15 Leistungsbeurteilung

von Bernd Marcus und Heinz Schuler

Inhaltsübersicht

1 Einleitung

Berufliche Leistung von abhängig Beschäftigten läßt sich als Beitrag zu den Zielen einer Organisation definieren.

Wann immer Menschen ihre Arbeitskraft gegen Entgelt zur Verfügung stellen, wird von ihnen erwartet, diese im Sinne dessen (eines Kunden, Klienten, der Gesellschaft) einzusetzen, der dafür eine Gegenleistung erbringt. Im Falle von abhängig Beschäftigten besteht diese Erwartung darin, *einen Beitrag zu den Zielen einer Organisation zu leisten*. Daß dies legitim sei und daß die Organisationsmitglieder in erster Linie an diesem Beitrag gemessen werden – und nicht an anderen vorstellbaren Kriterien wie Alter *(Senioritätsprinzip)* oder Abstammung *(Adelsprinzip)* – ist Kerngedanke des unser Wirtschaftsleben dominierenden Leistungsprinzips *(meritokratisches Prinzip)*. Letztlich sind alle in diesem Band beschriebenen personalpsychologischen Maßnahmen, ob Personalauswahl, -zuordnung, -entwicklung oder -führung, darauf gerichtet, berufliche Leistungen von Individuen, Teams oder dem gesamten Mitarbeiterstamm zu verbessern. Ihr Erfolg sollte auch in erster Linie an diesem Kriterium gemessen werden, was, ebenso wie die meritokratische Zuteilung von materiellen oder ideellen Gütern, voraussetzt, daß die Leistung selbst in möglichst geeigneter Weise gemessen wird. Gleichzeitig stellt diese Leistungsmessung – i.d.R. in Form einer subjektiven Beurteilung – jedoch keinen passiven Meßvorgang dar, sondern wirkt ihrerseits wiederum als Intervention – mit nicht nur erwünschten Folgen, wie noch zu zeigen sein wird.

Leistungsbeurteilung (LB) ist kein passiver Meßvorgang, sondern hat immer auch den Charakter einer Intervention.

Leistungsbeurteilung läßt sich in Organisationen, die sich am Leistungsprinzip orientieren, nicht wirklich umgehen, setzt jedoch nicht unbedingt eine explizite Form oder gar einen formalisierten Ablauf voraus. Dennoch verfügt in der Praxis die Mehrzahl zumindest der größeren deutschen Unternehmen über formalisierte (Personal-) Beurteilungssysteme. Dabei überwiegen standardisierte, oft organisationsübergreifende Formulare, in denen direkte Vorgesetzte ihre Mitarbeiter in regelmäßigen, oft ein-, manchmal mehrjährigen Abständen anhand einiger, manchmal vieler sehr allgemein gehaltener Merkmale auf im Mittel fünfstufigen Skalen einschätzen (zusammenfassend Gerpott & Domsch, 1995). Dieser Vorgang und v.a. das sich daran regelmäßig anschließende Mitarbeitergespräch (vgl. Kapitel 16) zählt zu den von beiden Seiten am wenigsten geschätzten Führungsaufgaben. Darüber hinaus war die Praxis betrieblicher Leistungsbeurteilung immer wieder Gegenstand heftiger Kritik von wissenschaftlicher Seite (z. B. Deming, 1986; Neuberger, 1980), die im Extrem auf die Forderung hinauslief, auf formale Beurteilungen gänzlich zu verzichten.

Hier wird die Auffassung vertreten, daß diese Forderung, bei aller teilweise berechtigten Kritik, überzogen ist. Betriebliche Leistungsbeurteilung kann ein personalpolitisches Instrument sein, dessen Nutzen die Risiken bei weitem überwiegt, wenn bestimmte Punkte beachtet werden, auf die in den folgenden Abschnitten näher einzugehen sein wird. Dazu zählt nicht nur eine methodisch ausgefeilte Fundierung, auf die sich die personalpsychologische Forschung lange Zeit konzentriert hat, sondern auch bspw. die Hinzuziehung weiterer Quellen außer dem direkten Vorgesetzten, die Beachtung der Tatsache, daß Menschen nicht zu objektiven Meßinstrumenten mutieren, wenn sie andere oder sich selbst beurteilen (sondern damit z. B. auch durchaus eigennützige Ziele verfolgen) oder die Vermeidung des Versuchs, mit ein- und demselben Beurteilungssystem inkompatible Ziele zu verfolgen. Wir beschränken uns in diesem Kapitel weitgehend auf die formale Beurteilung individueller Arbeitsleistungen. Auf das damit eng zusammenhängende Feedback sowie die Beurteilung von Gruppenleistungen wird an anderer Stelle in diesem Lehrbuch eingegangen (vgl. Kapitel 16 und 18). Zunächst betrachten wir jedoch den Gegenstand der Beurteilung, die berufliche Leistung, etwas näher.

2 Das Konstrukt der beruflichen Leistung

2.1 Leistungskriterien

Wenn wir in der Einleitung berufliche Leistung als Beitrag zu den Zielen einer Organisation definiert haben, so war diese Umschreibung noch sehr abstrakt. Leistung ist ein hypothetisches Konstrukt, das als solches nicht direkt beobachtet werden kann. Jeder Versuch, dieses Konstrukt in konkreter Weise zu erfassen, führt über die Festlegung einer Meßvorschrift für operationale Leistungsindikatoren[1] – wir messen *Leistungskriterien*. Kriterien sind immer mehr oder weniger unvollkommene Annäherungen an ein Konstrukt; die Idee, ein „endgültiges Kriterium" als optimal gewichtetes Kompositorium aller Elemente anzustreben (Thorndike, 1949), findet heute nur noch wenig Unterstützung.

Leistung ist ein hypothetisches Konstrukt, das sich nur indirekt über Kriterien erfassen läßt.

