeiten" werden aus diesem Grund ganz sicher nicht den Anforderungen gerecht, die an die Konstruktvalidität einer Eigenschaftsmessung gestellt werden (vgl. die Konstruktvaliditätsproblematik im Assessment Center; vgl. Kapitel 6).

Bei der Konstruktion eines biographischen Fragebogens (Prototyp des *biographischen Verfahrensansatzes*) könnte sich z. B. das Item „Ich lese in meiner Freizeit viele Romane" als trennscharf zwischen erfolgreichen und weniger erfolgreichen Verkäufern herausstellen (Verneinung des Items = erfolgreicher Verkäufer). Zur „blinden" empirischen Leistungsprognose ist es nicht notwendig, zu wissen, worauf dieser Zusammenhang beruht. Will man ihn allerdings erklären, ist es erforderlich, auf die psychologische Konstruktebene zurückzugreifen. Das Item kann als indirektes Maß für die Extraversionsausprägung der Person interpretiert werden (Verneinung des Items = hohe Ausprägung in Extraversion).

Die Beschaffenheit von beruflichen Kriterien

Die Konzeption des Konstrukts „berufliche Leistung" schwankt in unterschiedlichen Anwendungen zwischen der Gestaltung als heterogenes „kompositorisches Kriterium" und der Umsetzung als homogenes „multiples Kriterium".

In der rechten unteren Ecke des Rahmenmodells ist die „berufliche Leistung" dargestellt. Ähnlich wie der „psychologische Konstruktbereich" ist Leistung nur indirekt über entsprechende Indikatoren erfaßbar. Die Konstruktion der Leistungskriterien und ihr Zusammenhang mit dem hypothetischen Konstrukt „berufliche Leistung" (*Zusammenhang 5*) ist allerdings nicht direkt mit der Situation bei psychologischen Eigenschaftskonstrukten gleichzusetzen. Traditionell können zwei Extrempositionen unterschieden werden (vgl. Binning & Barrett, 1989): ein „(ökonomisches) kompositorisches Kriteriumsmodell" und ein „multiples Kriteriumsmodell". Beim „kompositorischen Kriteriumsmodell"-Ansatz wird ein Gesamtscore von a priori als unabhängig und gleichberechtigt angesehenen Einzelkriterien gebildet: Von jedem Kriterium wird angenommen, daß es einen eigenen Leistungsaspekt der Arbeitstätigkeit abbildet (*Zusammenhang 6*) und somit separat unter Umgehung eines allgemeinen Konstruktbegriffs umgesetzt werden muß. Beim „multiplen Kriteriums"-Ansatz wird hingegen analog dem psychologischen Konstruktbereich ein aus der Arbeitstätigkeit abgeleitetes (*Zusammenhang 7*) Konstrukt „berufliche Leistung" mit teilweise überschneidenden Facetten definiert. Von den Indikatoren wird analog zum psychologischen Konstruktbereich dann eine hohe Konvergenz verlangt.

Bei der wissenschaftlichen Untersuchung von Leistungsbeurteilungen wird z.Z. eher ein Mischmodell favorisiert, in dem konzeptionell und statistisch trennbare Unterdimensionen der Leistung unterschieden werden (vgl. Kapitel 15). In der Praxis reicht die Kriteriumsmessung aber meist nicht über eine globale Vorgesetztenbeurteilung hinaus.

Aktuell wird ein berufliches Leistungsmodell favorisiert, bei dem konzeptionell trennbare und nur mäßig korrelierende Leistungsunterdimensionen unterschieden werden.

Kasten 3:
Extraversionsrelevante Leistungskriterien

Beispiel für extraversionsrelevante Leistungskriterien

Die einschlägigen Metaanalysen zur kriteriumsbezogenen Validität von Extraversion (vgl. Kapitel 5) weisen für Extraversion einen (relativ niedrigen) positiven Zusammenhang für allgemeine Leistungskriterien aus (zwischen .08 und .16). Die Moderatoranalyse von Barrick und Mount (1991) offenbart dabei einen höheren Zusammenhang für verkäuferische Berufe (korrigiertes r = .18). Ganz im Sinne der Prädiktor-Kriteriums-Symmetriehypothese (vgl. z. B. den Abschnitt zur manuellen Arbeitsprobe im Kapitel 6) kann der vergleichsweise hohe Zusammenhang von Extraversion mit Trainingserfolgskriterien interpretiert werden. Nach Barrick und Mount zielten die meisten dieser Trainings auf eine Aktivierung, also extraversionstypisches Verhalten, ab.

Zu beachten sind natürlich immer auch komplexere Zusammenhänge, z. B. kurvilineare Wirkzusammenhänge (eine sehr hohe Extraversionsausprägung dürfte eher interaktionshinderlich als förderlich sein) oder mögliche situative Moderatoreinflüsse (vgl. z. B. Kapitel 5, zum Einfluß von Entlohnungssystemen).

Der Bezug zu den traditionellen Validitätsfacetten

Die traditionellen Validitätsfacetten lassen sich als Teilfragestellungen des Rahmenmodells identifizieren.

Die traditionellen Validitätsfacetten finden sich als Teilfragestellungen des Gesamtmodells wieder. Konstruktvaliditätsfragen zielen auf den *Zusammenhang 2* ab, wobei der Konstruktbezug bei simulationsorientierten und biographiebezogenen Verfahren konstruktionsbedingt nur indirekt erfolgen kann. Dementsprechend können die traditionellen diagnostischen Analysestrategien (vgl. Abschnitt 2.1) ohne weitere Modifikationen nur für konstruktorientierte Verfahren eingesetzt werden.

Auch Kontentvaliditätsaspekte können in direkter Form zunächst nur für die konstruktorientierten Verfahren geprüft werden. Relevant für die psychologische Konstrukterfassung sind nur die berufsbezogenen Aspekte des Konstrukts. Im Sinne eines „deduktiven Ansatzes“ (Klimoski, 1993) sollten dann Indikatoren abgeleitet werden, die repräsentativ für die am Arbeitsplatz relevanten Erlebens- und Verhaltensaspekte sind. Kontentvalide Aspekte werden in diesem Sinne also als spezifische Operationalisierungsstrategien der Konstrukte verstanden. Strittig ist allerdings (Guion, 1977; Tenopyr, 1977), ob die so erreichte Repräsentativität für die Konstruktvalidität der Messungen notwendig ist, oder ob damit nur die Akzeptanz bei Laien (i.S. einer „Augenscheinvalidität“) erhöht wird.

Eine andere Form der Kontentvalidität wird bei simulationsorientierten Verfahren problematisiert. Im Sinne eines „induktiven“ Konstruktionsansatzes sind sie zunächst repräsentativ hinsichtlich der leistungsrelevanten Tätigkeitsmerkmale des Arbeitsplatzes, und erst danach wird geschlußfolgert, welche Konstrukte dabei involviert sein könnten.

Die kriteriumsbezogene Validitätsanalyse (*Zusammenhang 1*) wurde bereits bei der Darstellung des Prädiktor-Kriterium-Submodells ausführlicher behandelt. Festzuhalten bleibt hier, daß häufig Validitätsaussagen abgeleitet werden, die sich eher auf den nicht direkt beobachtbaren Zusammenhang der beiden hypothetischen Konstruktbereiche „psychologisches Eigenschaftskonstrukt“ und „berufliche Leistung“ beziehen (*Zusammenhang 8*). Das Rahmenmodell verdeutlicht, daß dabei verschiedene andere involvierte Zusammenhänge vernachlässigt werden. So können z. B. bei mangelhaftem kriteriumsbezogenen Validitätsnachweis eines konstruktbezogenen Verfahrens (niedriger *Zusammenhang 1*) neben einem tatsächlich geringen Zusammen-

hang (niedriger *Zusammenhang 8*) die Gründe auch in einer mangelhaften Konstruktoperationalisierung (niedriger *Zusammenhang 2*), einer mangelhaften Leistungserfassung (niedriger *Zusammenhang 5*) oder in Mängeln der durchgeführten Arbeitsanalyse (niedriger *Zusammenhang 3* oder *7* auf der Konstrukt- bzw. *4* und *6* auf der Operationalisierungsebene) liegen.

Potentielle Ursachen für eine geringe Validität von Extraversionsmessungen für Berufserfolg

Kasten 4:
Potentielle Ursachen für eine geringe Validität von Extraversionsmessungen für Berufserfolg

Die für Extraversion gefundenen kriteriumsbezogenen Validitätskoeffizienten zwischen .08 und .16 sind zwar metaanalytisch abgesichert, aber hinsichtlich ihrer absoluten Größe eher bescheiden. Wo liegen mögliche Ursachen für diese Befundlage?

- Das Rahmenmodell würde zunächst eine Überprüfung der Konstruktoperationalisierung (*Zusammenhang 2*) nahelegen. Da in den meisten Studien mit etablierten Persönlichkeitsverfahren gearbeitet wurde, sind Mängel in diesem Bereich eher unwahrscheinlich.
- Eine mangelhafte Leistungserfassung (*Zusammenhang 5*) wäre schon eher wahrscheinlich, da in den meisten Studien relativ unreflektiert auf globale, nicht weiter strukturierte Leistungsbeurteilungen durch Vorgesetzte zurückgegriffen wird.

Eine weitere Ursache kann eine unzureichende Berücksichtigung der konkreten Arbeitstätigkeit sein. Hierbei können Defizite auf der Konstrukt- (*Zusammenhang 3* und *4*) oder Leistungsseite (*Zusammenhang 6* und *7*) eine Rolle spielen.

- Extravertierte Personen können u.a. mit den Begriffen „gesprächig", „offen", „unternehmungslustig" oder „aufgeregt" beschrieben werden. Während die ersten beiden Attribute erkennbar tätigkeitsrelevant für verkäuferische Berufe sind, beziehen sich die letzten beiden Aspekte eher auf allgemeine Verhaltensäußerungen, die besser im Freizeitverhalten beobachtet werden können. Allgemeine Persönlichkeitsinventare würden Extraversion also potentiell mit einem Bedeutungsüberschuß für berufliche Tätigkeiten erfassen. Entsprechend leitet sich hieraus die mögliche Forderung zur Konstruktion von konstruktorientierten Verfahren ab, bei denen die erfaßten Konstruktinhalte explizit auf die Arbeitstätigkeit abgestimmt werden. Allerdings müßte die Aktualität der Verfahren dann wegen möglicher Veränderungen der Tätigkeitsanforderungen über die Zeit regelmäßig überprüft werden.
- Neben der prinzipiellen Tauglichkeit (= Validität) des Leistungsbeurteilungssystems spielt in diesem Zusammenhang auch die Einschlägigkeit der erhobenen Kriterien für Extraversion eine Rolle (vgl. Kasten 3). Höhere Zusammenhänge ergeben sich erwartungsgemäß in unserem Verkaufspersonalbeispiel für verhaltensnahe Kriterien mit Bezug zu interpersonalen Verkaufssituationen.

2.3 Validierung als Hypothesentestung

Der Validierungsprozeß kann als Hypothesenprüfverfahren zu bestimmten Zusammenhangsannahmen im eignungsdiagnostischen Rahmenmodell verstanden werden.

Die Beschreibung des Rahmenmodells sollte gezeigt haben, daß die einzelnen Validitätsaspekte nur willkürliche Einzelaspekte aus dem Gesamtrahmen der Validität in der beruflichen Eignungsdiagnostik ansprechen. Für eine umfassende Validitätsprüfung müssen alle Aspekte gleichermaßen berücksichtigt werden. Im Sinne von Landy (1986) oder Hogan und Nicholson (1988) definiert sich der Validierungsprozeß dann als Hypothesenprüfverfahren zu bestimmten Zusammenhangsannahmen. Erhobene empirische Daten (z. B. kriterienbezogene Validitätsangaben) beziehen sich dabei nicht isoliert auf einen einzigen Zusammenhang, sondern sind immer vor dem Hintergrund der übrigen (Zusammenhangs-)Annahmen zu interpretieren. Hohe Zusammenhänge bei kriterienbezogenen Validitätsstudien sind so nur

zu erwarten, wenn zuvor eine valide Prädiktor- und Kriteriumskonstruktion erfolgt ist.

Während eine praxisorientierte Eignungsdiagnostik naturgemäß ihr Hauptaugenmerk auf den manifesten (korrelativen) Zusammenhang zwischen den eignungsdiagnostischen Prädiktoren und beruflichen Leistungsmessungen legen wird (*Zusammenhang 1*), wird eine wissenschaftlich orientierte Eignungsdiagnostik darüber hinausgehende Analysen zum „eigentlichen" Zusammenhang der Konstrukte (*Zusammenhang 8*) und seiner Wirkgrundlage anstellen müssen.

Eine praxisorientierte Eignungsdiagnostik wird sich mit der Maximierung des manifesten Zusammenhangs zwischen eignungsdiagnostischen Prädiktoren und Leistungskriterien begnügen. Für eine wissenschaftliche Eignungsdiagnostik ist eine weitergehende Aufklärung der Konstruktzusammenhänge notwendig.

3 Personalentscheidungen: Selektion und Klassifikation

Nachdem im vorherigen Abschnitt die Validitätsanalyse im Kontext der beruflichen Eignungsdiagnostik im Vordergrund stand, sollen in diesem Unterkapitel unterschiedliche Personalentscheidungssituationen diskutiert werden. Die beiden Abschnitte sind inhaltlich durch eine Mittel-Zweck-Relation miteinander verbunden: Eine valide Diagnose (Abschnitt 2) ist notwendig, um eine richtige Personalentscheidung (Abschnitt 3) zu treffen.

Zunächst soll eine Typologie von Entscheidungssituationen aufgestellt werden, bevor exemplarisch die Problemfelder von Klassifikationsentscheidungen diskutiert werden.

3.1 Eine Typologie von Personalentscheidungssituationen

Cronbach und Gleser (1965) stellen in ihrem Buch „Psychological tests and personnel decisions" eine einfache Typologie für Personalentscheidungssituationen mit Zuordnungen von Personen zu Arbeitsplätzen vor (p. 16). Über sechs Fragen mit jeweils zwei Antwortalternativen wird das vorliegende Problem einer von 2^6=64 Möglichkeiten zugeordnet.

Cronbach und Gleser (1965) teilen die möglichen Personalentscheidungssituationen in 2^6=64 verschiedene Gruppen ein.

Frage 1:
(a) Wird der Nutzen einer Entscheidung für alle Personen gleichsinnig bewertet? oder
(b) Werden unterschiedliche Nutzenwerte für jede Person verwendet?

Diese Frage zielt auf die Unterscheidung zwischen einer institutionellen (= 1a) oder individuumsspezifischen Entscheidungssituation (= 1b). Bei der ersten Alternative werden Entscheidungen hinsichtlich vieler Personen mit einem gleichbleibenden Werte- bzw. Entscheidungssystem getroffen (z. B. Selektion von Bewerbern für einen Arbeitsplatz in einem Unternehmen, Zuordnung bereits eingestellter Mitarbeiter zu unterschiedlichen Arbeitsplätzen). Diese Entscheidungsart steht im Mittelpunkt der meisten in diesem Zusammenhang behandelten Problemsituationen. Bei der zweiten Alternative muß ein Individuum eine Entscheidung zu unterschiedlichen Wahlmöglichkeiten treffen (z. B. kann ein Bewerber vor der Wahl stehen, ein Jobangebot einer Organisation anzunehmen oder sich doch für ein Konkurrenzangebot zu entscheiden). Obwohl diese Problemkonstellation konzeptionell ähnlich gehandhabt werden kann wie institutionelle Entscheidungen, stehen explizit darauf abzielende Forschungen erst am Anfang (vgl. z. B. Smith, Farr & Schuler, 1993).