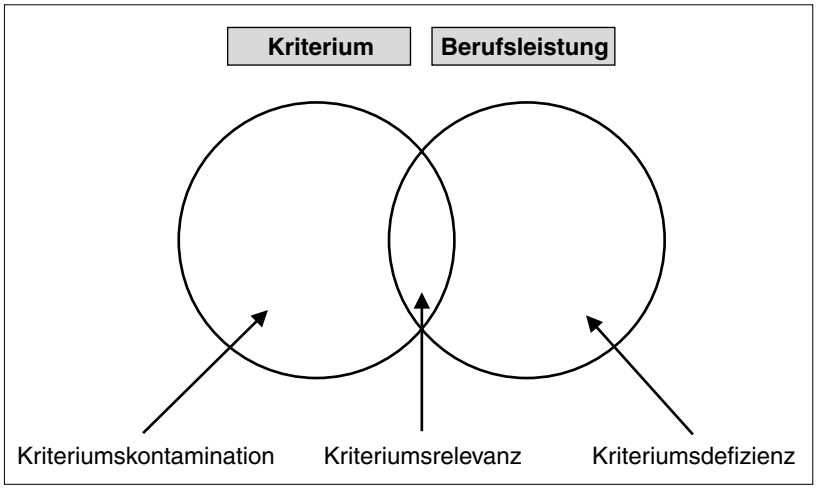


Abbildung 1:
Kriteriumsrelevanz, -defizienz und -kontamination

Das bedeutet natürlich nicht, daß Kriterien nichts mit der eigentlichen Leistung zu tun hätten. Bspw. zählt Verkaufen sicher zu den wichtigsten Aufgaben eines Außendienstmitarbeiters, und seine Verkäufe lassen sich relativ einfach über die Umsatzerlöse erfassen. In gewissem Maße sind Umsatzerlöse also *relevante* Kriterien für die Leistung des Vertreters. *Mit Kriteriumsrelevanz bezeichnen wir das Ausmaß, in dem das Kriterium Aspekte des Leistungskonstrukts erfaßt.* Andererseits soll ein Außendienstmitarbeiter vielleicht nicht nur verkaufen, sondern bspw. auch Kunden langfristig binden, den Markt beobachten und das Unternehmen mit Informationen versorgen. Über seine Leistung auf diesen Gebieten geben die Umsätze keinerlei Auskunft; sie sind in dieser Hinsicht *defizient*. *Kriteriumsdefizienz wird der Teil der tatsächlichen Leistung genannt, der vom Kriterium nicht erfaßt wird.* Schließlich werden auch die Umsatzerlöse ihrerseits nicht nur von der Leistung des Vertreters beeinflußt. Unterschiedliches Potential der Verkaufsbezirke, Aktivitäten der Konkurrenz oder die allgemeine und branchenspezifische Konjunkturlage bestimmen den Verkaufserfolg mit und liegen außerhalb der Macht des einzelnen Mitarbeiters; seine Umsätze sind zu einem gewissen Grade *kontaminiert*. *Mit Kriteriumskontamination ist der Teil des Kriteriums gemeint, der etwas anderes als das angestrebte Konstrukt erfaßt.* Abbildung 1 zeigt die Zusammenhänge zwischen Relevanz, Defizienz und Kontamination.

Leistungskriterien sind unvollkommene Annäherungen an das Konstrukt der Leistung, da sie tatsächliche Leistung nur unvollständig abbilden (Defizienz) und zusätzlich Irrelevantes erfassen (Kontamination).

[1] Die *inhaltliche* Bestimmung der Leistungsdimensionen oder -indikatoren ist eine in erster Linie anforderungsanalytische Fragestellung (vgl. Kapitel 3).

So gut wie alle Kriterien, ob „hart" oder „weich", sind in mehr oder weniger starkem Maße defizient und kontaminiert, aber auch relevant, sofern sie nicht völlig unsinnig sind. Das Ausmaß, in dem diese „Kriterien für Kriterien" erfüllt sind, ist allerdings häufig schwer zu quantifizieren und wird in der Literatur recht unterschiedlich operationalisiert (vgl. Bernardin & Beatty, 1984, und Borman, 1991, für Übersichten und weitere Gütemaße für Kriterien). Weiter unten in diesem Kapitel werden noch einige etwas konkretere Maße der Urteilsqualität diskutiert.

Leistung läßt sich auf den Ebenen des Potentials, des Verhaltens und der Ergebnisse beschreiben.

Kriterien können sehr unterschiedliche Gestalt annehmen. Ein besonders wichtiges unter den zahlreichen diskutierten Ordnungsprinzipien zu deren inhaltlicher Klassifikation ist die Unterscheidung nach *Beschreibungsebenen*; im einfachsten Fall aufgeteilt nach den Bereichen *Potential* (Eigenschaften, Fähigkeiten, Kenntnisse etc.), *Verhalten* und *Ergebnisse* (siehe ausführlich: Schuler, 1989). Verstanden wird diese Dreiteilung häufig als Kausalkette, wobei die unterschiedliche Ausprägung des Potentials (Bsp.: Fr. Müller ist in hohem Maße belastbar) Unterschiede bei der Prozeßvariable „Verhalten" bedingt (Fr. Müller arbeitet unter Termindruck effizient), was schließlich zu unterschiedlichen Ergebnissen führt (Fr. Müller hält Termine grundsätzlich ein). Dieser theoretisch elegante Zusammenhang leidet allerdings in der Praxis unter erheblichen Defizienz- und Kontaminationsproblemen auf allen drei Ebenen sowie unter der Tatsache, daß Menschen diese Bereiche bei ihrer Urteilsbildung und -abgabe nicht immer wirklich unterscheiden.

Wir hatten im Beispiel mit den Umsatzerlösen eines Vertreters bereits ein *Ergebniskriterium* kennengelernt, das für sich den Vorzug der *Objektivität* – im Sinne der Unabhängigkeit von einem Beurteiler – in Anspruch nehmen kann. Dies war, trotz der angesprochenen Probleme, insofern ein untypisches Beispiel, als sich objektive Maße von vergleichbarer oder gar größerer Relevanz nur in Ausnahmefällen finden lassen (z. B. bei Akkordarbeitern). Universell anwendbare objektive Kriterien wie bspw. Fehlzeiten erfassen nur einen sehr kleinen Ausschnitt der relevanten Arbeitsleistung und sind damit hochgradig defizient. Versucht man andererseits, schwer beobachtbare Leistungen (z. B. eines Topmanagers) an weniger defizienten objektiven Kriterien (z. B. dem Unternehmenserfolg) zu messen, so erkauft man dies i.d.R. mit einer erheblichen Kontamination. Ergebnisse werden deshalb häufig nicht objektiv gemessen, sondern, ebenso wie meist auch die Kriterien der beiden anderen Ebenen, *subjektiv* beurteilt, womit sie allerdings anfällig für alle potentiellen Fehlerquellen des Urteilsprozesses werden. Ein typischer Anwendungsfall ist bspw. die Beurteilung von Erreichungsgraden für (zuvor ausgehandelte) Ziele im Rahmen des Management by Objectives (vgl. unten). Wesentliche Vorzüge von Ergebniskriterien sind die augenscheinliche (oft aber nur scheinbare) Nähe zum Letztkriterium, den Zielen der Organisation, und, besonders bei objektiven Kriterien, deren Unpersönlichkeit. Letzteres kann Beurteilungsgesprächen viel von ihrem Konfliktpotential nehmen, allerdings auf Kosten der Ursachenanalyse. Die Meßgenauigkeit (Reliabilität) ist dagegen – entgegen der intuitiven Einschätzung – nicht besser als die von Verhaltens- oder Eigenschaftsmaßen (Visweswaran, 1993), was v.a. mit der Instabilität von Arbeitsergebnissen zusammenhängt. Als Hauptproblem wird jedoch oft die Beeinflußbarkeit durch Umstände außerhalb der Person angesehen, weshalb manche Autoren (z. B. Campbell, McCloy, Oppler & Sager, 1993) die Verwendung von Ergebnismaßen als Leistungskriterium ablehnen, während v.a. Betriebswirte insbesondere Zielerreichungsgraden als Kriterium positiv gegenüberstehen (z. B. Becker, 1994).

Ergebnisse liegen nahe am Letztkriterium, sind aber häufig sehr stark kontaminiert.

Größere Unterstützung finden dagegen *Verhaltensmaße*, die oftmals als einzig rechtfertigbares Kriterium für individuelle Leistung angesehen werden. Tatsächlich bildet die Beobachtung von Arbeitsverhalten die geeignetste Basis für ursachenorientiertes Feedback, was in Verbindung mit verhaltensorientierter Zielsetzung eine sehr wirksame Methode der Verhaltenssteuerung ist (vgl. Kapitel 13). Als Grundlage gezielter Personalentwicklung (vgl. Kapitel 9 und 10) sind Verhaltenskriterien unverzichtbar. Die Hoffnung auf eine erhebliche Verbesserung der psychometrischen Eigenschaften von Be-

urteilungsinstrumenten hat sich dagegen nicht in dem erwarteten Ausmaß erfüllt. Dies liegt in erster Linie daran, daß Beurteiler zwischen Verhalten und Eigenschaften nicht wirklich unterscheiden, vielmehr oft Verhalten aus einmal gebildeten Eigenschaftsurteilen rekonstruieren, v.a. dann, wenn sie untrainiert sind und die Instrumente lediglich Eigenschaftszuschreibungen in Tätigkeitssätze umformulieren („arbeitet gewissenhaft" statt „Gewissenhaftigkeit"). Verhaltensnähere Urteile erfordern einen erheblichen Aufwand für Anforderungsanalyse (vgl. Kapitel 3), Verfahrenskonstruktion und Beurteilertraining (siehe unten). Zudem wird die Vorgabe eines „Idealverhaltens" und die damit verbundene Kontrolle oftmals, besonders von erfahrenen Mitarbeitern und Führungskräften, als einengend und nicht mehr zeitgemäß empfunden.

Verhaltenskriterien eignen sich als Grundlage von Feedback, sind aber oft von Eigenschaftsbeurteilungen kaum zu trennen.

Die *Beurteilung von Eigenschaften* wie Leistungsbereitschaft oder Gewissenhaftigkeit wird in der Literatur mit bemerkenswerter Einhelligkeit abgelehnt (eine Ausnahme hiervon: Kavanagh, 1971). Wichtigste Kritikpunkte sind der Vorwurf, einer unkontrollierten Laienpsychologie Vorschub zu leisten sowie die mangelnde Relevanz von Persönlichkeitseigenschaften für das eigentliche Leistungskonstrukt. Dessenungeachtet sind Eigenschaftsbeurteilungen in der Praxis nach wie vor stark verbreitet, vor allem dort, wo die Inhalte der Leistungsbeurteilung tarifvertraglich festgelegt werden, was einen hohen Allgemeinheitsgrad bedingt und damit die Anwendung von spezifischen Verhaltenskriterien praktisch unmöglich macht. Vom praktischen Standpunkt spricht für Persönlichkeits- und andere Potentialmaße, daß elaborierte Verhaltensskalen bei erheblich größerem Entwicklungsaufwand nur wenig anderes erfassen als einige sehr allgemeine Fähigkeiten und Eigenschaften, die zudem über verschiedene Beurteiler generalisierbar sind (siehe Kasten 1).

Kasten 1:
Persönliche Konstrukte beruflicher Leistung

Borman (1987) untersuchte bei 25 erfahrenen Vorgesetzten von Angehörigen unterschiedlichster Bereiche der U.S. Army, welche „personal constructs" von beruflicher Leistung letztlich in ihre Beurteilungen einfließen. Er verwendete in dieser methodisch anspruchsvollen Studie ein arbeitsanalytisch fundiertes (mit Hilfe der Critical Incident Technique, Flanagan, 1954), verhaltensnahes Beurteilungsinstrument. Fast drei Viertel der Varianz in den Beurteilungen ließ sich durch sechs allgemeine, *nicht* beurteilerspezifische Faktoren erklären (sinngemäß übersetzt): *Arbeitsinitiative, Verantwortlichkeit, technische Fähigkeiten, Organisation, mitarbeiterorientierte Führung, aufgabenorientierte Führung.*

Unverzichtbar sind Potentialbeurteilungen bei Personalentscheidungen mit Prognosecharakter wie der über den Wechsel in eine andere Position und bei der langfristigen Personalplanung. Solche originär eignungsdiagnostischen Aufgaben lassen sich aber durch die Adaption von Auswahlverfahren wie Assessment Center oder Persönlichkeits-, Fähigkeits- und Kenntnistests (vgl. Kapitel 5 und 6) oft besser unterstützen als durch klassische Vorgesetztenbeurteilungen.

Eigenschaftskriterien erfassen eher Potentiale als eigentliche Leistung, eignen sich aber gerade deshalb für Prognosen.

Merke:

Leistung läßt sich auf den Ebenen des Potentials, des Verhaltens und der Ergebnisse beschreiben, wobei keiner dieser Bereiche generell vorzuziehen oder abzulehnen ist. Vielmehr kommt es hier auf einen jeweils zweckentsprechenden Einsatz an.

2.2 Die Struktur beruflicher Leistung

Lange Zeit (und teilweise auch heute noch) ging die Lehrmeinung zur Struktur beruflicher Leistung dahin, daß diese ein multidimensionales (lies: un-

überschaubares) Konstrukt sei. Die mit großer Hartnäckigkeit auftretenden Überschneidungen als unabhängig postulierter Urteilsdimensionen wurden in den Bereich menschlicher Unzulänglichkeit (Halo) verwiesen. Zudem führt die (in vielerlei Hinsicht verdienstvolle) anforderungsanalytische Orientierung der Personalpsychologie dazu, Leistung als etwas für jeden Arbeitsplatz spezifisch zu Definierendes zu betrachten. Solche Vielfalt ist aber praktischen Zwecken nicht immer dienlich. Bereits Schmidt und Kaplan (1971) kamen zu dem Schluß, daß die gesonderte Betrachtung einzelner Urteilsdimensionen (*multiple criteria*) zwar für ein tieferes wissenschaftliches Verständnis durchaus hilfreich ist, praktische Entscheidungen aber letztlich ein globales Leistungsmaß erfordern, ermittelt über eine gewichtete, lineare Kombination einzelner Elemente (*composite criterion*). Erst in jüngster Zeit wurden allerdings systematische Versuche unternommen, zu generellen empirischen und theoretischen Aussagen über das Konstrukt beruflicher Leistung zu kommen. Die Ergebnisse sprechen dafür, daß Leistung zwar i.d.R. mehrdimensional, aber keineswegs unüberschaubar komplex ist (manche Beurteilungssysteme in der Praxis enthalten 20 oder mehr „Dimensionen") und daß das Ausmaß an Generalisierbarkeit über verschiedene Arbeitsplätze beträchtlich ist. Wir beschränken uns im folgenden auf die Darstellung einer großen empirischen Studie und zweier, mehr oder weniger darauf aufbauender theoretischer Ansätze.