Frage 2:
(a) Wird für jede Person eine eigenständige Entscheidung vorgenommen? oder
(b) Sind Entscheidungen zu mehreren Personen voneinander abhängig?

Die Alternative 2b tritt beispielsweise in Kraft, wenn in einer Auswahlsi-

tuation bestimmte Quoten vorgegeben werden. Die Entscheidung, eine Person einzustellen oder nicht, hängt dann davon ab, wie viele andere Personen bereits eingestellt bzw. abgelehnt wurden.

Frage 3:
(a) Wird eine Person nur einer der gegebenen Zuweisungsalternativen zugeordnet? oder
(b) Kann eine Person auch mehreren Alternativen zugeordnet werden?

Die sechs dichotomen Fragen sollten die jeweils vorliegende Personalentscheidungssituation hinreichend gut charakterisieren.

In einer Auswahlsituation wird eine Person z. B. einer der beiden „Treatment"-Alternativen „angenommen" vs. „abgelehnt" zugewiesen. Bei einer Personalentwicklungsdiagnose kann eine Person hingegen möglicherweise mehreren Trainingsmaßnahmen zugewiesen werden.

Frage 4:
(a) Ist eine der erlaubten Treatmentalternativen „Ablehnung"? oder
(b) Verbleiben alle Personen in der Organisation?

Hier wird das Unterscheidungsmerkmal zwischen „Selektion" und „Klassifikation" angesprochen. Bei Selektionen wird nur eine Teilstichprobe der Personen in die Organisation aufgenommen und plaziert. Bei Klassifikationen wird allen vorhandenen Personen ohne Ausnahme ein Arbeitsplatz zugewiesen.

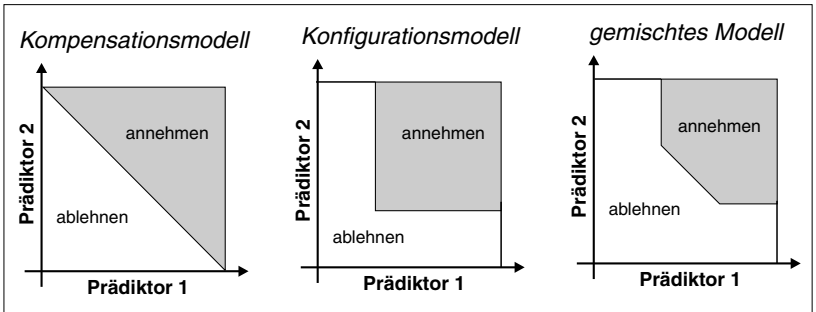
Frage 5:
(a) Basiert die Entscheidung auf einer univariaten Information? oder
(b) Liegt eine multivariate Informationsbasis vor?

Als Entscheidungsgrundlage kann zum einen ein einzelner Wert (z. B. ein Kompositorium aus einzelnen Testergebnissen) dienen. Eine Klassifikation reduziert sich so zu einer einfachen „Plazierung", bei der verschiedene cut-off-Werte die Treatmentzuweisung regeln. Beispielsweise kann eine Teilnehmergruppe bei einer Fremdsprachenschulung auf der Grundlage eines Sprachtests (= univariate Informationsbasis) unterteilt werden in Teilgruppen mit abgestufter Sprachkompetenz. Jede Teilgruppe erhält dann einen dem Wissensstand angemessenen Sprachunterricht.

Andererseits werden häufig mehrere entscheidungsrelevante Informationen gesammelt, die nicht ohne weiteres über einen Summenwert verknüpft werden können (multivariate Informationsbasis). Die Verknüpfung der Prädiktoren hängt dann von der Natur der Variablen bzw. der Festlegung des Entscheiders ab. In Kasten 5 sind in Anlehnung an Cronbach und Gleser (1965) bzw. Schuler (1996) drei Kombinationsmöglichkeiten für zwei Prädiktoren in einer Personalauswahlsituation dargestellt.

Mögliche Kombination von zwei Prädiktoren bei einer Auswahlsituation

Kasten 5:
Mögliche Kombinationen von zwei Prädiktoren bei einer Auswahlentscheidung



Die Verrechnung von zwei Prädiktoren erfolgt in den meisten Personalauswahlsituationen in Form eines Kompensationsmodells (vgl. linke Abbildung). Schwächen bei einem Prädiktor können durch Stärken bei einem anderen Prädiktor kompensiert werden. Andererseits können aber auch Mindeststandards definiert werden, die beide Prädiktoren minde-

stens erreichen müssen (Konfigurationsmodell; mittlere Abbildung). Realistisch dürfte wohl ein gemischtes Modell sein (rechte Abbildung), bei dem beide Prädiktoren Mindeststandards erfüllen müssen, darüber hinaus aber eine gegenseitige Kompensation möglich ist.

Ein einfaches Beispiel soll dieses verdeutlichen: Ein Verkäufer sollte neben verkaufsförderlicher Extraversion auch ein gewisses Maß an Gewissenhaftigkeit aufweisen. Gegebenenfalls kann ein sehr gewissenhafter Verkäufer mangelnde Extraversion ausgleichen und so erfolgreich arbeiten (= Kompensation). Allerdings müssen die Ausprägungen auf den Prädiktorkonstrukten gewisse (über eine Anforderungsanalyse ermittelten) Mindestwerte überschreiten (= durch die zusätzlichen Konfigurationsanforderungen resultiert ein gemischtes Modell).

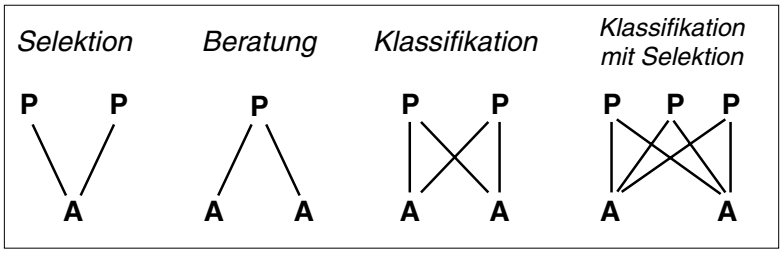
Frage 6:
(a) Werden endgültige Entscheidungen getroffen? oder
(b) Werden noch weitere Informationen vor einer endgültigen Entscheidung eingeholt?

Die letzte Frage zielt auf einstufige oder mehrstufige Personalentscheidungen ab. Bei mehrstufigen Personalauswahlverfahren wird der Bewerberpool beispielsweise schrittweise reduziert, indem immer aufwendigere (und validere) Verfahren eingesetzt werden (vgl. Kapitel 6).

Die in der Praxis auftretenden Formen der Zuordnung von Personen und Arbeitsplätzen sollten anhand dieser sechs Fragen hinreichend charakterisiert werden. Vier häufig auftretende Zuordnungssituationen sind in Kasten 6 beschrieben.

Kasten 6:
Typische Formen der Zuordnung von Personen (P) und Arbeitsplätzen (A) (nach Schuler, 1996, S. 153)

Einige typische Formen der Zuordnung von Personen und Arbeitsplätzen



Bei der *Selektion* wird für die Besetzung eines Arbeitsplatzes eine Auswahl aus mehreren Personen getroffen. Ohne weitergehende Informationen zur konkreten Vorgehensweise liegt im einfachsten Fall eine Personalentscheidung des Typs abaaaa vor (gleichsinnige Nutzenbewertung, abhängige Entscheidung, nur eine Zuweisungsalternative, mit Ablehnung, univariate Information, endgültige Entscheidung). Die *Beratung* stellt eine Inverse der Selektion dar. Für eine Person wird ein geeigneter Arbeitsplatz ausgesucht. Im einfachsten Fall liegt eine Entscheidung des Typs baaaaa vor. Bei der *Klassifikation* werden n Personen auf n Arbeitsplätze verteilt (im einfachsten Fall Typ ababaa). Bei der *Klassifikation mit Selektion* schließlich findet eine Auswahl unter mehreren Personen bei der Zuweisung zu mehreren Arbeitsplätzen statt (im einfachsten Fall Typ abaaaa).

Im weiteren soll beispielhaft die Problemsituation der Klassifikation etwas ausführlicher dargestellt werden.

3.2 Klassifikationsentscheidungen: Zuordnung von *n* Personen zu *m* Arbeitsplätzen

Bei der Klassifikation werden *n* Personen ohne Selektion *m* verschiedenen Arbeitsplätzen (bei *n=m*) oder Arbeitsplatzgruppen (bei *n>m*) zugewiesen.

Bei einer betrieblichen Klassifikationsentscheidung werden *n* Personen ohne Selektion *m* verschiedenen Arbeitsplätzen (bei *n=m*) oder Arbeitsplatzgruppen (bei *n>m*) zugewiesen. Wenn die Entscheidung auf mehreren Kriterien (multivariate Informationsbasis) basiert, liegt eine „echte" Klassifikation vor, ansonsten wird von einer „Plazierung" gesprochen.

Eine solche Situation kann z. B. nach einer internen Umstrukturierung der Organisation eintreten, wenn den bisherigen Mitarbeitern neue Arbeitstätigkeiten zugeordnet werden müssen. Im militärischen Bereich ergibt sich diese Problemstellung, wenn die Rekruten nach einer allgemeinen Grundausbildung verschiedenen Waffengattungen zugeteilt werden sollen.

Eine geeignete Klassifikationsstrategie muß einen tragbaren Kompromiß finden zwischen dem institutionellen Ziel, jeden Arbeitsplatz mit dem bestgeeigneten Mitarbeiter zu besetzen, und dem individuellen Ziel jedes Mitarbeiters, einen seinem Potential entsprechenden Arbeitsplatz zu finden.

Bei der Klassifikation entsteht ein Spannungsfeld zwischen institutionellen und individuellen Zielkriterien: Die Organisation ist daran interessiert, jeden Arbeitsplatz mit dem bestgeeigneten Mitarbeiter zu besetzen. Jeder Mitarbeiter wird daran interessiert sein, einen Arbeitsplatz zu besetzen, der seinem Potential am ehesten entspricht. Eine Zuordnungsstrategie wird einen Kompromißweg zwischen diesen Zielsetzungen finden müssen. In Kasten 7 ist die Problematik an einem einfachen Beispiel mit einer Variable veranschaulicht (ursprünglich aus Ghiselli & Brown, 1955, hier in der adaptierten Version von Schuler, 1996).

Kasten 7:
Drei Klassifikationsstrategien für die Zuordnung von Personen zu Arbeitsplätzen (nach Schuler, 1996, S. 158)

Beispiel für drei Klassifikationsstrategien für die Zuordnung von Personen zu Arbeitsplätzen

	Arbeitsplatz 1	Arbeitsplatz 2	Arbeitsplatz 3
Minimalqualifikation	7	4	1
Qualifikation von Person A	8	7	9
Qualifikation von Person B	6	5	4
Qualifikation von Person C	2	3	1
Klassifikation I nach Leistungserwartung	A	A	A
Klassifikation II nach individueller Qualifikation	B	C	A
Klassifikation III nach anforderungsgemäßer Besetzung aller Stellen	A	B	C

In diesem vereinfachten Beispiel soll eine univariate Information (z. B. ein Kompositorium aus unterschiedlichen Eignungstestinformationen) als Ausgangsbasis dienen. Die drei Arbeitsplätze erfordern jeweils spezifische Minimalqualifikation. Gleichzeitig weisen die drei zuzuordnenden Personen unterschiedliche Qualifikationen für die Arbeitsplätze auf.

Zunächst könnte der institutionelle Nutzen maximiert werden, indem die Stellen nur mit den Personen besetzt werden, die voraussichtlich die größte Leistung erzielen werden (*Klassifikation I*). In diesem Fall müßte Person A für alle Arbeitsplätze nominiert werden, wobei sie für Arbeitsplatz 2 und besonders 3 eindeutig überqualifiziert ist.

Eine andere Strategie wird mit *Klassifikation II* verfolgt. Hier wird individuumszentriert jeder Person der Arbeitsplatz zugewiesen, für den sie

ihre höchste Qualifikation aufweist. Person B und Person C besetzen zwar die Plätze, für die sie ihre höchste Qualifikation aufweisen, beide erfüllen aber nicht die Mindestqualifikation. Sie werden auf Dauer vermutlich überfordert sein.

Klassifikation III verkörpert eine sogenannte „cut and fit"-Strategie (vgl. Cascio, 1997): Alle Arbeitsplätze werden mit zumindest minimalqualifizierten Personen besetzt. Gleichzeitig werden die Personen nach Möglichkeit Arbeitsplätzen zugewiesen, die am ehesten ihrer Qualifikation entsprechen. Nachteil dieser Kompromißlösung ist, daß keine Person bestmöglich zugeordnet wird.

In der Praxis läuft diese Strategie auf eine sukzessive Selektionstechnik hinaus: Die Arbeitsplätze werden gemäß ihrer Wichtigkeit (= geforderte Mindestqualifikation) der Reihe nach bearbeitet. Für den wichtigsten Arbeitsplatz wird die qualifizierteste Person ausgewählt, die gleichzeitig die Mindestanforderungen erfüllt. Danach wird aus dem Pool der übriggebliebenen Personen die Person ausgewählt, die die Anforderungen des zweitwichtigsten Arbeitsplatzes am besten erfüllt usw. Sollten Arbeitsplätze mangels qualifizierter Personen vakant bleiben, kann entweder die bisherige Ordnung überarbeitet werden (einen wichtigeren Arbeitsplatz vakant lassen und dafür mehrere weniger wichtige Plätze besetzen), nichtqualifizierte Personen zugewiesen oder geeignete Personen nachnominiert werden.

Bei einer „cut and fit"-Klassifikationsstrategie werden alle Arbeitsplätze mit zumindest minimalqualifizierten Personen besetzt. Gleichzeitig werden die Personen nach Möglichkeit Arbeitsplätzen zugewiesen, die am ehesten ihrer Qualifikation entsprechen.

Im Spezialfall einer „Klassifikation mit Selektion" sinkt der Klassifikationsaufwand mit steigender Personenanzahl bei gleichbleibender Anzahl von zu besetzenden Arbeitsplätzen. Durch die große Zahl verfügbarer Personen kann voraussichtlich jedem Arbeitsplatz eine ausreichend qualifizierte Person zugewiesen werden.

Wichtiger werden Klassifikationsstrategien natürlich mit einer steigenden Anzahl von zu besetzenden Arbeitsplätzen. Die Effizienz (verstanden als Verhältnis zwischen dem Aufwand der Klassifikation und dem zu erwartenden Nutzen im Vergleich zu einer Zufallszuordnung) steigt mit der Heterogenität der Tätigkeitsanforderungsprofile und der Personenqualifikationen.