Die Komplexität des Leistungskonstrukts wird häufig überschätzt.

In den achtziger Jahren wurde bei den U.S.-Streitkräften eine der umfangreichsten personalpsychologischen Studien überhaupt durchgeführt: das *U.S. Army Selection and Classification Project* (*Project A*) (vgl. Themenheft von *Personnel Psychology, 43 (2)*). Neben der Entwicklung und Validierung von Auswahl- und Klassifikationsverfahren waren weitere, für das Thema dieses Kapitels einschlägige Ziele: die Entwicklung organisationsweiter Leistungskriterien, die Identifizierung der Konstrukte, die das Universum berufsrelevanter Informationen ausmachen und die Entwicklung eines allgemeinen Modells beruflicher Leistung in Positionen für qualifizierte Berufsanfänger (Campbell, 1990a). Der Umfang der dafür gesammelten Daten ist enorm: Die wichtigsten Berechnungen beruhen auf einer Stichprobe von N = 9.430 aus 19 verschiedenen Berufen, bei der mehr als 200 Leistungsindikatoren und eine (nicht nur im übertragenen Sinne) erschöpfende Prädiktorbatterie erhoben wurde. In einem mehrstufigen Prozeß exploratorischer und konfirmatorischer Faktorenanalysen wurde schließlich ein generelles, also positionsübergreifendes Modell beruflicher Leistung mit sehr guter Anpassung – d.h. Modell und Daten paßten zusammen – identifiziert. Danach setzt sich die Leistung der untersuchten Armeeangehörigen aus fünf inhaltlich interpretierbaren Faktoren zusammen (Campbell, McHenry & Wise, 1990; daneben fanden sich methodenspezifische Faktoren):

In einer umfangreichen Studie der U.S. Army (Project A) wurde eine fünffaktorielle Struktur des Leistungskonstrukts gefunden.

I. *Tüchtigkeit bei der Erfüllung wesentlicher Aufgaben:* Ausmaß, in dem das Individuum *fähig* ist, den zentralen Anforderungen der *spezifischen Position* gerecht zu werden
II. *Allgemeine soldatische Tüchtigkeit:* Ausmaß, in dem das Individuum generelle, *positionsübergreifende* Aufgaben in der Armee bewältigen kann
III. *Einsatz und Führung:* Ausmaß, in dem das Individuum *bereit* ist, sich bei allen Aufgaben zu engagieren und andere bei ihrer Aufgabenerfüllung zu unterstützen
IV. *Persönliche Disziplin:* Ausmaß, in dem sich das Individuum an militärische Regeln hält, Selbstkontrolle und Integrität im alltäglichen Umgang demonstriert
V. *Körperliche Fitneß und äußere Erscheinung*

Die Faktoren waren – mit Ausnahme des letzten – nach Auspartialisierung der Methodenvarianz untereinander moderat korreliert (r zwischen .20 und .45). Es bleibt zu fragen, inwieweit sich diese Ergebnisse über den militärischen Bereich hinaus generalisieren lassen; akkurat gebügeltem Hemd über gestähltem Bizeps (Faktor V) mag im Zivilleben nicht immer ein einheitli-

cher Wert beigemessen werden. Hunt (1996) fand bei einer noch größeren Stichprobe von Einzelhandelsverkäufern acht Faktoren, bei denen zumindest die Themen Einsatzbereitschaft und Disziplin – hier aufgespalten in mehrere Subfaktoren – aus dem Project A wiederkehrten. Die hier teilweise sehr hohe Interkorrelation der Faktoren legt aber eine einfachere Struktur nahe.

Campbell und Kollegen (Campbell, 1990b; Campbell et al., 1993) entwikkelten, aufbauend auf den Project A-Daten, eine generelle *Theorie beruflicher Leistung* (siehe Abbildung 2). Darin werden acht unabhängige Leistungskomponenten spezifiziert, die, in unterschiedlicher Gewichtung und inhaltlicher Ausdifferenzierung, das Universum menschlicher Leistung im Beruf abdecken. Campbell et al. (1993) legen Wert auf die Feststellung, daß sie mit „Leistung" ausschließlich beobachtbares Arbeitsverhalten, nicht aber z. B. Arbeitsergebnisse meinen. Potentiale spielen dagegen als *Determinanten* der Leistung eine zentrale Rolle in der Theorie. Hier wird eine *multiplikative* Verknüpfung der Elemente deklaratives (statisches) Wissen, prozedurales Wissen und Fertigkeiten („gewußt wie") sowie der Motivation angenommen, was bedeutet, daß *jede* dieser Determinanten wenigstens zu einem Mindestmaß vorliegen muß, damit eine Leistung zustandekommt. Gleichzeitig wird die Aufzählung aber als erschöpfend postuliert; es existieren also keine relevanten individuellen Unterschiede über die drei genannten Determinanten hinaus, die ihrerseits aber wieder jeweils eine Funktion vorgelagerter Ursachen sind. Über die empirische Bewährung der Theorie läßt sich gegenwärtig noch wenig berichten. Eine von den Urhebern selbst vorgenommene, konfirmatorische (modellprüfende) Analyse von Project A-Daten (McCloy, Campbell & Cudeck, 1994) fand einige Bestätigung für den Drei-Determinanten-Teil der Theorie, wobei Motivation die bei weitem bedeutsamste Komponente war. Die Annahmen über das Zustandekommen dieser Determinanten erscheinen dagegen noch recht spekulativ und – mit z.T. identischen oder beliebigen Argumenten in den Funktionen – auch konzeptionell weniger durchdacht.

Campbell et al. (1993) sehen Leistung durch die Interaktion dreier Determinanten bestimmt.