Eine intuitive „cut and fit"-Strategie (vgl. Kasten 7) wird mit steigender Anzahl von genutzten Informationen (vgl. Frage 5: multivariate Informationsbasis) und zu besetzenden Arbeitsplätzen wegen der zu großen Komplexität nur suboptimale Lösungen ergeben. Es gab deshalb schon früh Bemühungen, diese Problemsituationen über statistische Modellbildung zu lösen (z. B. Brodgen, 1959). Zusätzlich können auch Ansätze aus anderen Forschungsbereichen genutzt werden, z. B. aus der pädagogischen Psychologie oder der allgemeinen Diagnostik (vgl. beispielsweise Janke, 1982).

Es wurde schon früh versucht, über statistische Modelle eine „cut and fit"-Klassifikation zu optimieren.

Alley (1994) stellt ein allgemeines Rahmenmodell zu den aktuell verwendeten statistischen Klassifikationssystemen vor. Es ist in Abbildung 2 wiedergegeben.

Alley unterscheidet fünf Phasen:

- In der ersten Phase wird zunächst definiert, welche Messungen als Prädiktoren dienen, welche Berufserfolgskriterien berücksichtigt werden und welche Typen von Arbeitsplätzen unterschieden werden sollen. Vor diesem Hintergrund wird ein geeigneter Datensatz ausgewählt, der als Grundlage für alle weiteren Berechnungen dient.
- In der zweiten Phase werden für jeden Arbeitsplatztyp getrennt Parameter in typischerweise linearen Funktionen zur Vorhersage der Kriterien durch die Prädiktoren geschätzt.
- Diese Schätzungen werden in der dritten Phase verwendet, um Leistungserwartungen für alle Personen über alle Arbeitsplatztypen hinweg zu treffen. Resultat ist eine n × m-Matrix mit der spezifischen Leistungserwartung einer Person für einen bestimmten Arbeitsplatztyp in jeder Zelle.
- In der vierten Phase wird über einen geeigneten statistischen Algorithmus jeder Person ein geeigneter Arbeitsplatz(-typ) zugewiesen. Ziel ist hier eine Maximierung der zu erwartenden beruflichen Leistung. Welche statistischen Verfahren hierfür eingesetzt werden, ist beim jetzigen Stand der

Forschung weitgehend Ermessenssache. Johnson und Zeidner (Johnson & Zeidner, 1991; Scholarios, Johnson & Zeidner, 1994) greifen beispielsweise auf die von ihnen entwickelte „differential assignment theory“ zurück, eine Verknüpfung der älteren Ansätze von Brodgen (1959) und Horst (1955) mit einem Monte Carlo-Simulationsansatz. Gegebenenfalls muß in einem Zwischenschritt (Phase 4a) noch die ursprüngliche Unterteilung der Arbeitsplatzeinteilungen überarbeitet werden, um so die Klassifikationssysteme zu vereinfachen und/oder die maximierte Leistung zu steigern.

- In der letzten Phase 5 werden schließlich standardisierte Prädiktorkompositorien abgeleitet, die als Zuweisungskriterien für zukünftige Klassifikationsentscheidungen dienen.

Ein Rahmenmodell für statistische Klassifikationssysteme

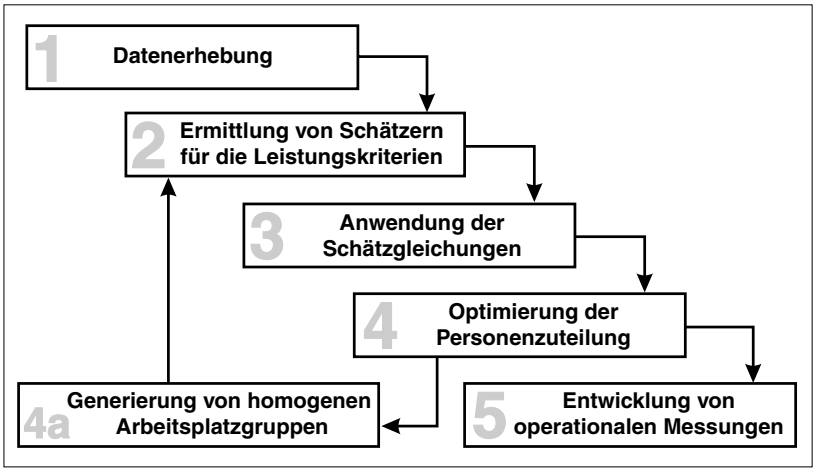


Abbildung 2:
Ein Rahmenmodell für statistische Klassifikationssysteme (nach Alley, 1994, p. 432)

Systematische Klassifikationsstrategien werden bevorzugt beim (amerikanischen) Militär angewendet.

Eine Durchsicht der neueren organisationspsychologischen Literatur zeigt ein auffallendes Desinteresse an der Thematik. So trägt beispielsweise ein Überblickskapitel von Guion (1991) im „Handbook of industrial and organizational psychology“ den vielversprechenden Titel „Personnel assessment, selection, and placement“. Guion verweist allerdings lapidar auf fehlende neuere Entwicklungen und stellt die Notwendigkeit von Klassifikationsstrategien in Abrede, da die meisten Organisationen kaum systematische Prozeduren über Personalselektion hinaus betreiben würden. Eine Ausnahme stellt auch für Guion der militärische Bereich dar, und nicht ohne Grund finden sich auch hier die (übriggebliebenen?) Verfechter von systematischen Klassifikationssystemen (vgl. mehrere Beiträge in Rumsey, Walker & Harris, 1994).

Die Darstellung von Zeidner und Johnson (1994) zeigt, daß sich die Klassifikationstheorie auch gegen inhaltstheoretische Vorbehalte wehren muß. Gestützt auf die metaanalytische Befundlage plädieren einige Wissenschaftler (z. B. Ree & Carretta, 1998) für die alleinige Verwendung von allgemeinen kognitiven Fähigkeitstests als Berufserfolgsvorhersagen (vgl. Kapitel 5). Damit würde sich der multiple Prädiktorsatz auf eine einzige Meßgröße reduzieren. Die resultierende Plazierungsproblematik wäre wesentlich leichter zu lösen als die komplexere Klassifikationssituation. Die Darstellung von Zeidner und Johnson (1994) zielt dementsprechend auf eine Entkräftung dieser Argumentationslinie ab (Nachweis der inkrementellen Validität von multiplen Messungen, Diskussion der statistischen Robustheit von Regressionsparametern usw.).

4 Evaluation von Personalentwicklungsmaßnahmen

Unter „Evaluation“ wird die Erfolgskontrolle von Interventionsprogrammen verstanden. Sie umfaßt die Bewertung des eingesetzten Konzepts, des Untersuchungsplans, der Implementierung und der Wirksamkeit.

Am Anfang des Kapitels (Abschnitt 2) wurde mit der „Validierung eignungsdiagnostischer Verfahren“ ein relativ engbegrenztes Forschungsgebiet zur Güteüberprüfung von psychologischen Testverfahren dargestellt. „Validierung“ kann als Spezialfall einer „Evaluation“ angesehen werden, hier ganz allgemein verstanden als „Erfolgskontrolle von Interventionsprogrammen“. Die Erfolgskontrolle umfaßt die Bewertung des eingesetzten Konzepts, des Untersuchungsplans, der Implementierung und der Wirksamkeit des untersuchten Interventionsprogramms.

Die prinzipiellen Schritte einer Evaluation und die auftretenden Grundprobleme sind relativ unabhängig vom konkreten Forschungsgegenstand. Sie sollen deshalb im folgenden kurz in allgemeiner Form skizziert werden, um einen Generalüberblick zur Materie zu geben. Danach soll die Evaluation personzentrierter Personalentwicklung (vgl. Kapitel 9 und 10) als spezielle Form sozialer Interventionsprogramme genauer behandelt werden. Hierfür wird ein Rahmenmodell zur Trainingseffektivität vorgestellt, in dem auf die Besonderheiten dieses Evaluationsgebiets ausführlich eingegangen wird. Zwei Evaluationsbeispiele konkretisieren die Ausführungen.

4.1 Grundelemente und -probleme einer Evaluation

In Kasten 8 sind in einer Zusammenschau die wesentlichen Arbeitsschritte einer Evaluation beschrieben (vgl. für ausführlichere Darstellungen z. B. Fink, 1995; Rossi & Freeman, 1993; Wottawa & Thierau, 1998).

Kasten 8:
Arbeitsschritte einer Evaluation

Die prinzipiellen Arbeitsschritte einer Evaluation sind relativ unabhängig vom konkreten Forschungsgegenstand.

1. Klärung der inhaltlichen Fragestellung des zu evaluierenden Programms
Zunächst ist eine genaue Erfassung des zu evaluierenden Interventionsprogramms notwendig. So müssen die Entscheidungsträger („Wer steuert das Programm?“) und die Zielgruppen („Wer soll von dem Programm profitieren?“) möglichst genau beschrieben werden. Das Programm selbst muß hinsichtlich seiner Struktur, der geplanten und tatsächlich ablaufenden Prozesse und der kurz-, mittel- und langfristig resultierenden Ergebnisse untersucht werden.
2. Ableitung der Evaluationsfragen
Danach müssen die angestrebten Ziele der Evaluation offengelegt und konkretisiert werden. Rossi und Freeman (1993) unterscheiden hierbei drei typische Einsatzbereiche für Evaluation: – In der Planungsphase der Konzepterstellung und Gestaltung der Intervention dient Evaluation *einer exakteren Erfassung* und gleichzeitig *eindeutigen Eingrenzung der zu behandelnden* Probleme und der genauen Zielsetzung des durchzuführenden Programms, z. B. hinsichtlich der Zielpopulation und der Interventionsausgestaltung. – Bei neu angelaufenen oder bereits längere Zeit bestehenden Programmen dient sie der *Effektivitätsprüfung einzelner Programmteilabläufe und -ergebnisse* sowie der *Überprüfung des Interventionseffekts.* Ein Schwerpunkt liegt hierbei auf der Erfassung des „program impact“ i.S. von Einfluß- und Effektbereich sowie Dauer der Wirkung verschiedener Variablen. – Eine dritte Anwendung kann in der *Feinjustierung bestehender Programme* liegen, bei der nur Subelemente zur Verbesserung der Effektivität und Effizienz modifiziert werden, die prinzipielle Zielsetzung aber nicht in Frage gestellt wird.

Als Ausgangspunkt für die Konkretisierung der Evaluationsfragen kann ein aus der Programmanalyse abgeleitetes, am aktuellen wissenschaftlichen Forschungsstand orientiertes theoretisches Modell (ein sog. „impact"-Modell) dienen, in dem möglichst alle relevanten Wirkgrößen auf das Programmgeschehen berücksichtigt werden.
Auf dieser Basis können dann Evaluationsziele abgeleitet werden, die
– die Programmzielsetzungen möglichst vollständig abbilden,
– gleichzeitig aber realistisch bleiben,
– meßbare Größen und Operationalisierungen zu ihrer Überprüfung angeben und
– hinsichtlich ihrer Bedeutsamkeit gewichtet sind.
Dies sollte unter Einbindung aller betroffenen Parteien erfolgen, um auf einer möglichst breiten Konsensbasis arbeiten zu können.

3. Planung der Evaluationsstudie

Durch verschiedene Zielsetzungen ergeben sich bei spezifischen Ausgestaltungen unterschiedliche Gewichtungen der Teilphasen.

Die nunmehr festgelegten Evaluationsziele müssen nachfolgend in ein Evaluationsmodell überführt werden. In ihm werden die konzeptionellen Eckpfeiler der Evaluation festgelegt und eine Entscheidung hinsichtlich des theoretischen Hintergrunds getroffen, vor dem die Evaluationsinhalte interpretiert werden sollen. Sollte sich z. B. auf dieser Grundlage ein nicht-empirisches Vorgehen (vgl. die Beispiele in Wottawa, 1996) genügen, ist die Datensammlung nach Erfassung dieser nicht-empirischen Informationen beendet und es kann mit der Vorbereitung der Berichterstattung (*Phase 5*) begonnen werden. Meistens stellt sich dies allerdings als noch nicht ausreichend heraus, und so schließen sich Überlegungen an, wie die fehlenden Informationen gewonnen werden können. Zur Erfassung steht prinzipiell das gesamte Methodenrepertoire der empirischen Sozialforschung zur Verfügung (vgl. Mohr, 1995, oder Bierhoff & Rudinger, 1996, für eine Übersicht zu möglichen quasi-experimentellen Versuchsplänen).

4. Datenerhebung und -auswertung

Über eine geeignete Stichprobenziehung muß gewährleistet sein, daß die in den Evaluationszielen (*Phase 2*) festgelegte Zielpopulation in allen Erhebungsbedingungen repräsentativ erfaßt wird. Über eine Gegenüberstellung der aufgestellten Ziele und der eingesetzten Methoden kann sichergestellt werden, daß jedes Evaluationsziel über ein oder mehrere Instrumente adäquat erfaßt wird. Ein strukturierter Auswertungsplan muß eine suffiziente Auswertung vom eingesetzten Tiefeninterview bis hin zum vorgelegten standardisierten Testverfahren gewährleisten.

5. Ergebnisrückmeldung und Berichterstattung

Prinzipiell sind zwei Arten von Rückmeldung der Evaluationsergebnisse denkbar:

Bei einer *formativen Evaluation* dienen die Rückmeldungen von Zwischenergebnissen als Grundlage für Modifikationen im laufenden Programm. Sie unterstützen somit eine Optimierung des Programms.

Bei einer *summativen Evaluation* wird eine Gesamtrückmeldung nach Beendigung des Programms gegeben, ohne daß Evaluationserkenntnisse Einfluß auf den Programmablauf hatten. Sie dient somit einer Qualitätsbeurteilung des Programms.

Die Berichterstattung muß zielgruppenspezifisch hinsichtlich des statistischen und inhaltlichen Vorwissens, des berichteten Inhalts und des Anspruchsgrads erfolgen. Über eine gedankliche Vorwegnahme möglicher Konsequenzen können Entscheidungshilfen, z. B. in Form von Verbesserungsvorschlägen zum Programm oder Hinweisen auf mögliche Nachfolgeuntersuchungen, gegeben werden.

6. Prozeßbegleitende Maßnahmen

Zu den begleitenden Maßnahmen gehören alle Aktivitäten des Projektmanagements, die nicht direkt mit den Evaluationszielen in Verbindung stehen, für einen erfolgreichen Ablauf aber zwingend notwendig sind. Hier-

zu zählt z. B. die Erstellung und Kontrolle eines Zeitplans, die Bestimmung des personellen und finanziellen Projektbedarfs, die Koordination zwischen den Evaluatoren und den betroffenen Programmträgern, die Überprüfung der vereinbarten Qualitätsstandards usw.

7. Meta-Evaluation zum Abschluß

Zum Abschluß der eigentlichen Evaluation sollte sich eine „Evaluation der Evaluation" (= Meta-Evaluation) anschließen, in der kritisch reflektiert wird, ob alle Bedürfnisse der Beteiligten durch die Evaluationsstudie und ihre Berichterstattung gedeckt wurden.
Dazu gehört auch eine kritische Reflexion des Evaluationsdesigns:
- Welche Programmbereiche konnten nur teilweise erfaßt, welche Evaluationsziele konnten nicht oder nur unzureichend erfüllt werden?
- In welchen Programmphasen kriselte das Evaluationsmanagement?
- Wie hätten alternative Evaluationsdesigns aussehen können?
- Wurden entsprechende Revisionsmöglichkeiten im Programm vorgesehen?
- Ergeben sich sinnvolle Ansatzpunkte für anschließende Evaluationsprojekte? usw.