Eine noch stärker auf den Inhalt beruflicher Leistung konzentrierte Theorie stammt von Borman und Motowidlo (1993). Diese Autoren profitierten gleichfalls von den vielfältig analysierbaren Project A-Daten, bauen aber auch auf den Konzepten *prosozialen Verhaltens* in Organisationen (Brief &

PC_i = **DK** (Fakten, Prinzipien, Ziele, Selbstkenntnis) x **PKS** (kognitive, physische, psychomotorische und soziale Fertigkeiten; „Selbstmanagement") x **M** (Leistungsentscheidung, Anstrengungsniveau und Ausdauer)

PC = performance components (Leistungskomponenten)
i = Laufindex von PC (von 1 bis 8); im einzelnen:
1 = positionsspezifische Tüchtigkeit zur Aufgabenerfüllung
2 = generelle Tüchtigkeit zur Aufgabenerfüllung
3 = Kommunikationsfähigkeit (schriftlich und mündlich)
4 = Anstrengung (Ausmaß und Konstanz)
5 = persönliche Disziplin
6 = Kooperation und Unterstützung von Kollegen / der Arbeitsgruppe
7 = Mitarbeiterführung
8 = Management/Administration (nicht direkt personenbezogene Führungsaufgaben wie Organisation)
DK = declarative knowledge = *f* (Fähigkeiten, Persönlichkeit, Interessen, Ausbildung etc.)
PKS = procedural knowledge and skill = *f* (Fähigkeiten, Persönlichkeit, Interessen, Ausbildung, Erfahrung, Übung etc.)
M = motivation = *f* (Operationalisierung von der gewählten Motivationstheorie abhängig)

Deklaratives Wissen, prozedurales Wissen und Fertigkeiten sowie Motivation sind in Campbell et al.'s Theorie multiplikativ verknüpft.

Abbildung 2:
Campbell et al.'s (1993) Theorie beruflicher Leistung

Borman und Motowidlo (1993) unterscheiden zwischen reiner Aufgabenleistung und einem auf das Umfeld der Arbeitsaufgaben bezogenen Bereich.

Motowidlo, 1986) und des *organizational citizenship behavior* (Organ, 1988) auf. Kern der Theorie ist der Gedanke, daß neben der reinen, „technischen" Aufgabenerfüllung, die sich anforderungsanalytisch *positionsspezifisch* fundieren und schließlich in Stellenbeschreibungen fixieren läßt, ein wichtiger *genereller* Bereich beruflicher Leistung existiert, der über die *aufgabenbezogene Leistung* (task performance) hinausgeht, gänzlich andere Ursachen hat und deshalb in Arbeitsanalysen gewöhnlich übersehen wird. Borman und Motowidlo (1993) prägten dafür den Begriff der „*umfeldbezogenen Leistung*" (contextual performance). Die Autoren unterscheiden hier weiter zwischen der Leistung von Führungskräften und Nicht-Führungskräften. In Abbildung 3 werden die wesentlichen Inhalte und Unterschiede von aufgaben- und kontextbezogener Leistung zusammengefaßt:

	aufgabenbezogene Leistung	**umfeldbezogene Leistung**
Kennzeichnung:	Tätigkeiten, die zum formalen Gegenstand der Arbeit gehören; direkt ergebnisbezogen; jobspezifisch	Tätigkeiten, die über die formalen Arbeitsinhalte hinausgehen; indirekt ergebnisunterstützend; allgemeingültig
Beispiel:	fehlerfreier Einbau eines Armaturenbretts	Unterstützung ungeübter Kollegen
Inhalt:	durch Arbeitsanalyse / Stellenbeschreibung festgelegt	*bei Führungskräften:* – Bindung an die Organisation (commitment) – Aufrechterhaltung guter Arbeitsbeziehungen – Repräsentation der Organisation nach außen – Anstrengung/Ausdauer b. d. Zielverfolgung *bei Nicht-Führungskräften:* – Bindung an die Organisationsziele – Bereitschaft zu freiwilligem Einsatz über den formalen Arbeitsinhalt hinaus – Kooperation und Unterstützung anderer – Einhaltung von Regeln der Organisation – Anstrengung/Ausdauer b. d. Aufgabenerfüllung
Ursachen/ Determinanten:	Kenntnisse, Fähigkeiten und Fertigkeiten, Erfahrung	Persönlichkeit und Motivation

Abbildung 3:
Borman & Motowidlos (1993) Konzept der aufgaben- vs. umfeldbezogenen Leistung

Neben positionsspezifischen existieren auch universelle Komponenten beruflicher Leistung.

Erste empirische Untersuchungen unterstützen die These unterschiedlicher Prädiktor- / Kriteriumsbeziehungen für die beiden Leistungsbereiche (Borman & Motowidlo, 1997; Hattrup, O'Connell & Wingate, 1998; Motowidlo & VanScotter, 1994) und liefern marginale Unterstützung für die Konstruktvalidität der Unterscheidung; d.h. Kriterien *innerhalb* der task- bzw. contextual-Domäne korrelieren etwas höher (um .70) als jene *zwischen* diesen Dimensionen (um .55; Conway, 1996). Auch hier wird sich erst in Zukunft ein deutlicheres Bild abzeichnen.

In die Bemühungen um das theoretische und empirische Verständnis des Konstrukts „berufliche Leistung" ist also seit Beginn der neunziger Jahre erhebliche Bewegung gekommen. Dazu zählen bspw. auch die von der metaanalytischen Schule der Psychometrie geprägten pfadanalytischen *Kausalmodelle*, die jedoch die Prädiktorseite stärker betonen (siehe z. B. Schmidt & Hunter, 1992), oder die Unterscheidung von *Job-* und *Rollenelementen* (Ilgen & Hollenbeck, 1991), die der Theorie von Borman und Motowidlo ähnelt. Zum gegenwärtigen Zeitpunkt läßt sich daraus allenfalls ein Zwischenfazit ziehen: Berufliche Leistung scheint sich, auch wenn sie durch vielfältige Indikatoren erfaßt wird, durch eine durchaus überschaubare Anzahl von Dimensionen beschreiben zu lassen, die untereinander in mäßiger bis moderater Höhe korreliert sind. Einige Bereiche haben dabei eine generelle Bedeutung für einen großen Teil aller beruflichen Positionen.

3 Funktionen der Leistungsbeurteilung

Mit Beurteilungssystemen wird in Leistungsorganisationen eine Vielzahl von Zielen (Zwecken, Funktionen) verfolgt. Für deren Klassifikation wurde in der Literatur eine Reihe dichotomer Ordnungsprinzipien vorgeschlagen. Diese unterscheiden u.a. zwischen *individuellen* (Beurteiler *und* Beurteilte) und *organisationalen* Zielen (Ilgen, 1993), zwischen *vergangenheits-* und *zukunftsgerichteter* Betrachtung (Leistungs- vs. Potentialbeurteilung), zwischen *manifesten* (offenen) und *latenten* (unausgesprochenen) Zwecken oder zwischen *personalpolitischen* (administrativen) und *führungspolitischen* (Verhaltenssteuerungs-) Zielen, wobei hier in den letzten Jahren eine Schwerpunktverlagerung vom ersten auf den zweiten Bereich zu konstatieren ist (Domsch & Gerpott, 1992; Murphy & Cleveland, 1995). Schuler (1989) nennt dagegen zehn Funktionsbereiche von Leistungsbeurteilungen, deren Unterscheidung sich vorwiegend daran orientiert, welche Zwecke aus Gründen praktischer Unvereinbarkeit auseinandergehalten werden *sollten*. Eine ähnliche Intention verfolgen Cleveland, Murphy, und Williams (1989), deren Einteilung von ursprünglich zwanzig einzelnen Funktionen in vier große Bereiche (Tabelle 1) für sich in Anspruch nehmen kann, in einer konfirmatorischen Faktorenanalyse empirisch bestätigt worden zu sein.

Mit Leistungsbeurteilungen werden zahlreiche Ziele verfolgt.

Tabelle 1:
Funktionen der Leistungsbeurteilung (nach Cleveland, Murphy und Williams, 1989)

1. interpersonale Entscheidungen:
(z. B. Entgeltfindung, Beförderung, Kündigung auf der Grundlage unzureichender bzw. herausragender Leistungen)
2. intrapersonale Entscheidungen:
(z. B. Feedback, Verhaltenssteuerung, Beratung; Identifizierung individueller Stärken und Schwächen als Grundlage von Personalentwicklungsmaßnahmen)
3. Erhaltung des (Organisations-) Systems:
(z. B. Personalplanung, Planung des Organisations- und Personalentwicklungsbedarfs, Evaluation der Personalpolitik und von Zielerreichungsgraden, Erhaltung der Autoritätsstruktur)
4. Dokumentation:
(z. B. Kriterien für Validierungsstudien, Dokumentation personeller Entscheidungen und deren Begründung gemäß rechtlicher und tariflicher Anforderungen)

All diese Funktionen lassen sich durch geeignete Gestaltung von Beurteilungssystemen zumindest sinnvoll unterstützen, allerdings – das gilt es festzuhalten – nicht alle auf einmal. Ein wichtiges Ergebnis anwendungsorientierter Forschung zur Leistungsbeurteilung besteht in der Einsicht, daß einige ihrer Ziele nicht miteinander vereinbar sind oder zumindest konfligieren. Der stärkste Konflikt besteht dabei zwischen administrativen Entscheidungen wie Gehaltsbestimmung oder Beförderung, bei denen die Verteilung begehrter und knapper Ressourcen von der Günstigkeit der Beurteilung abhängt, und persönlichen Entwicklungszielen aus dem Komplex Beratung, Förderung und Verhaltensfeedback, deren Erreichung entscheidend von der Freimütigkeit des Eingestehens auch von Schwächen und einer offenen, vertrauensvollen Atmosphäre im Personalgespräch profitiert. Beurteilungssysteme, bei denen versucht wird, an dieser Stelle zwei Fliegen mit einer Klappe zu schlagen (was in der Praxis nicht ungewöhnlich ist, vgl. Cleveland et al., 1989), legen ihr Scheitern – oder zumindest erhebliche Schwierigkeiten –

Der Kardinalfehler vieler LB-Systeme in der Praxis besteht darin, unvereinbare Ziele mit einem einzigen System erreichen zu wollen.

bereits im Keim an. Wenn also in den folgenden Abschnitten von Quellen, Verfahren und Qualitätsmaßstäben der Beurteilung die Rede sein wird, so ist stets zu fragen, für welchen *Zweck* sich deren Anwendung empfiehlt oder nicht. Die entsprechende Einschätzung kann dabei von Fall zu Fall recht unterschiedlich ausfallen.

4 Quellen der Beurteilung

4.1 Objektive Quellen

Objektive Leistungsdaten sind zwar unabhängig vom Beurteiler, oft aber auch hochgradig defizient und kontaminiert.

Wenn es um die Auswahl von Leistungsindikatoren geht, liegt zunächst der Rückgriff auf Quellen wie das Rechnungswesen oder Produktionsaufzeichnungen nahe, die weitgehend frei von persönlichen Einschätzungen sind. Neben rein quantitativen Arbeitsergebnissen wie Stückzahlen oder Umsatzvolumina läßt sich auch die Arbeitsqualität in manchen Fällen über zählbare Indikatoren wie Ausschußquoten oder Kundenbeschwerden indirekt beurteilen. Wie wir in Abschnitt 2.1 gesehen haben, werden hier jedoch die Vorzüge der Einfachheit und Objektivität oft dadurch überkompensiert, daß die Relevanz in keinem Verhältnis zu Defizienz und Kontamination steht. Vielfach, besonders bei komplexen Tätigkeiten, läßt sich Leistung kaum objektiv messen. Die Entwicklung der Mikroelektronik hat es jedoch in den letzten Jahren möglich gemacht, Produktivität und teilweise auch objektives Verhalten in dem großen und wachsenden Bereich der Computerarbeitsplätze minutengenau und detailliert zu erfassen (*Electronic Performance Monitoring*, EPM). Erste Erfahrungen mit dieser elektronischen Leistungsüberwachung zeigen allerdings, daß EPM nicht unbedingt zu Leistungsverbesserungen, ziemlich sicher aber zu erheblichem Streß führt (Aiello & Kolb, 1995). Die Einführung solcher Systeme kann sich also durchaus kontraproduktiv auswirken.

Wo objektive Kriterien vorliegen, beträgt ihre Korrelation mit subjektiven Beurteilungen, metaanalytisch geschätzt (Bommer, Johnson, Rich, Podsakoff & MacKenzie, 1995), im Bereich um .40. Eine andere Metaanalyse (Viswesvaran, 1993) kam zu einer höheren Schätzung (.57), wobei beide Bereiche substantiell auf einem Faktor höherer Ordnung luden. Objektive Leistungsindikatoren können also, wo sinnvoll erhebbar, einen Beitrag zur Leistungsmessung liefern. Ihre alleinige Verwendung scheint jedoch nur in Ausnahmefällen angezeigt. Die Freiheit von subjektiven Urteilen legt aber ihre – zumindest ergänzende – Anwendung als Grundlage der Bestimmung leistungsabhängiger Entgeltanteile nahe.

4.2 Subjektive Quellen

Menschen können Leistung umfassender beurteilen, sind aber in ihren Einschätzungen immer subjektiv.

Menschliche Beurteiler sind zumindest grundsätzlich in der Lage, durch objektive Indikatoren nicht erfaßte Bereiche (Defizienz) und äußere Umstände, die die Leistung beeinflussen (Kontamination), in ihren Einschätzungen zu berücksichtigen. Vor dem Hintergrund dieser Plausibilitätsüberlegung überraschend ist allerdings der empirische Befund (Steel & Mento, 1986), daß situative Bedingungen stärker in *subjektive* Beurteilungen einfließen als in objektive Indikatoren, erstere also stärker kontaminiert waren. Unzweifelhaft ist dagegen, daß sich viele Aspekte menschlicher Leistung überhaupt nur durch menschliche Einschätzungen erheben lassen.