Zum Abschluß dieses allgemeinen Evaluationsüberblicks sollen drei Problem- bzw. Diskussionsfelder angesprochen werden, die bei jeder praktischen Umsetzung relevant sind.

Unterschiedliche Evaluationsparadigmen

Die bestehenden Evaluationsansätze können grob nach qualitativer und quantitativer Orientierung unterschieden werden.

In der einschlägigen Evaluationsliteratur wird der interdisziplinäre Charakter der Evaluationsarbeit stark betont (vgl. Rossi & Freeman, 1993). Dies impliziert neben der gewollten Bereicherung der Arbeit durch unterschiedliche Abstraktionsebenen und Schwerpunktsetzungen bei der Analyse und Interpretation aber auch ein Aufeinandertreffen unterschiedlicher Ziele und Beurteilungskriterien. Antoni (1993) unterscheidet z. B. unter Rückgriff auf Cook und Reichardt (1979) generell zwischen einer qualitativen und einer quantitativen Orientierung:

Beim qualitativen Ansatz liegt mit der Verwendung unstandardisierter Verfahren wie z. B. Tiefeninterviews oder Gruppendiskussionen der Schwerpunkt auf einer unmittelbaren und vielschichtigen Datenerfassung. Der Evaluator versteht sich hier als Subjekt der Situation und ist durch die verwendeten Methoden (z. B. teilnehmende Beobachtung) in das Geschehen z.T. aktiv involviert. Dadurch soll die Einzigartigkeit der Situation nachvollzogen werden, allerdings unter bewußtem Verzicht auf Generalisierbarkeit. Da Einflußgrößen nicht kontrolliert werden, fehlt ein wesentlicher Bestimmungspunkt für Kausalinterpretationen.

Der quantitative Ansatz favorisiert den Einsatz fundierter quantitativer Methoden, wie z. B. testtheoretisch konstruierte Fragebogen oder psychophysiologische Meßverfahren. Der Schwerpunkt liegt hier auf der möglichst reliablen, replizierbaren Erfassung von Fakten. Der Evaluator soll ein neutraler, möglichst objektiv agierender Außenseiter sein, der über eine hypothetisch-deduktive Herangehensweise Kausalannahmen überprüfen möchte. Die abgeleiteten Schlußfolgerungen zielen auf eine Generalisierbarkeit der Ergebnisse über die spezifische Situation hinaus. Hauptkritikpunkt an diesem Ansatz ist hier die Tendenz zum Reduktionismus, der über der Detailanalyse das Ganze zu vernachlässigen droht.

Die Indikation für einen quantitativen oder qualitativen Ansatz ergibt sich aus dem Konkretisierungsgrad der Evaluationsziele.

Die Indikation für einen qualitativen und gegen einen quantitativen Ansatz (oder umgekehrt) ergibt sich aus dem Konkretisierungsgrad der Evaluationsziele: Wenn diese noch gar nicht feststehen oder noch sehr unkonkret sind, wenn viele Einflußvariablen erfaßt werden sollen und deren Wirkzusammenhang unbekannt ist, dann ist der qualitative, nicht-experimentelle Ansatz vorzuziehen. Andererseits erscheinen „...quantitative experimentelle Methoden um so eher geeignet ..., je enger die Fragestellung auf die Erfassung von einigen wenigen Variablen und Kausalbeziehungen ausgerichtet ist, je mehr es um die Bestätigung von bereits relativ sicher vermuteten Interventionsme-

chanismen bzw. -Effekten geht und je stabiler und unwichtiger die Rahmenbedingungen sind und je angemessener eine kontrollierte Manipulation der Maßnahmen ist“ (Antoni, 1993, S. 333).

Denkbar ist natürlich auch ein sequentielles Vorgehen, bei dem zunächst qualitative Verfahren eingesetzt werden, um die Evaluationsziele hinreichend für quantitative Methoden zu konkretisieren.

Parteieninteressen bei der Evaluation

In der Praxis ist eine Evaluation in jeder konzeptionellen Phase den zum Teil konträren Interessen der Beteiligten ausgesetzt.

Aus der bisherigen Darstellung könnte der Schluß gezogen werden, eine Programmevaluation sei eine von allen Parteien gleichermaßen erwünschte wie unterstützte Aktion. In der Praxis findet sie aber nicht im neutralen Raum statt, sondern ist in jeder konzeptionellen Phase den z.T. konträren Interessen der Beteiligten ausgesetzt.

Bereits die Initiierung von Evaluation ist davon betroffen. So führen Thierau, Stangel-Meseke und Wottawa (1999) sowie Nork (1991) für den Personalentwicklungsbereich zahlreiche Gründe für mangelnde Evaluationsaktivität der Unternehmensleitung, des Evaluators, der Trainer bzw. Dozenten und der Teilnehmer an, die von mangelndem Interesse über fehlende Fachkenntnisse bis hin zu Bewertungsangst reichen. Eine laufende Evaluation muß sich ebenso mit dem bestehenden politischen Interessengefüge auseinandersetzen. Zwar soll z. B. die Ableitung von Evaluationszielen unter Mitwirkung aller Involvierten (hier: Unternehmensleitung, Durchführende, Teilnehmer) erfolgen, viele der Zielsetzungen sind aber an bestimmte Interessengruppen gebunden, stehen sich konträr gegenüber und werden teilweise gar nicht offen artikuliert. Möglicherweise entsteht am Ende eine allgemeine und vage Zielvereinbarung als Kompromißlösung, dem Evaluator fehlt dann aber eine wichtige Arbeitsgrundlage.

Weitere mögliche „politische Einflußnahmen auf Evaluation“ (vgl. Antoni, 1993) sind z. B. die Vergabe von Evaluationsstudien zur bloßen Entscheidungsverzögerung, die Evaluationssteuerung hinsichtlich einseitig positiver oder negativer Programmeffekte, der Versuch der Delegation der Entscheidungsverantwortung vom Auftraggeber hin zum Evaluator oder eine rein symbolhafte Evaluationsanwendung ohne praktische Relevanz. Wenn ein Evaluator solche Einflußnahmen auch niemals ausschalten kann, so muß er zumindest darauf vorbereitet sein und die möglichen Konsequenzen für die eigene Arbeit antizipieren können.

Wissenschaftlichkeit von Evaluation

Evaluationen erfolgen zweckorientiert und dienen meistens keiner wissenschaftlichen Hypothesenprüfung im Sinne der Grundlagenforschung.

Evaluationsforschung ist nicht gleichzusetzen mit „normaler“ wissenschaftlicher Arbeit. Bortz und Döring (1995) greifen auf Herrmanns (1976) Unterscheidung zwischen wissenschaftlichen und technologischen Theorien zurück, um dies zu erläutern:

Wissenschaftliche Theorien dienen danach zur Beschreibung, Erklärung und Vorhersage von Sachverhalten. Hier ist besonders die Grundlagenforschung gefragt, da sie zunächst nicht nach Nutzen oder Anwendungsmöglichkeiten ihres Wissens fragt, sondern primär auf die Sammlung von Hintergrundwissen ohne funktionale Zielsetzung abzielt. Technologische Theorien geben hingegen konkretes Handlungswissen zur praktischen Umsetzung der wissenschaftlichen Theorien. Sie finden Anwendung in der Angewandten, Interventions- und Evaluationsforschung. Während die Angewandte Forschung allgemein die Nutzung von Grundlagenerkenntnissen für praxisrelevante Problemfelder anstrebt, dient die Interventionsforschung direkt zur Ableitung konkreter Maßnahmen. Evaluationsforschung kann nun wiederum als Bewertung der Intervention verstanden werden.

Daraus ergibt sich, daß von Evaluationsprojekten eine konkrete Aussage zu einem vorliegenden Sachverhalt verlangt wird (Wottawa, 1996). Sie sind in übergreifende Entscheidungsprozesse eingebunden, und alle während der Evaluation erhobenen Variablen sollen einen zielgerichteten Entscheidungsbezug aufweisen: Evaluationen erfolgen zweckorientiert und dienen meistens keiner wissenschaftlichen Hypothesenüberprüfung im Sinne der Grundlagenforschung. Die Ergebnisse einer Evaluationsstudie sind oft sin-

gulär, da sich zumeist nicht die Möglichkeit einer Replikation ergibt. Die Praxis setzt auch der in der Grundlagenforschung üblichen Methodik schnell Grenzen: So wird eine experimentelle Versuchsplanung aus technischen Gründen schwierig (wie kann bei vorgegebenen Arbeitsgruppen eine Randomisierung erfolgen?) und teilweise ethisch problematisch sein (z. B. die Einführung einer unbehandelten Kontrollgruppe). Ein Doppelblind-Ansatz (weder Evaluationsleiter noch Teilnehmer kennen die Untersuchungshypothesen) wird nicht anzustreben sein, da eine Evaluation über eine möglichst umfassende Offenlegung der Arbeiten eine möglichst große Akzeptanz und Kooperation erzielen muß. Viele wissenschaftliche Grundkonzepte sind also nur begrenzt umsetzbar. Dies sollte allerdings nicht dazu führen, daß die Komplexität der Praxissituation, die ein Abwägen des sinnvollen und machbaren Vorgehens notwendig macht, als Begründung für methodisch schlechte oder lückenhafte Evaluationen mißbraucht wird.

Die Komplexität der Praxissituation, die ein Abwägen des sinnvollen und machbaren Vorgehens notwendig macht, darf aber nicht als Begründung für methodisch mangelhafte Evaluationen dienen.

Bei der Anwendung auf die Evaluation von personzentrierten Personalentwicklungsverfahren kann das allgemeine Evaluationsvorgehen ergänzt werden um ein relativ ausführliches Wissen zu Effektivitätsbedingungen von Trainings. Darauf soll im weiteren eingegangen werden.

4.2 Einflußvariablen der Trainingseffektivität

Zunächst sollen zwei relativ intensiv untersuchte Forschungsgebiete bei Trainingseffektivitätsstudien behandelt werden: unterschiedliche Wirkungsebenen von Trainings sowie relevante Randbedingungen des Transfers von Trainingsinhalten. Danach erfolgt eine Integration der Befunde in einem Modell zu allgemeinen Effektivitätsbedingungen von Trainings.

Inhalte und Ebenen von Trainingsergebnissen

Zur Klärung der inhaltlichen Fragestellung des zu evaluierenden Programms müssen zunächst die Trainingsziele erfaßt werden: Was sollen die Teilnehmer durch das Training lernen, welchen Professionalitätsgrad sollen sie dabei erreichen und unter welchen Bedingungen sollen sie das gelernte Wissen anwenden? In der prominentesten Taxonomie hierzu unterscheidet Kirkpatrick (z. B. Kirkpatrick, 1987) vier Ebenen: Reaktion, Lernen, Verhalten und Ergebnisse.

Kirkpatrick (1987) unterscheidet zwischen den vier Ebenen *Reaktion, Lernen, Verhalten* und *Ergebnis*, in denen sich Konsequenzen des Trainings zeigen können.

- *Reaktion* beschreibt die Zufriedenheit der Trainingsteilnehmer, ihre emotionalen Wertungen („Das Training war gut“) und subjektiven Nutzenabwägungen („Das Training hat mir etwas gebracht“). Diese Informationen werden zumeist über Fragebogen oder Interviews direkt nach dem absolvierten Training erhoben.
- *Lernen* umschreibt die Gruppe der Kriterien, die das Verständnis und die Aneignung der Trainingsinhalte durch die Teilnehmer erfassen sollen. Kraiger, Ford und Salas (1993) unterscheiden hierbei kognitive Ergebnisse (z. B. verbessertes fachliches Wissen, bessere Wissensorganisation, Anwendung adäquaterer kognitiver Strategien), verbesserte Fertigkeiten (z. B. schneller und fehlerfreier Arbeitsablauf, erhöhter Verselbständigungsgrad von Arbeitsabläufen) und affektive Änderungen (Einstellungen, arbeitsbezogene Motivation u.ä.). Für die Erfassung dieser Kriterien bieten sich entsprechende standardisierte Testverfahren bzw. Selbstauskünfte an.
- Über Kriterien der *Verhaltensebene* wird die Umsetzung des Gelernten am Arbeitsplatz überprüft. Es wird erhoben, inwieweit die Teilnehmer vom konkreten und idealtypischen Trainingsinhalt auf nur prinzipiell ähnliche und in einen größeren Zusammenhang eingebettete Problemfälle generalisieren können.
- Die *Ergebnisse* werden auf der abstraktesten Ebene durch globale organisationale Leistungskriterien, wie betriebliche Produktivitätskennziffern oder Bilanzzahlen, beschrieben.

In einem Literaturüberblick diskutieren Alliger und Janak (1989) mehrere Vorannahmen zu den Kirkpatrick-Kriterien, die bei der Verwendung des Modells häufig anzutreffen sind:

- Die Annahme, daß jede nachfolgende Kriterienebene informativer sei als die vorhergehenden, spiegelt eine eingeengte Werthaltung wider. Qualitative und individuelle Informationen werden gegenüber ökonomischen Geldkriterien geringgeschätzt. Zudem werden dabei spezifische Zielsetzungen von Trainings (z. B. einfache Information der Teilnehmer ohne gezielte Verhaltensänderung) übergangen.
- Die Annahme, jede nachfolgende Ebene sei kausal durch die vorhergehenden beeinflußt, läßt sich zumindest für die Reaktionsebene nicht bestätigen: Ein Training kann zwar unterhaltsam und humorig gestaltet sein, muß aber nicht gleichzeitig einen Lerneffekt aufweisen. Im Gegensatz dazu kann ein Training von den Teilnehmern als langweilig und wenig unterhaltsam eingestuft werden, gleichzeitig aber doch lehrreich sein und verhaltensändernd wirken.
- Die alternative Annahme, alle Ebenen seien positiv miteinander korreliert, konnte nur ansatzweise bestätigt werden. Alliger und Janak (1989) fanden in ihrem 30-Jahres-Literaturüberblick (von der Erstveröffentlichung der Kriterien 1959 bis 1989) nur relativ geringe positive Korrelationen zwischen den letzten drei Ebenen (Lernen, Verhalten und Ergebnisse) von r=.18 bis r=.40.

Bei der Erfassung der unterschiedlichen Kriterienebenen scheint es einen Bruch zwischen Praxis und Wissenschaft zu geben: Während in der Praxis wegen der einfachen Zugänglichkeit häufig nur Reaktionsmaße erhoben werden, finden sich durch das vorgeschaltete Reviewverfahren in wissenschaftlichen Publikation eher Maße der übrigen Ebenen, meistens aber nicht mehr als zwei gleichzeitig.

Randbedingungen des Transfers von Trainingsinhalten

Einflüsse auf den Transfer von Trainingsinhalten lassen sich unterteilen in Effekte des Trainingsdesigns, der Teilnehmermerkmale sowie von Variablen der Arbeitsumgebung und der organisationalen Bedingungen.

Der Transfer von Trainingsinhalten ist von verschiedenen Bedingungen abhängig. Baldwin und Ford (1988; eine Überarbeitung findet sich bei Noe & Ford, 1992) haben in ihrer Literaturübersicht drei verschiedene solcher Einflußgrößen zusammengefaßt:

Die Gruppe *Trainingsdesign* beinhaltet alle Bedingungen, bei denen zur Verbesserung der Trainingsgestaltung auf gesicherte Lernprinzipien zurückgegriffen wurde, z. B. die exakte Simulation des späteren Anwendungsbereichs im Training oder zeitlich verteiltes statt massiertes Lernen. In der Gruppe *Teilnehmermerkmale* werden alle Untersuchungsbefunde zu relevanten Fähigkeiten, Persönlichkeitseigenschaften und Motivationsaspekten zusammengefaßt. Die dritte Gruppe umfaßt *Variablen der Arbeitsumgebung und organisationale Bedingungen,* z. B. Unterstützung durch den Vorgesetzten, Anwendungsmöglichkeiten des Trainingsinhalts oder Einfluß von Unternehmenskulturvariablen.

Eine umfassendere Diskussion dieser Einflußbedingungen findet sich in Hesketh (1997; vgl. auch die Kommentierungen im gleichen Band).

Ein Modell zu allgemeinen Effektivitätsbedingungen von Trainings

Das Tannenbaum-Modell zur Trainingseffektivität integriert bisherige Ansätze und gibt einen Generalüberblick zu möglichen Einflüssen.

Tannenbaum und Mitarbeiter (z. B. Cannon-Bowers, Salas, Tannenbaum & Mathieu, 1995; Tannenbaum, Mathieu, Salas & Cannon-Bowers, 1991) haben ein allgemeines Modell zu Effektivitätsbedingungen von Trainings entwickelt, das die älteren Befunde miteinander verbindet. Es ist in Abbildung 3 dargestellt.

Die Kirkpatrick-Kriterienebenen finden sich hier im zentralen, fett markierten Bereich wieder. Tannenbaum et al. unterscheiden bei ihrer Darstellung im Gegensatz zu Kirkpatrick (1987) zwei Verhaltensbereiche: Mit seiner *Trainingsleistung* zeigt der Teilnehmer unmittelbar, daß er die Trainingsinhalte beherrscht. In der späteren *Arbeitsleistung* zeigt sich aber erst, ob er in der Lage ist, das Wissen unter geänderten Bedingungen und unter Einfluß unterschiedlicher Situationsvariablen anzuwenden. Bei der Modellierung der Zusammenhänge zwischen den Kriterienebenen folgen Tannenbaum et al. den einschlägigen Ergebnissen von Alliger und Janak (1989): Die Reaktionen der Teilnehmer stehen unverbunden neben der Ler-

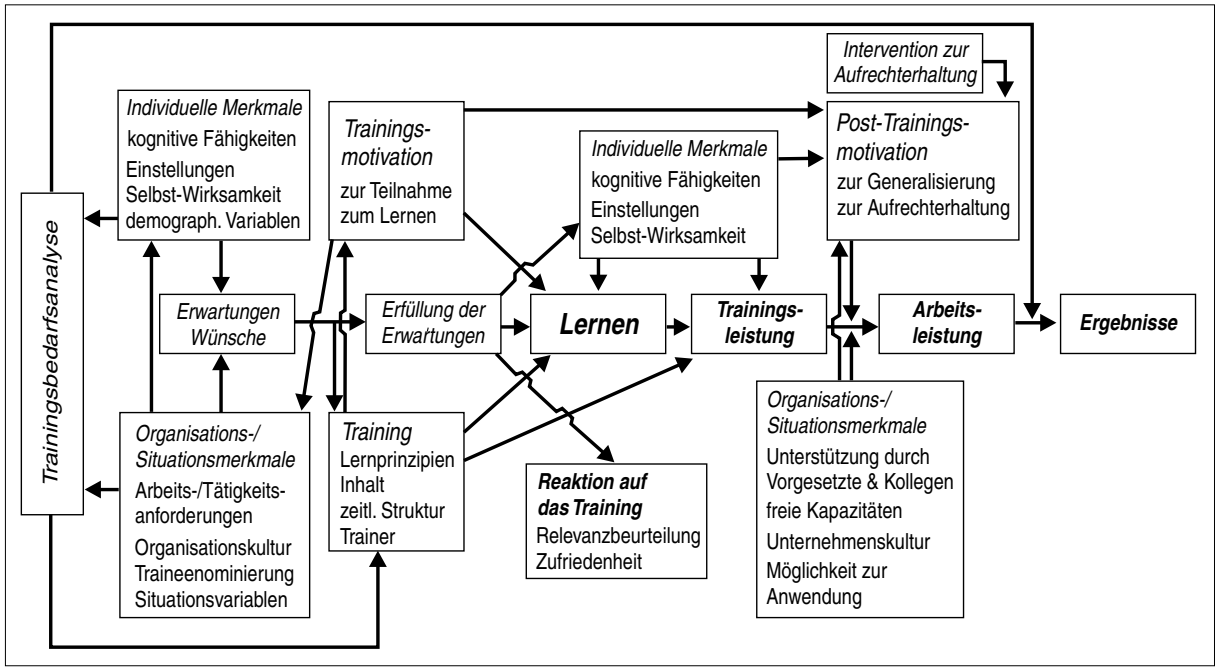


Abbildung 3:
Das Tannenbaum-Rahmenmodell zur Trainingseffektivität

nen-Verhalten-Ergebnis-Kette. Nach neueren metaanalytischen Befunden (Alliger, Tannenbaum, Bennett, Traver & Shotland, 1997) muß hierbei allerdings auch der Abstraktionsgrad der Fragen berücksichtigt werden. So zeigen inhaltlich eingegrenzte Fragen zu subjektiven Nutzenbewertungen höhere Zusammenhänge (korrigiertes r=.18) mit Maßen der Arbeitsleistung als globale Bewertungen (korrigiertes r=.07).

Die von Baldwin und Ford (1988) eingeführten Randbedingungen sind im Tannenbaum-Modell weitergehend umschrieben und hinsichtlich ihrer Wirkung vor und nach dem Training differenziert: Im Vorfeld des Trainings bilden sich auf seiten des Individuums Erwartungen und Wünsche, deren Erfüllung zusammen mit den eigentlichen Merkmalen des Trainings maßgeblich für die weiteren Transferbedingungen ist. Die Trainingselemente haben direkten Einfluß nur auf das Lernen und die im Training gezeigte Leistung, während die individuellen und Organisations- und Situationsmerkmale sowohl im Vorfeld auf die Bildung der Erwartungen und Wünsche wirken als auch über ihren Einfluß auf die Teilnehmermotivation alle Trainingstransferebenen beeinflussen.

Das Tannenbaum-Modell liefert eine umfassende Beschreibung zu Trainingseffekten und beeinflussenden Variablen und kann somit Hilfestellung zur exakteren Definition von Evaluationszielen geben:

- Zum „Trainingserfolg“ können mehrere verschiedene Ebenen von Erfolgskriterien unterschieden werden, die jeweils spezifische Aspekte von „Erfolg“ abbilden. Eine Reaktionsmessung („Hat Ihnen das Training etwas gebracht?“) ist nur ein denkbares Kriterium mit begrenztem Informationswert.
- Neben dem eigentlich zu evaluierenden Training wirkt sich noch eine Vielzahl anderer Variablen auf den Erfolg des Trainings aus. Wenn eine umfassende Bedingungsanalyse angestrebt wird, müssen sie ausdrücklich berücksichtigt und ggf. in das Evaluationsdesign eingebunden werden. Anders ist beispielsweise kein erfolgreicher interorganisationaler Transfer von Trainingsverfahren zu gewährleisten (vgl. z. B. Goldstein, 1993).

Das Tannenbaum-Modell liefert eine umfassende Beschreibung des Trainingsgeschehens und kann somit Hilfestellung zur exakteren Definition von Evaluationszielen geben.

4.3 Zwei Beispiele für PE-Evaluation

Nachfolgend sollen anhand von zwei Beispielen mögliche Evaluationsumsetzungen dargestellt werden. Die beiden Studien sind dabei bewußt unterschiedlich hinsichtlich ihrer Zielsetzung ausgewählt:

Die Studie von Thierau (1991) ist als formative Evaluation zur Analyse und Verbesserung einer Weiterbildungsmaßnahme für Führungskräfte angelegt. Die Studie von Latham und Frayne (1989) ist eine summative Evaluation eines Selbstmanagement-Trainings zur Reduktion betrieblicher Absentismusraten.

Beispiel für eine formative Evaluation

Beispiel 1: Thieraus (1991) Evaluation einer Weiterbildungsmaßnahme zur Mitarbeiterbeurteilung

Ausgangspunkt der Evaluationsstudie von Thierau (1991; Thierau et al., 1999) war der Wunsch eines Unternehmens der Versicherungsbranche, eine Weiterbildungsmaßnahme für Führungskräfte (Schulung zu einem betrieblichen Personalbeurteilungssystem, kurz PBS genannt) formativ zu evaluieren. Die Evaluation sollte ein Stärken/Schwächen-Profil des bestehenden Verfahrens liefern und konkrete Hinweise für Verbesserungsmöglichkeiten geben.

Partizipativ mit den Betroffenen wurden in einem Workshop zuerst die Evaluationsziele zu den PBS-Seminaren und zum PBS-Verfahren selbst aufgestellt:

- Überprüfung der Umsetzung der Weiterbildungsziele im Seminar;
- Subjektive Bewertung des Seminars durch die Teilnehmer;
- Subjektive Einschätzung des PBS durch die Führungskräfte;
- Überprüfung der Umsetzung bzw. Einführung des PBS durch die Führungskräfte.

Als Versuchspersonen der Experimentalgruppe dienten 74 Vorgesetzte, die das Beurteilerseminar als Teilnehmer absolvierten. Als Kontrollgruppe fungierten 15 Vorgesetzte des Unternehmens, die nicht am Seminar teilgenommen hatten, sowie 20 Bochumer Studenten. In Tabelle 1 sind die eingesetzten Verfahren getrennt nach den Evaluationszielen zusammengefaßt.

In der Berichterstattung wurden von den Evaluatoren u.a. folgende Ergebnisse an die Beteiligten rückgemeldet:

- Der Kenntnisstand der Führungskräfte zum Beurteilungssystem war vermutlich aufgrund einer unklaren Informationsvermittlung während der Schulungen noch nicht ausreichend. Das eingesetzte Entscheidungslabyrinth zur Kenntnisüberprüfung trennte dabei nicht zwischen Experimental- und Kontrollgruppe.

Tabelle 1:
Eingesetzte Verfahren in der Thierau (1991)-Studie

Kriterienebene	Evaluationsfragen	Instrument
Reaktion	Wie wird die Weiterbildungsmaßnahme bewertet? Ergeben sich Veränderungen über die Zeit?	Seminarbeurteilungsbogen (nach dem Training und in der Nachbefragung)
Reaktion	Wie sieht die Einstellung zu den Weiterbildungsmaßnahmen aus? Gibt es Veränderungen im Vergleich vorher/nachher?	Bedarfsanalyse (vor dem Training) Seminarbeurteilungsbogen
Lernen	Wie gut sind die Kenntnisse zum PBS-Verfahren?	Wissenstest
Verhalten	Wie sehen verwendete Kommunikationsstile aus?	Entscheidungslabyrinth
Verhalten	Wie ist der Stand der Einführung? Welche Hindernisse & Probleme gibt es?	Nachbefragung (3 Monate nach dem Training)

- Die einzelnen Bestandteile der Weiterbildungsmaßnahme (Workshop- und Trainingsteil) wurden durch die Teilnehmer unterschiedlich bewertet: Das Training schnitt generell besser ab.
- Trotz allgemein positiver Einstellung zur Personalbeurteilung wurden durch die Führungskräfte große Probleme bei der konkreten Anwendung des Beurteilungsverfahrens gesehen.
- Wegen vielfacher Gründe (z. B. Zeitmangel, Schwierigkeiten bei der Umsetzung der Schulungsinhalte, mangelnde Unterstützung seitens der Personalabteilung) befanden sich auch drei Monate nach der Weiterbildungsmaßnahme die meisten Führungskräfte noch in der Planungsphase zur Einführung des Beurteilungssystems.

Der formative Charakter der Evaluation ist eindeutig erkennbar, da die gefundenen Ergebnisse erst den Startpunkt weiterer Modifikationen der Schulungsmaßnahme bedeuten. Bedingt durch die ausgehandelten Zielsetzungen der Evaluation stützt sie sich in der Hauptsache auf Reaktionskriterien, wobei die Transfereffekte auf die anderen Kriteriumsebenen weitgehend offenbleiben müssen. Die Ergebnisse deuten an, daß für eine Verbesserung des Trainingserfolgs eine weitergehende Analyse der Rahmenbedingungen stattfinden muß, um z. B. Interventionen zur Einbindung der Trainingsinhalte vorzunehmen.

Beispiel für eine summative Evaluation

Beispiel 2: Latham undFraynes (1989) Evaluation eines Selbstmanagement-Trainings

Die Studie von Latham und Frayne (1989) ist als summative Evaluation angelegt. Die Effektivität des Trainings wird vorrangig im Hinblick auf mögliche Schlußfolgerungen für ähnlich angelegte Trainings untersucht.

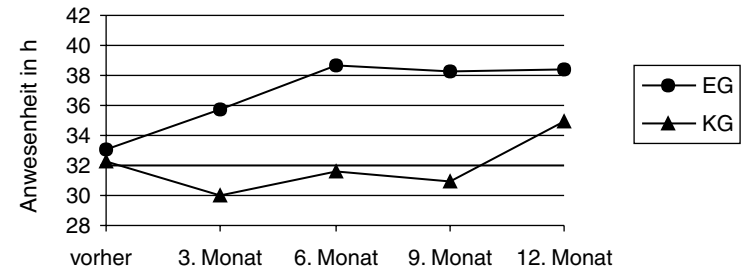
In der Studie nahmen zunächst 20 zufällig aus einer Gruppe von 40 Freiwilligen ausgewählte Angestellte einer Bundesbehörde an einem Gruppentraining teil, in dem orientiert am Selbstmanagement-Ansatz von Kanfer, Reinecker & Schmelzer (2000) Selbstkontrolltechniken für den Arbeitsplatz gelehrt wurden. Inhalte waren dabei z. B. Problemanalyse und Grundzüge der Zielsetzungsmethode, Selbstbeobachtungstrainings, praktische Übungen und eine Diskussion aufgetretener Probleme in der Gruppe. Das Training erstreckte sich über acht wöchentliche Sitzungen von jeweils einer Stunde Dauer, begleitend wurden ebenfalls wöchentlich 30minütige Einzelgespräche zu individuellen Problemen geführt.