Die bei weitem wichtigste Quelle subjektiver Beurteilungen ist nach wie vor der *direkte Vorgesetzte*. Dies entspricht der Konvention in hierarchischen Organisationen und wird im übrigen sowohl von Beurteilern wie von Beurteilten bevorzugt (vgl. Murphy & Cleveland, 1995, für eine ausführliche Diskussion dieses und der folgenden Aspekte). Ein Problem von Vorgesetz-

tenbeurteilungen[2], das sich in flachen Hierarchien mit zunehmenden Kontrollspannen in Zukunft tendenziell verstärken wird, ist die oft mangelnde Gelegenheit zu direkten Verhaltensbeobachtungen. Hinzu kommt, zumal bei hochqualifizierten Tätigkeiten wie in der Forschung und Entwicklung, die nicht immer ausreichende Vertrautheit mit den unmittelbaren Arbeitsaufgaben. Eine Schwierigkeit, die sich weniger auf die *Fähigkeit* als auf den *Willen* zur Abgabe korrekter Urteile bezieht, ist die Involviertheit des Vorgesetzten in die Folgen der Beurteilung. Vorgesetzte haben ein vitales Interesse daran, die Motivation ihrer Mitarbeiter nicht zu gefährden und gute persönliche Beziehungen aufrechtzuerhalten. Ferner wird die Leistung des Vorgesetzten selbst oft an den Beurteilungen seiner Gruppe gemessen; er oder sie möchte vielleicht geschätzte Mitarbeiter nicht durch Beförderung verlieren oder andere nur zu gern in andere Abteilungen „wegloben". Solche Ursachen für *mikropolitisches Verhalten* beeinflussen die Urteilsabgabe möglicherweise stärker als Mängel im Prozeß der Informationsaufnahme und -verarbeitung, obwohl die Evidenz hierfür bislang entweder anekdotischer (z. B. Longenecker, Sims & Gioia, 1987) oder indirekter Natur ist. So scheinen für reine Forschungszwecke abgegebene Urteile korrekter auszufallen als solche mit Folgen für administrative Entscheidungen. Dennoch wird der direkte Vorgesetzte in absehbarer Zeit kaum als Hauptquelle der Beurteilung abzulösen sein, deren Ergänzung von anderer Seite jedoch für manche Zwecke sinnvoll ist.

Trotz einiger Probleme sind direkte Vorgesetzte die weitaus wichtigste Urteilsquelle, sollten aber nicht die einzige sein.

Eine potentielle Urteilsquelle, die häufiger Gelegenheit zur Verhaltensbeobachtung hat als der direkte Vorgesetzte und der in manchen Bereichen auch größere Sachkompetenz zugetraut wird, sind unmittelbare Arbeitskollegen bzw. *Gleichgestellte*. Zudem läßt sich hier i.d.R. über mehrere Urteile aggregieren, was Reliabilitätsvorteile bedingt (die *individuellen* Urteile sind aber bei Kollegen im Mittel weniger reliabel [Viswesvaran, Ones & Schmidt, 1996]). Ein häufig erhobener Einwand gegen Gleichgestelltenbeurteilungen ist der vermutete Einfluß von Sympathie und anderen Affekten auf die Urteilsabgabe. Solche *affektiven Komponenten* wurden allerdings mglw. vorschnell als *Fehlerquelle* qualifiziert (Varma, DeNisi & Peters, 1996), und es besteht wenig Anlaß zu der Überzeugung, daß sie sich bei Kollegen stärker auswirken als bei anderen menschlichen Beurteilern. Potentiell gravierendere Probleme der Kollegenbeurteilung als deren mögliche psychometrische Mängel erwachsen aus dem Widerspruch zu hierarchischen Konventionen, den möglichen Rollenkonflikten („Kumpel" vs. „Richter") und den sich hieraus ergebenden Folgen für das Arbeitsklima innerhalb des Teams. Obwohl anekdotische Berichte über gute Akzeptanz seitens der *Beurteilten* vorliegen (Jochum, 1991), fühlen sich Gleichgestellte als *Beurteiler* wohler, wenn ihre Urteile ausschließlich Beratungs- und Förderungszwekken dienen und wenn sie lediglich einzelne, herausragende Kollegen nennen (peer nomination) und keine komplette Rangordnung (peer ranking) aufstellen müssen.

Die qualitativen und psychometrischen Vorzüge der Gleichgestelltenbeurteilung sollten gegen deren Konfliktpotential abgewogen werden.

Eine in letzter Zeit stärker beachtete Quelle der Beurteilung sind unterstellte *Mitarbeiter*. Der potentielle Beitrag von Mitarbeiterbeurteilungen liegt v.a. in deren unterschiedlicher Perspektive im Vergleich zu Vorgesetzten, wobei der Fokus weniger auf Ergebnisse und Sachaufgabenerfüllung als auf interpersonale Aspekte der Mitarbeiterführung gerichtet ist, wo sie eine kaum ersetzbare Quelle darstellen. Funktion von „Aufwärtsbeurteilungen" ist praktisch ausschließlich Verhaltensfeedback („upward feedback" ist ein gebräuchliches Synonym); deshalb werden sie *für diesen Zweck* von den Beurteilten auch kaum als Bedrohung der Autoritätsstruktur wahrgenommen (Bernardin, Dahmus & Redmon, 1993), obwohl nicht auszuschließen ist, daß solche Urteile – zumindest implizit – auch als Grundlage administrativer Ent-

[2] Terminologisch wählen wir hier jeweils die Perspektive des *Beurteilers*. In der Literatur findet sich häufig auch die umgekehrte Konvention, also bspw. der Begriff Vorgesetztenbeurteilung, wenn es um die Beurteilung von Vorgesetzten *durch* ihre Mitarbeiter geht. Weniger mißverständlich sind die v.a. bei Betriebswirten gebräuchlichen Begriffe Abwärts-, Seitwärts- und Aufwärtsbeurteilung.

scheidungen dienen. Die Qualität und Bereitschaft zur Abgabe von Mitarbeiterurteilen steht und fällt mit der Anonymität der Urteilsabgabe, weshalb sie i.d.R. in der Form schriftlicher Befragungen erhoben werden. Erfahrungen mit zunächst anonym durchgeführten Beurteilungen dürften allerdings auch die Bereitschaft zu offenem Feedback fördern. Auf Anonymität und psychometrische Qualität wirken sich, im Gegensatz zu Vorgesetztenbeurteilungen, steigende Kontrollspannen vorteilhaft aus.

Anonym erhobene Beurteilungen unterstellter Mitarbeiter können eine wichtige Feedbackfunktion für Führungskräfte erfüllen.