Kriterien aus dem Reaktions-, dem Lern-, dem kognitiven und dem Verhaltensbereich wurden drei, sechs, neun und zwölf Monate nach Trainingsbeginn erhoben. In Tabelle 2 sind sie genauer beschrieben. Parallel wurde eine Kriterienmessung bei einer unbehandelten Kontrollgruppe aus derselben Behörde (die übrigen 20 Freiwilligen) vorgenommen. Neun Monate nach Programmbeginn unterzogen sich diese ebenfalls dem Programm, so daß die Kriterienmessung der ersten Gruppe nach 12 Monaten mit der ersten Kriterienmessung der zweiten Gruppe nach drei Monaten zusammenfiel.

Tabelle 2:
Erhobene Kriterien in der Latham und Frayne (1989)-Studie

Kriterienebene	Evaluationsfragen	Instrument
Reaktion	Wie hoch ist die Zufriedenheit der Teilnehmer mit dem Trainingsprogramm? Ändert sich die Zufriedenheit über die Zeit?	5-Item-Skala
Lernen	Haben die Teilnehmer die Trainingsinhalte gelernt, und behalten sie sie über die Zeit?	Wissenstest mit 12 situativen Fragen
Kognition	Führt das Training zu einer dauerhaften Erhöhung der wahrgenommenen Selbstwirksamkeit?	15-Item-Selbstwirksamkeitsskala 15-Item-Skala zu Ergebniserwartungen
Verhalten	Erhöht sich die Anwesenheitszeit am Arbeitsplatz, und ändern sich die Gründe der Abwesenheitszeiten?	Stechuhrkontrolle vertrauliche Interviews

Bei den Ergebnissen zeigte sich eine gleichbleibend hohe Zufriedenheit mit dem Training und ein konstanter Wissensstand auch ein Jahr später. Bei den kognitiven und den Verhaltensmaßen ließ sich eine verzögerte positive Wirkung nach sechs Monaten verzeichnen (vgl. Abbildung 4 für den beispielhaften Verlauf der Anwesenheitszeiten).



Angegeben ist die durchschnittliche Anwesenheit in einer 40-Stunden-Arbeitswoche. Die 12. Monat-Messung der Experimentalgruppe (EG) fällt zusammen mit der 3. Monat-Messung der Kontrollgruppe (KG).

Abbildung 4:
Verlauf der Anwesenheitszeiten (nach Latham & Frayne, 1989)

Die Studie von Latham und Frayne (1989) ist sicherlich vorbildlich hinsichtlich der Erfassung unterschiedlicher Kriterien über mehrere Meßzeitpunkte hinweg: Die Fragebogenverfahren bestehen zwar nur aus relativ wenig Items, können aber mit akzeptablen Reliabilitäten zwischen .8 und .9 aufwarten. Durch die Hinzuziehung einer Kontrollgruppe, die in einem späteren Stadium ebenfalls das Training mit einem anderen Leiter absolvierte, konnte die Generalisierbarkeit der Ergebnisse über den Zeitpunkt und den spezifischen Trainer hinaus nachgewiesen werden.

Die Stichprobengröße mit insgesamt 40 Personen ist allerdings gering, außerdem wurden diese freiwillig rekrutiert, so daß eine Selektivität hinsichtlich des Zielkriteriums möglich ist. Größter Nachteil ist die unzureichende Erfassung zusätzlicher Randvariablen, die vermutlich ebenfalls Einfluß auf die Abwesenheitszeiten haben: So ist durchaus eine „organisations-spezifische Abwesenheitskultur“ denkbar (vgl. z. B. Nicholson & Johns, 1985), die normativen Einfluß auf die einzelnen im Vorfeld und auch nach dem Training hat und eine Generalisierung der Ergebnisse über die Organisation hinaus behindern kann.

5 Ökonomische Nutzenanalyse

Zur Einschätzung der Praktikabilität von Verfahren treten neben die Validität auch Aspekte wie die Kosten der Verfahrenskonstruktion oder -auswahl, der erforderliche Aufwand bei der Durchführung und der zu erwartende Nutzen durch die Anwendung.

Für den Einsatz in der Praxis reicht der Validitätsnachweis eines Personalauswahlverfahrens oder die nachgewiesene Effektivität einer Personalentwicklungsmaßnahme meistens nicht aus. Zur Einschätzung der Praktikabilität von Verfahren treten neben die Validität auch Aspekte wie die Kosten der Verfahrenskonstruktion oder -auswahl, der erforderliche organisatorische Aufwand bei der Durchführung und der zu erwartende Nutzen durch die Anwendung. Im weiteren soll der Ansatz der ökonomischen Nutzenanalyse vorgestellt werden, bei dem über den Weg einer monetären Nutzenprognose („Wie groß ist voraussichtlich der durch den Einsatz des Verfahrens erwirtschaftete Mehrwert?“) eine Entscheidungshilfe für oder gegen den Einsatz eines Verfahrens gegeben werden soll. Zunächst werden wieder die allgemeinen Grundlagen behandelt. Danach werden zwei prototypische Modelle vorgestellt und mögliche Weiterentwicklungen angesprochen.

5.1 Grundlagen von Nutzenanalyse-Modellen

Nutzenanalysen können als Spezialfall von „multiattribute utility (MAU)"-Modellen der Entscheidungstheorie angesehen werden.

Nach Boudreau (1991) können Nutzenanalyse-Modelle als Spezialfall von sog. „multiattribute utility (MAU)"-Modellen der Entscheidungstheorie angesehen werden. Alle MAU-Modelle basieren auf *vier Grundelementen:*

1. Es werden verschiedene *Entscheidungsalternativen* („decision options") verglichen.
   In der Personalauswahl würden z. B. verschiedene eignungsdiagnostische Verfahren einander gegenübergestellt.
2. Die Entscheidungsalternativen sind durch eine Reihe von *Alternativenmerkmalen* („set of attributes") charakterisiert.
   Bei Personalauswahlverfahren können z. B. Validität, Kosten bei der Konstruktion und der Durchführung, Akzeptanz bei den Durchführenden usw. als relevante Merkmale unterschieden werden.
3. Es wird eine *Nutzenskala* („utility scale") definiert, die die Ausprägung jedes Attributs jeder Entscheidungsalternative angibt.
   Die Validität eines Verfahrens wird üblicherweise über einen entsprechenden Koeffizienten angegeben, die anfallenden Kosten in Geldwerten. Die Akzeptanz kann z. B. über ein testtheoretisch konstruiertes Verfahren erhoben und in Form eines „Zufriedenheitsscores" angegeben werden.
4. Die Eigenschaftsausprägungen der Entscheidungsalternativen werden in einer *Payoff-Funktion* miteinander zu einem allgemeinen Nutzenwert („overall utility value") kombiniert. Die Gewichtung der Entscheidungsalternativen und die gewählte mathematische Funktion hängen dabei von den subjektiven Gewichtungspräferenzen des Entscheidungsträgers ab. Wenn eine Firma z. B. die Akzeptanz eines Verfahrens besonders hoch gewichtet, sollte sich dies in der gewählten Payoff-Funktion in Form einer hohen Gewichtung dieses Aspekts widerspiegeln.

Drei Grundklassen von Alternativenmerkmalen, die in jedem Nutzenanalyse-Modell enthalten sind

Drei Grundklassen von Alternativenmerkmalen werden von Boudreau (1989) unterschieden, die nach seiner Ansicht in jedem Nutzenanalyse-Modell enthalten sind:

- *Quantität* bezeichnet die Anzahl der Mitarbeiter und die Anzahl der Zeitabschnitte, die durch das analysierte Programm betroffen sind.
- *Qualität* umschreibt die Konsequenzen (pro Person oder Zeitabschnitt), die sich durch das analysierte Programm ergeben
- Über die *Kosten* werden die zur Implementation und Aufrechterhaltung des Programms benötigten Ressourcen beschrieben.

In allen definierten Payoff-Funktionen wird das Produkt aus Quantitäts- und Qualitätsmerkmalen gebildet, von dem dann die Kosten subtrahiert werden.

Nutzenanalyse-Modelle können als spezifische Form von Entscheidungsmodellen oder als Weiterführung der (klassischen) Testtheorie verstanden werden.

Wenn *Nutzenanalyse-Modelle als spezifische Form von Entscheidungsmodellen* verstanden werden, so hängt die Brauchbarkeit der Modelle von ihrer Fähigkeit zur adäquaten Beschreibung, Vorhersage, Erklärung und Verbesserung von Entscheidungen (hier: im Personalbereich) ab. Die Entscheidungssituation wird offengelegt, die als wichtig angesehenen Merkmale der Entscheidungsalternativen werden explizit genannt und ihre angenommene Bedeutsamkeit in der Definition der Payoff-Funktion offengelegt. Es werden in systematischer Form Alternativenpräferenzen deduziert, und für Diskussions- und Kritikansätze werden mögliche Orientierungspunkte gegeben (vgl. Russell, Colella und Bobko, 1993, für eine ausführlichere Diskussion dieser Zielsetzung).

Im testtheoretischen Sinne können *Nutzenanalyse-Modelle als Weiterführung der (klassischen) Testtheorie* verstanden werden. Sie beschreiben durch die Erweiterung des einfachen Validitätskoeffizienten die Spezifika einer organisationspsychologischen Testanwendung und berücksichtigen relevante Randbedingungen, deren Nichtbeachtung zu gravierenden Fehlentscheidungen führen würde.

5.2 Einige Nutzenanalyse-Modelle

Die im folgenden dargestellten Nutzenanalyse-Modelle bauen historisch gesehen aufeinander auf. Während beim Taylor-Russell-Modell die Erweiterung der Testtheorie deutlich wird, sind bei dem Brodgen-Cronbach-Gleser-Modell die im letzten Abschnitt eingeführten MAU-Grundelemente in besonders klarer Form erkennbar.

Beim Taylor-Russell-Modell werden neben dem Validitätskoeffizienten auch die Grundquote der qualifizierten Bewerber und die angestrebte Selektionsquote berücksichtigt.

Taylor-Russell-Modell (T-R-Modell)
Taylor und Russell (1939) definieren in ihrem Modell eine Erfolgsquote („success ratio"), die den Anteil der Erfolgreichen an den über ein eignungsdiagnostisches Verfahren ausgewählten Bewerbern angibt. Die Erfolgsquote wird in Beziehung gesetzt zu drei Einflußgrößen:

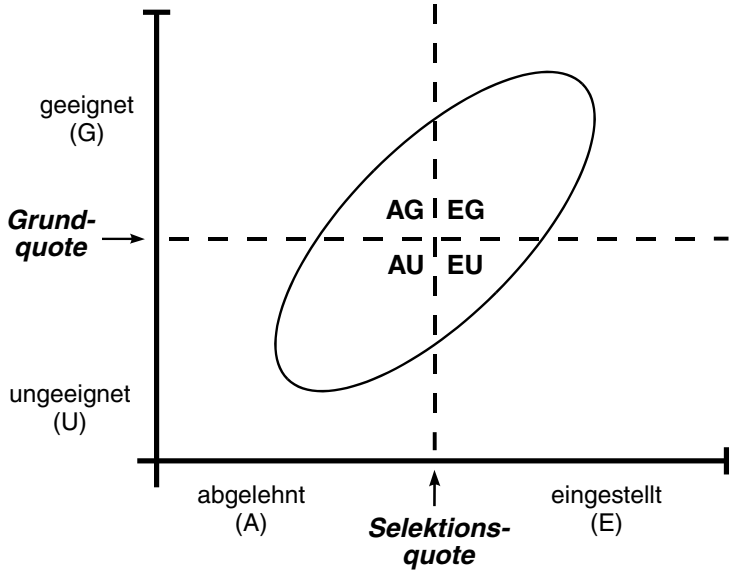
- dem *Validitätskoeffizienten* des eingesetzten Verfahrens,
- der *Grundquote*, die den „wahren" Anteil der Geeigneten unter den Bewerbern angibt,
- der *Selektionsquote*, die den Anteil der Ausgewählten unter den Bewerbern angibt.

Taylor und Russell stellten ein ausführliches Tabellenwerk zusammen, in dem die Erfolgsquote in Abhängigkeit von verschiedenen Ausprägungskombinationen dieser Einflußgrößen angegeben wird (ein Nachdruck findet sich bei Cascio, 1991). Im Kasten 9 ist eine graphische Interpretation des Modells wiedergegeben.

Das Modell stellt eine wichtige Erweiterung gegenüber der bloßen Inspektion des Validitätskoeffizienten dar: So ist die Validität eines Verfahrens um so wichtiger, je niedriger die Grundquote und je größer die Selektionsquote ist. Andererseits ist das Validitätskriterium weniger bedeutsam, wenn die Grundquote hoch und die Selektionsquote niedrig ist.

Graphische Interpretation des Taylor-Russell-Modells

Kasten 9:
Graphische Interpretation des Taylor-Russell-Modells (nach Cascio, 1991)



Die beiden dichotomen Einflußgrößen „Selektionsquote (E / (E+A))" und „Grundquote (G / (G + U))" stellen Abszisse und Ordinate des Diagramms dar. Durch die Überschneidungsbereiche der Quotenmarkierungen werden vier Quadranten gebildet:

– EG „true positive": Anteil der zu Recht (weil geeignet) Eingestellten
– AG „false negative": Anteil der zu Unrecht (weil eigentlich geeignet) Abgelehnten

– AU „true negative“: Anteil der zu Recht (weil ungeeignet) Abgelehnten
– EU „false positive“: Anteil der zu Unrecht (weil ungeeignet) Eingestellten

Die eingezeichnete Ellipse beschreibt die Gruppe der untersuchten Bewerber mit ihrer Kombination aus erzieltem eignungsdiagnostischen Prädiktorwert und „wahrem“ Berufserfolg. Die Validität des eingesetzten Auswahlverfahrens ist durch die Bauchigkeit der Ellipse und die damit erfaßten EG/AG/AU/EU-Größen wiedergegeben.

Die Erfolgsquote wird über eine Payoff-Funktion beschrieben, die das Verhältnis der ausgewählten Geeigneten zu allen Eingestellten (EG / (EG+EU)) in Beziehung setzt und gleichzeitig einen linearen Zusammenhang zwischen Prädiktor und Kriterium (= Verwendung eines Korrelationskoeffizienten als Validitätsindikator) annimmt.

Änderungen der Einflußgrößen wirken sich unter Konstanthaltung der übrigen Bedingungen folgendermaßen auf die Erfolgsquote aus:

– Eine höhere Validität des Auswahlverfahrens führt zu einer schmaleren Ellipse, d.h. damit werden die „false“-Bereiche AG und EU verkleinert, und die Erfolgsquote steigt.
– Eine höhere Grundquote (z. B. durch ein verbessertes Vorauswahlverfahren) äußert sich graphisch durch eine Absenkung der horizontalen cutoff-Gerade. Dadurch vergrößert sich u.a. der EG- und verkleinert sich der AU-Bereich, d.h. die Erfolgsquote steigt.
– Eine niedrigere Selektionsquote (weniger Einstellungen) äußert sich graphisch durch Verrücken der vertikalen cutoff-Gerade nach rechts. Dadurch werden zwar mehr Geeignete abgelehnt (Vergrößerung des AG-Bereichs), aber gleichzeitig auch weniger Ungeeignete eingestellt (Verkleinerung des EU-Bereiches). Damit steigt die Erfolgsquote.

Der allgemeine Nutzenwert errechnet sich aus der Differenz zwischen der Erfolgsquote des analysierten Verfahrens (bei spezifizierter Grund- und Selektionsquote sowie bekannter Validität) und einem Vergleichsverfahren (z. B. einfache Zufallsauswahl). Wenn beispielsweise die Selektionsquote bei einem Auswahlverfahren auf .10 festgelegt ist, die Grundquote .50 beträgt und das Verfahren eine Validität von r = .30 aufweist, so ergibt sich nach der Taylor-Russell-Tabelle eine Erfolgsquote von .71. Die Erfolgsquote einer einfachen Zufallsauswahl würde der Grundquote (.50) entsprechen. Somit ergibt sich ein Nutzenwert von .71 – .50 = .21, d.h. es resultiert beim Einsatz des Verfahrens eine Nutzensteigerung von 21% im Vergleich zur Zufallsauswahl.

Eine höhere Validität, eine höhere Grundquote sowie eine niedrigere Selektionsquote führen (ceteris paribus) zu einer steigenden Erfolgsquote.

Zu beachten ist allerdings, daß einige für eine Alternativenentscheidung wichtige Variablen im Modell nur vereinfacht oder überhaupt nicht berücksichtigt werden. Bezogen auf die im letzten Abschnitt eingeführten Grundmerkmale Quantität, Qualität und Kosten bedeutet das: Quantitätsinformationen (absolute Anzahl der Bewerber, beeinflußte Zeitabschnitte) werden nicht, Qualitätsinformationen nur zum Teil (drei Einflußgrößen werden verarbeitet) und vereinfacht berücksichtigt, so ist z. B. die Dichotomisierung des Kriteriums in „geeignet“ vs. „ungeeignet“ kritisch. Kostenaspekte werden im Grundmodell gar nicht betrachtet.

Brodgen-Cronbach-Gleser-Modell (B-C-G-Modell)

Von den Weiterentwicklungen des T-R-Modells ist das im wesentlichen von Brodgen (1949) sowie Cronbach und Gleser (1965) entwickelte Modell am bedeutendsten. Es fußt auf dem Prinzip der linearen Regression und stellt die konzeptionelle Grundlage der meisten späteren Nutzenmodelle dar. In Kasten 10 ist die Payoff-Funktion des Modells (erweiterte Form nach Funke, Schuler & Moser, 1995) dargestellt.

Das Brodgen-Cronbach-Gleser-Modell fußt auf dem Prinzip der linearen Regression und stellt die konzeptionelle Grundlage der meisten späteren Nutzenmodelle dar.

In der Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells sind Qualitäts-, Quantitäts- und Kostenterme gut abgrenzbar.

Kasten 10:
Payoff-Funktion des Brodgen-Cronbach-Gleser-Modells

$$\Delta U = N_E \cdot T \cdot SD_y \cdot r_{xy} \cdot \overline{z_x} - C \cdot N_B$$

mit

ΔU	:	Nutzenzuwachs durch das Verfahren (in Geldeinheiten, z. B. US-Dollar)
N_E	:	Anzahl der Eingestellten
T	:	Anzahl der berücksichtigten Zeiteinheiten (z. B. Jahre)
SD_y	:	Standardabweichung des Kriteriums in Geldeinheiten
r_{xy}	:	Validitätskoeffizient
$\overline{z_x}$	:	durchschnittlicher standardisierter Prädiktorwert der Ausgewählten
C	:	Kosten pro Bewerber
N_B	:	Anzahl der Bewerber

Für eine Erklärung der Formel soll wieder auf die drei Grundelemente aller Nutzenmodelle zurückgegriffen werden:

Der *Qualitätsterm* wird über die Produktkette aus SD_y, r_{xy} und $\overline{z_x}$ gebildet. Das Produkt $r_{xy} \cdot \overline{z_x}$ wird hierbei als (lineare) Vorhersage für die durchschnittliche standardisierte Kriteriumsleistung der Ausgewählten, $\overline{z_y}$, verwendet. Die geschätzte Standardabweichung des Kriteriums in Geldeinheiten SD_y dient als Skalierungsfaktor, um die Kriteriumsleistung zu gewichten: Bessere Leistungen der Ausgewählten implizieren auch einen höheren monetären Nutzen.

Der *Quantitätsanteil* wird durch das Produkt aus N_E und T gebildet. In ihm werden die betroffenen Personen-Zeit-Einheiten beschrieben: Je mehr Personen ausgewählt werden und je größer ihre Verweildauer in der Organisation ist, desto größer ist auch der monetäre Nutzen des eingesetzten Verfahrens. Von diesem Bruttonutzen werden die *Kosten* des Verfahrens $C \cdot N_B$ abgezogen. Das Ergebnis ist ΔU, der inkrementelle Nettonutzen des eingesetzten Verfahrens, in diesem Fall bezogen auf eine einfache Zufallsauswahl ($r_{xy} = 0$).

In Kasten 11 ist ein hypothetisches Anwendungsbeispiel des Modells gegeben.

Kasten 11:
Anwendungsbeispiel zum Brodgen-Cronbach-Gleser-Modell

Ein einfaches Anwendungsbeispiel zum Brodgen-Cronbach-Gleser-Modell

Als ein hypothetisches Anwendungsbeispiel soll die Auswahl von Bankkaufleuten (nach Abschluß der Ausbildung) dienen:

Von 585 Bewerbern (= N_B) sollen 120 Personen (= N_E) eingestellt werden. Aus Dokumentenanalysen ist bekannt, daß die durchschnittliche Verweildauer von Bankkaufleuten in dieser Bank durchschnittlich 10.2 Jahre (= T) beträgt.

Zur Auswahl wird ein standardisiertes Testverfahren verwendet, das eine prognostische Valdität von .35 (= r_{xy}) aufweist und pro Anwendung 270 DM (= C) kostet. Die als geeignet eingestuften Bewerber erzielen bei diesem Test einen durchschnittlichen standardisierten Testwert von .5 (= $\overline{z_x}$), d.h. ihr Testwert liegt durchschnittlich eine halbe Standardabweichung über dem Mittelwert aller Bewerber. Ein Schätzverfahren (s.u.) soll ergeben haben, daß eine Standardabweichung des Nutzens der Leistung eines Bankkaufmanns bzw. einer Bankkauffrau 45000 DM pro Jahr (= SD_y) beträgt.

Quantität = Eingestellte · Verweildauer = 120 Personen · 10.2 Jahre = 1224 Personen-Jahre

Qualität = durchschnittlicher Testscore · Validitätskoeffizient · Standardabweichung
= .5 · .35 · 45000 DM = 7875 DM pro Personen-Jahr

Nutzen = (Quantität · Qualität) – Kosten
= (1225 Personen-Jahre · 7875 DM pro Personen-Jahr) – (585 · 270 DM)
= 9.488.925 DM

Die Verwendung des Testverfahrens würde also im Vergleich zu einer bloßen Zufallsauswahl (r_{xy}=0) für die nächsten zehn Jahre einen inkrementellen Nutzenzuwachs von ca. 9 ½ Millionen DM erwarten lassen.

Die Schätzung der Standardabweichung des Kriteriums in Geldeinheiten stellt immer noch eine Achillesferse der Nutzenanalysen dar.

Eine besondere Schwierigkeit des Ansatzes liegt in der Schätzung des SD_y-Parameters, der Standardabweichung des Kriteriums in Geldeinheiten. Boudreau (1991) unterscheidet hier mehrere Ansätze, von denen die prominentesten Verfahren kurz charakterisiert werden sollen (vgl. hierzu auch Holling, 1998):

- *Kostenrechnungsverfahren*: Hier wird möglichst jeder durch ein Individuum geleisteten Produktionseinheit (Dienstleistung, Produkt usw.) ein Geldwert zugewiesen, abhängig von ihrem Beitrag zum Ertrag der Organisation. Die Standardabweichung dieser individuellen Produktivitätsindikatoren dient als SD_y-Schätzer. Trotz des hohen Aufwands unterliegen die Ergebnisse dieses Ansatzes durch willkürliche Festlegungen meistens großer Beliebigkeit (Boudreau, 1991, p. 652).
- *Globale Einschätzung*: Experten werden gebeten, den Geldwert verschiedener Prozentrangausprägungen von Mitarbeiter-Leistungen zu schätzen. Bei Annahme einer Normalverteilung der Leistungseinschätzungen dienen dann die Differenzen (Prozentrang 50 – Prozentrang 15) sowie (Prozentrang 85 – Prozentrang 50) als SD_y-Schätzer. Auch hier sind die Befunde gemischt, trotzdem gehören globale Einschätzungen zu den am häufigsten eingesetzten Schätzverfahren.
- *Individualisierte Schätzung*: In dem prominentesten Ansatz dieser Herangehensweise, der sog. „Cascio-Ramos estimate of performance in dollars (CREPID)"-Methode und deren Weiterführungen, wird in einem zweistufigen Prozeß vorgegangen (vgl. Funke & Barthel, 1995, S. 827):
  1. Die Gesamtleistung jedes Mitarbeiters ergibt sich durch Addition der Leistung in einzelnen hinsichtlich Häufigkeit und Wichtigkeit gewichteten Arbeitsaufgaben.
  2. Der Geldwert ergibt sich durch Multiplikation der Gesamtleistung beispielsweise mit dem Jahresgehalt des Mitarbeiters.

  Die Standardabweichung des Geldwerts in der Mitarbeiter-Stichprobe dient als SD_y-Schätzer. Funke und Barthel (1995) beurteilen besonders die Weiterentwicklungen des Verfahrens positiv, da diese auf eine deutlich verbesserte Ökonomie hinauslaufen und exaktere Schätzungen liefern als globale Einschätzungen.

Es bleibt festzuhalten, daß die SD_y-Schätzung ein z.Z. noch offenes Problem der Nutzenanalysen darstellt. Die Nutzenschätzungen unterliegen deshalb abhängig von den eingesetzten Schätzmethoden (konservative vs. optimistische Verfahren) relativ großen Schwankungen.

Weitere Nutzenanalyse-Modelle

Weiterentwicklungen bestehen in den meisten Fällen nur in einer Ergänzung des Brodgen-Cronbach-Gleser-Ansatzes um weitere Alternativenmerkmale.

Weiterentwicklungen des B-C-G-Modells bestehen in den meisten Fällen in einer Ergänzung des Grundansatzes um weitere Alternativenmerkmale. So setzen z. B. Schuler, Funke, Moser und Donat (1995) für ihre Nutzenkalkulation eines Auswahlverfahrens für Mitarbeiter im Forschungs- und Entwicklungsbereich industrieller Großunternehmen ein Modell in Anlehnung an Boudreau (1983) sowie Cronshaw und Alexander (1985) ein. In ihm wird das B-C-G-Modell um einige betriebswirtschaftliche Parameter erweitert (Schuler et al., 1995, S. 181):

- fixe Testkosten (für Testentwicklung, Schulung, Lizenzgebühren usw.) und variable Testkosten (wegen schwankender Bewerberzahlen) bei der Anwendung;

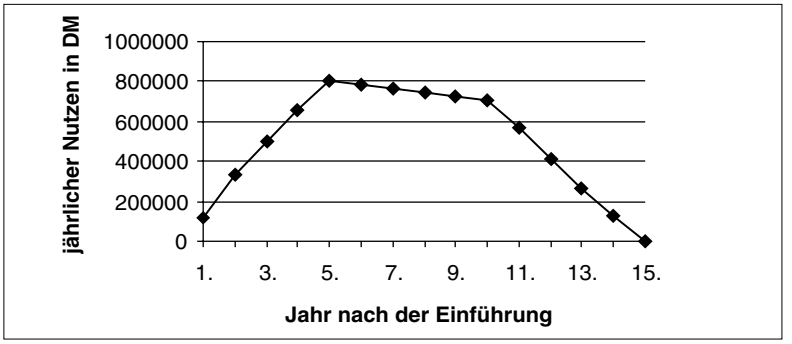
- Berücksichtigung mehrerer Anwendungsperioden des Verfahrens;
- Kohortenbetrachtung der kontinuierlichen Mitarbeiterzu- und -abgänge als Dynamisierung der Verweildauer;
- Berücksichtigung der Ablehnungswahrscheinlichkeit eines Stellenangebotes durch ausgewählte Mitarbeiter;
- Berücksichtigung variabler Kosten der Mehrleistung, bedingt z. B. durch besondere Gratifikationen;
- Steuern auf die Gewinne.

Diese zusätzlichen Parameter werden fix gesetzt oder auf der Grundlage der bestehenden Datenlage geschätzt. Die Ergebnisse der Studie von Schuler et al. sind in Kasten 12 dargestellt.

Beispiel für eine Nutzenanalyse zu einem Personalauswahlverfahren im Bereich Forschung und Entwicklung

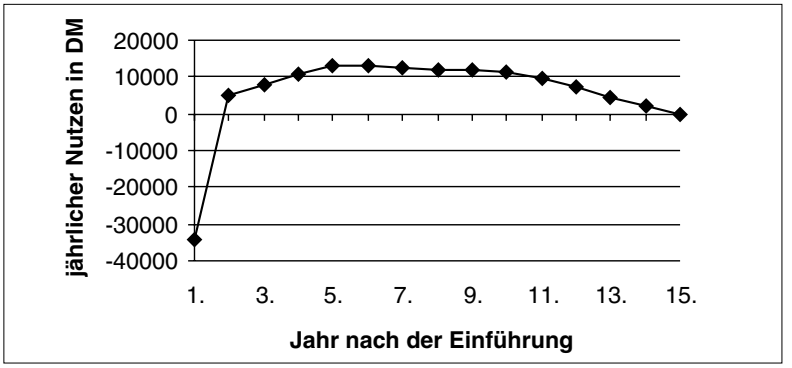
Kasten 12:
Nutzenanalyse zu einem F&E-Auswahlverfahren (nach Schuler et al., 1995, S. 175-199)

In der ersten Abbildung ist der zu erwartende jährliche monetäre Nutzen des entwickelten Auswahlverfahrens bei 10jährigem Einsatz und einer jährlichen Einstellung von 25 Personen angegeben. Es ergibt sich ein Gesamtnutzen von 7,5 Mio DM nach Steuern.



Weitere Randbedingungen des angenommenen Regelfalls:
- Selektionsrate 1:10
- Verweildauer: 5 Jahre
- Validitätsvorsprung des neuen Verfahrens: r = .20
- Gesamtsteuersatz: 66,66%
- durchschnittlicher standardisierter Prädiktorwert der Ausgewählten: $\overline{z_x} = 1{,}55$
- Standardabweichung der Leistung: SD_y = 100.000 DM
- Summe der fixen Kosten: 109.000 DM
- Summe der variablen Kosten: 340 DM pro Bewerber

In der zweiten Abbildung ist der zu erwartende jährliche monetäre Nutzen für ein „worst case"-Szenario angegeben: Es werden für kurzfristige Bedingungen im Prinzip noch mögliche, aber sehr negative Parameter gesetzt. Bei 10jährigem Einsatz und einer jährlichen Einstellung von 5 Personen ergibt sich noch ein Gesamtnutzen von 84.000 DM.



Weitere Randbedingungen des „worst case"-Szenarios:
- Selektionsrate 1:2
- durchschnittlicher standardisierter Prädiktorwert der Ausgewählten: $\overline{z}_x = 0{,}49$
- Standardabweichung der Leistung: SD_y = 27.000 DM
- übrige Angaben s.o.

Übertragbarkeit der Modelle auf den Personalentwicklungsbereich

Zwar werden Nutzenanalysen vorrangig bei Personalauswahlverfahren eingesetzt, prinzipiell sind die Modelle aber ohne weiteres auch auf den Personalentwicklungsbereich übertragbar.

Auch wenn in der bisherigen Darstellung und in den gewählten Beispielen immer Bezug genommen wurde auf die Personalauswahl, sind Nutzenanalysen prinzipiell auch auf den Personalentwicklungsbereich anwendbar. Hierfür müssen nur die Alternativenmerkmale inhaltlich reinterpretiert werden: Beispielsweise wird im „characteristic-changing model"-Ansatz (Landy, Farr & Jacobs, 1982; Schmidt, Hunter & Pearlman, 1982) die „Gruppe der Eingestellten" reinterpretiert als „Gruppe der Trainierten" und verglichen mit einer Gruppe untrainierter Organisationsmitglieder. Gleichzeitig wird anstelle der Ergebnisprognose die aus den eingesetzten Teststatistiken abgeleitete standardisierte Effektgröße d verwendet (vgl. Cascio, 1989, für eine PE-orientierte Darstellung von Nutzenanalysen).

Zusammenfassung

Zusammenfassung

Dieses Kapitel befaßte sich mit der Erfolgsüberprüfung von personalpsychologischer Arbeit. Hierfür wurde zunächst die traditionelle Verwendung des Validitätsbegriffs untersucht. Die Übertragung der konventionellen Konzepte auf die Problemsituation der beruflichen Eignungsdiagnostik wurde kritisch hinterfragt. Danach wurde ein einfaches Rahmenmodell zur Validität in der beruflichen Eignungsdiagnostik eingeführt, mit dem die Validitätskontrolle als erweiterte Hypothesenprüfung reinterpretiert werden kann.

Als nächster Schritt wurde die Verwendung diagnostischer Informationen für Personalentscheidungen untersucht. Eine Typologie zeigte, daß die Selektion geeigneter Bewerber in einer Personalauswahlsituation nur einen Spezialfall denkbarer Entscheidungssituationen darstellt. Die Problemlage bei Klassifikationsentscheidungen wurde darauffolgend beispielhaft vertieft. Neben unterschiedlichen Klassifikationsstrategien wurde ausführlicher auf das Grundprinzip von statistischen Klassifikationssystemen eingegangen.

Im Abschnitt „Evaluation von Personalentwicklungsmaßnahmen" wurde dann ein umfassendes Konzept zur Erfolgskontrolle von Interventionsprogrammen untersucht. Nach der Darstellung einiger elementarer Evaluationsbestandteile wurde ausführlicher auf relevante Einflußvariable der Trainingseffektivität eingegangen, die für eine Evaluation in diesem Bereich unbedingt berücksichtigt werden müssen.

Den Abschluß bildete ein Abschnitt zu ökonomischen Nutzenanalysen für Personalauswahl- und Personalentwicklungsprogramme. Nach einer Charakterisierung der bestehenden Ansätze in diesem Bereich als spezifische Formen von Entscheidungsmodellen wurden zwei prominente Modelle (das Taylor-Russell-Modell und der Brodgen-Cronbach-Gleser-Ansatz) vorgestellt.

In jedem Abschnitt wurden zunächst einige Grundbegriffe des jeweiligen Themengebiets nähergebracht und der prinzipielle Ansatz skizziert. Nach Möglichkeit wurde ein Rahmenmodell vorgestellt, das mit einigen Beispielen näher erläutert wurde. Ziel war dabei immer, Informationen für eine systematisierte Analyse der personalpsychologischen Entscheidungssituation zu geben.

Weiterführende Literatur

Weiterführende Literatur

Binning, J.F. & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 621-745). Palo Alto, CA: Consulting Psychologists Press.

Funke, U. & Barthel, E. (1995). Nutzenanalysen von Personalauswahlprogrammen. In W. Sarges (Hrsg.), *Management Diagnostik* (S. 820-833). Göttingen: Hogrefe.

Johnson, C.D. & Zeidner, J. (1991). *The economic benefits of predicting job performance: Vol. 2. Classification efficiency*. New York: Praeger.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

Nork, M.E. (1991). *Management Training: Evaluation – Probleme – Lösungsansätze*. München: Hampp.

Literatur

Literatur

Alley, W.E. (1994). Recent advances in classification theory and practice. In M.G. Rumsey, J.H. Walker & J.H. Harris (Eds.), *Personnel selection and classification* (pp. 431-442). Hillsdale, NJ: Lawrence Erlbaum.

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

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

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psychology, 37*, 1-15.

Antoni, C.H. (1993). Evaluationsforschung in der Arbeits- und Organisationspsychologie. In W. Bungard & T. Herrmann (Hrsg.), *Arbeits- und Organisationspsychologie im Spannungsfeld zwischen Grundlagenforschung und Anwendung* (S. 309-337). Bern: Huber.

Baldwin, T.T. & Ford, J.K. (1988). Transfer of training: A review and directions for future research. *Personnel Psychology, 41*, 63-105.

Barrick, M.R. & Mount, M.K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology, 44*, 1-26.

Bierhoff, H.W. & Rudinger, G. (1996). Quasi-experimentelle Untersuchungsmethoden. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 47-58). Weinheim: Beltz-PVU.

Binning, J.F. & Barrett, G.V. (1989). Validity of personnel decisions: A conceptual analysis of the inferential and evidential bases. *Journal of Applied Psychology, 74*, 478-494.

Borkenau, P. & Ostendorf, F. (1993). *NEO-Fünf-Faktoren Inventar (NEO-FFI)*. Göttingen: Hogrefe.

Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin: Springer.

Boudreau, J.W. (1983). Economic considerations in estimating the utility of human resource productivity improvement programs. *Personnel Psychology, 36*, 551-557.

Boudreau, J.W. (1989). Selection utility analysis: A review and agenda for future research. In M. Smith & I.T. Robertson (Eds.), *Advances in selection and assessment* (pp. 227-257). Chichester, England: John Wiley & Sons.

Boudreau, J.W. (1991). Utility analysis for decisions in human resource management. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 621-745). Palo Alto, CA: Consulting Psychologists Press.

Fortsetzung Literatur

Brogden, H.E. (1949). When testing pays off. *Personnel Psychology, 2*, 171-183.

Brogden, H.E. (1959). Efficiency of classification as a function of number of jobs, percent rejected, and the validity and intercorrelation of job performance estimates. *Educational and Psychological Measurement, 19*, 181-190.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-105.

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*, 141-164.

Cascio, W.F. (1989). Using utility analysis to assess training outcomes. In I.L.Goldstein (Ed.), *Training and development in organizations* (pp. 63-88). San Francisco, CA: Jossey-Bass.

Cascio, W.F. (1991). *Costing human resources: the financial impact of behavior in organizations*. Boston, MA: Kent.

Cascio, W.F. (1997). *Applied psychology in personnel management*. Englewood Cliffs, NJ: Prentice-Hall.

Cook, T.D. & Reichardt, C.S. (1979). *Quantitative and qualitative methods in evaluation research*. Beverly Hills, CA: Sage.

Cronbach, L.J. (1990). *Essentials of psychological testing*. New York: Harper & Row.

Cronbach, L.J. & Gleser, G.C. (1965). *Psychological tests and personnel decisions*. Urbana: University of Illinois Press.

Cronbach, L.J. & Meehl, P.E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281-302.

Cronshaw, S.F. & Alexander, H.A. (1985). One answer to the demand for accountability: Selection utility as an investment decision. *Organizational Behavior and Human Decision Processes, 35*, 102-118.

Edmundson, E.W., Koch, W.B. & Silverman, S. (1993). A facet analysis approach to content and construct validity. *Educational and Psychological Measurement, 53*, 351-368.

Fink, A. (1995). *Evaluation for education and psychology*. Thousand Oaks, CA: Sage.

Funke, U. & Barthel, E. (1995). Nutzenanalysen von Personalauswahlprogrammen. In W. Sarges (Hrsg.), *Management Diagnostik* (S. 820-833). Göttingen: Hogrefe.

Funke, U., Schuler, H. & Moser, K. (1995). Nutzenanalyse zur ökonomischen Evaluation eines Personalauswahlprojekts für Industrieforscher. In T.J. Gerpott & S.H. Siemers (Hrsg.), *Controlling von Personalprogrammen* (S. 139-171). Stuttgart: Schäffer-Poeschel.

Ghiselli, E.E. & Brown, C.W. (1955). *Personnel and industrial psychology* (2nd ed.). New York: McGraw-Hill.

Goldstein, I.L. (1993). *Training in organizations: Needs assessment, development, and evaluation*. Pacific Grove, CA: Brooks/Cole.

Guion, R.M. (1977). Content validity – the source of my discontent. *Applied Psychological Measurement, 1*, 1-10.

Guion, R.M. (1991). Personnel assessment, selection, and placement. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (Vol. 2, pp. 327-395). Palo Alto, CA: Consulting Psychologists Press.

Häcker, H., Leutner, D. & Amelang, M. (1998). *Standards für pädagogisches und psychologisches Testen*. Göttingen: Hogrefe.

Herrmann, T. (1976). *Die Psychologie und ihre Forschungsprogramme*. Göttingen: Hogrefe.

Hesketh, B. (1997). Dilemmas in training for transfer and retention. *Applied Psychology: An International Review, 46*, 317-386.

Hogan, R. & Nicholson, R.A. (1988). The meaning of personality test scores. *American Psychologist, 43*, 621-626.

Holling, H. (1998). Utility analysis of personnel selection: An overview and empirical study based on objective performance measures. *Methods of Psychological Research Online* [On-line serial], *3*(1). Retrieved January 25, 2000 from the World Wide Web: http://www.mpr-online.de.

Fortsetzung Literatur

Horst, P. (1955). A technique for the development of a multiple absolute prediction battery. *Psychological Monographs, 69*, 1-22.

Janke, W. (1982). Klassenzuordnung. In K.-J. Groffmann & L. Michel (Hrsg.), *Grundlagen psychologischer Diagnostik. Enzyklopädie der Psychologie B/II/1* (S. 376-466). Göttingen: Hogrefe.

Johnson, C.D. & Zeidner, J. (1991). *The economic benefits of predicting job performance: Vol. 2. Classification efficiency*. New York: Praeger.

Kanfer, F.H., Reinecker, H.S. & Schmelzer, D. (2000). *Selbstmanagement-Therapie: ein Lehrbuch für die klinische Praxis*. Berlin: Springer.

Kirkpatrick, D.L. (1987). Evaluation of training. In R.L. Craig (Ed.), *Training and development handbook: A guide to human resource development* (pp. 301-319). New York: McGraw-Hill.

Klimoski, R. (1993). Predictor constructs and their measurement. In N. Schmitt & W.C. Borman (Eds.), *Personnel selection in organizations* (pp. 99-134). San Francisco, CA: Jossey-Bass.

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

Landy, F.J. (1986). Stamp collection versus science: Validation as hypothesis testing. *American Psychologist, 4*, 1183-1192.

Landy, F.J., Farr, J.L. & Jacobs, R.R. (1982). Utility concepts in performance measurement. *Organizational Behavior and Human Performance, 30*, 15-40.

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.

Lienert, G.A. & Raatz, U. (1994). *Testaufbau und Testanalyse*. Weinheim: PVU-Beltz.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment. *American Psychologist, 50*, 741-749.

Mohr, L.B. (1995). *Impact analysis for program evaluation*. London: Sage.

Mumford, M.D., Costanza, D.P., Connelly, M.S. & Johnson, J.F. (1996). Item generation procedures and background data scales: Implications for construct and criterion-related validity. *Personnel Psychology, 49*, 361-398.

Nicholson, N. & Johns, G. (1985). The absence culture and the psychological contract: Who's in control of absence? *Academy of Management Review, 10*, 397-407.

Noe, R.A. & Ford, J.K. (1992). Emerging issues and new directions for training research. *Research on Personnel and Human Resources Managment, 10*, 345-384.

Nork, M.E. (1991). *Management Training: Evaluation – Probleme – Lösungsansätze*. München: Hampp.

Ree, M.J. & Carretta, T.R. (1998). General cognitive ability and occupational performance. In C.L. Cooper & I.T. Robertson (Eds.), *International Review of Industrial and Organizational Psychology* (Vol. 13, pp. 159-184). New York: John Wiley & Sons.

Rossi, P.H. & Freeman, H.E. (1993). *Evaluation. A systematic approach*. London: Sage.

Rumsey, M.G., Walker, C.B. & Harris, J.H. (Eds.). (1994). *Personnel selection and classification*. Hillsdale, NJ: Lawrence Erlbaum.

Russell, C.J., Colella, A. & Bobko, P. (1993). Expanding the context of utility: the strategic impact of personnel selection. *Personnel Psychology, 46*, 781-801.

Schmidt, F.L., Hunter, J.E. & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work-force productivity. *Personnel Psychology, 35*, 333-347.

Scholarios, D.M., Johnson, C.D. & Zeidner, J. (1994). Selecting predictors for maximizing the classification efficiency of a battery. *Journal of Applied Psychology, 79*, 412-424.

Schuler, H. (1996). *Psychologische Personalauswahl. Einführung in die berufliche Eignungsdiagnostik*. Göttingen: Hogrefe.

Fortsetzung Literatur

Schuler, H., Funke, U., Moser, K. & Donat, M. (1995). *Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren*. Göttingen: Hogrefe.

Smith, M., Farr, J.L. & Schuler, H. (1993). Individual and organizational perspectives on personnel procedures: Conclusions and horizons for future research. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 333-351). Hillsdale, NJ: Lawrence Erlbaum.

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.

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology, 23*, 565-578.

Tenopyr, M.L. (1977). Content-construct confusions. *Personnel Psychology, 30*, 47-54.

Thierau, H. (1991). *Analyse und empirische Überprüfung wissenschaftlicher Evaluationskonzepte in der betrieblichen Weiterbildung. Dargestellt am Beispiel der Schulung von Führungskräften in Personalbeurteilung*. Unveröff. Diss., Ruhr-Universität Bochum.

Thierau, H., Stangel-Meseke, M. & Wottawa, H. (1999). Evaluation von Personalentwicklungsmaßnahmen. In Kh. Sonntag (Hrsg.), *Personalentwicklung in Organisationen* (2. Aufl., S. 261-286). Göttingen: Hogrefe.

Wottawa, H. (1996). Methoden der Evaluationsforschung. In E. Erdfelder, R. Mausfeld, T. Meiser & G. Rudinger (Hrsg.), *Handbuch Quantitative Methoden* (S. 551-566). Weinheim: Beltz-PVU.

Wottawa, H. & Thierau, H. (1998). *Lehrbuch Evaluation*. Bern: Huber.

Zeidner, J. & Johnson, C.D. (1994). Is personnel classification a concept whose time has passed? In M.G. Rumsey & C.B. Walker (Eds.), *Personnel selection and classification* (pp. 377-410). Hillsdale, NJ: Lawrence Erlbaum.