Eine Quelle mit konkurrenzlos direktem und umfassendem Zugang zum tatsächlichen Verhalten ist *der Beurteilte selbst*. Die unmittelbare Betroffenheit des Beurteilers von seinem eigenen Urteil macht Selbstbeurteilungen aber auch in besonderer Weise anfällig für absichtliche und unwillkürliche Verzerrungen, was sich insbesondere in der Abgabe übermäßig vorteilhafter Urteile niederschlägt (das Ausmaß dieser „Mildetendenz" wurde in der Vergangenheit jedoch vermutlich überschätzt [vgl. Moser & Krauß, 1997]). Dieser Tendenz läßt sich durch verschiedene Maßnahmen entgegenwirken, z. B. durch die Ankündigung einer Überprüfung der Urteile oder die Beschränkung auf Entwicklungsziele (Donat, 1991), offenbar aber nicht, wie früher vermutet, durch die Wahrung der Anonymität (Moser & Krauß, 1997). In der Praxis ist das Problem übertriebener Selbstbeurteilungen schon deshalb weniger virulent, weil sie i.d.R. der Vorbereitung von Beurteilungsgesprächen dienen, wo eine allzu starke Abweichung von der Einschätzung des Vorgesetzten zu einem für beide Seiten peinlichen Rechtfertigungsdruck führen würde. Ein wesentlicher Vorzug von Selbstbeurteilungen besteht darin, daß sie i.d.R. differenzierter ausfallen als Fremdeinschätzungen. Dies, und die erwähnte Verhaltensnähe, prädestiniert Selbstbeurteilungen als Grundlage der Personalentwicklung. Für andere, insbesondere administrative Funktionen kommen sie dagegen allenfalls als ergänzende Diskussionsbasis in Frage.

Selbstbeurteilungen eignen sich vor allem als Grundlage der Personalentwicklung.

Gelegentlich werden auch Außenstehende als Quellen der Beurteilung herangezogen, v.a. *Kunden* bei Mitarbeitern, die mit diesen in Kontakt kommen. Gemeint sind damit nicht die im Einzelhandel gefürchteten „Kontrollkäufer", sondern i.d.R. punktuelle, schriftliche Befragungen von „Normalkunden". Obwohl Kunden eine eigenständige, von mikropolitischen Erwägungen weitgehend freie Perspektive beisteuern können, scheitert eine regelmäßige Erhebung meist an Problemen der Zugänglichkeit und Opportunitätserwägungen. Relativ verbreitet zur Kontrolle von Vorgesetztenbeurteilungen ist dagegen die Hinzuziehung von *Vorgesetzten höherer Hierarchieebenen*, wobei aber die Gelegenheiten zur unmittelbaren Verhaltensbeobachtung auf ein Minimum beschränkt sind, weshalb sich hier die „Beurteilung" nicht selten in einem Akt des Abzeichnens erschöpft.

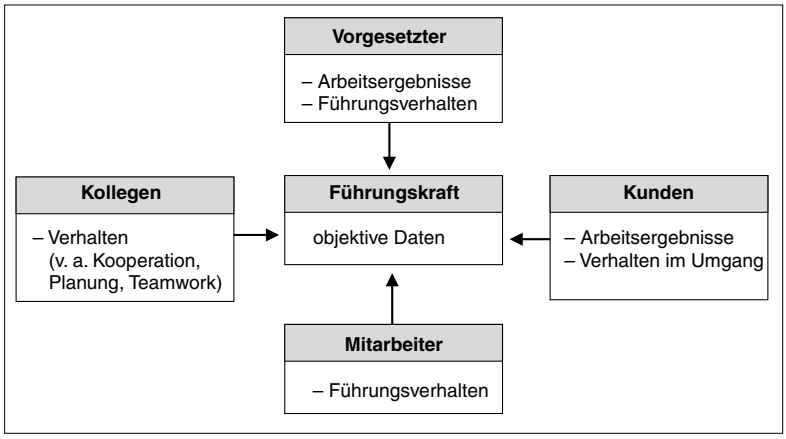
Auch Außenstehende und höhere Vorgesetzte kommen als Quelle der Beurteilung in Betracht.

Zur *Übereinstimmung* zwischen den verschiedenen Urteilsquellen liegt inzwischen eine Vielzahl von Einzelstudien und eine Reihe von Metaanalysen vor. Dabei zeigt sich regelmäßig, daß Fremdeinschätzungen unterschiedlicher Quellen, also z. B. Vorgesetztenurteile mit Kollegenbeurteilungen (Harris & Schaubroeck, 1988) und Mitarbeiterurteilen (z. B. Mount, 1984, in einer Einzelstudie), untereinander wesentlich höher (reliabilitätskorrigierte Differenz: r_{diff} um .30 bis .40) korrelieren als mit Selbstbeurteilungen (Harris & Schaubroeck, 1988; Mabe & West, 1982; Moser & Krauß, 1997; zusammenfassend Conway & Huffcutt, 1997). Die Stärke des Zusammenhangs zwischen Selbst- und Fremdbeurteilungen wird durch eine Reihe von Merkmalen der *Person*, des *Arbeitsplatzes*, des *Meßinstruments* und der *Durchführungsbedingungen* moderiert (Donat, 1991), so daß eine höhere Übereinstimmung durchaus erzielbar erscheint. Ob dies, im Sinne der Urteilsqualität, aber auch wünschenswert ist, ob also die Urteile der einen Quelle als Kriterium für die *Validität* der anderen herangezogen werden sollten – ein in der Literatur durchaus verbreitetes Vorgehen –, darf bestritten werden. Obwohl ein gewisses Maß positiver Beziehungen zwischen verschiedenen Indikatoren desselben Konstrukts sicher erwünscht ist, liegt der Vorzug unterschiedlicher Blickwinkel gerade darin, verschiedene (wenngleich tw. überlappende) Aspekte dieses Konstrukts zu erfassen. Ein v.a. in der Praktikerliteratur vielbeachteter Ansatz, der diese multiperspektivische Betrachtungsweise betont und zu integrieren versucht, wird in Kasten 2 beschrieben.

Fremdbeurteilungen unterschiedlicher Quellen stimmen stärker überein als Fremd- mit Selbstbeurteilungen. Vollständige Übereinstimmung würde aber auch vollständige Redundanz bedeuten.

Kasten 2:
360-Grad-Beurteilungen

Die Grundidee von 360-Grad-Beurteilungen ist frappierend einfach: Lasse die Leistung von Führungskräften (für Nicht-Führungskräfte wird das Verfahren wohl als zu aufwendig angesehen) aus allen erdenklichen Blickwinkeln einschätzen, wobei sich jede Quelle auf ihr zugängliche Dimensionen beschränken sollte, integriere diese Beurteilungen und melde sie dem betreffenden Manager zurück (siehe Abbildung):



Quellen und Inhalte des 360-Grad-Feedbacks (verändert nach London & Beatty, 1993, p. 355)

Vordringliches Ziel ist ein umfassendes individuelles Feedback und die damit verbundene Verhaltenssteuerung, obwohl auch organisationale Funktionen wie Personalplanung und Organisationsentwicklung unterstützt werden sollen. Ansätze zur Integration mehrerer Quellen hat es schon früher gegeben (in „Winkelterminologie": 180- oder 270-Grad-Beurteilungen); neu ist hier, neben dem griffigen Namen, die systematische Integration aller verwertbaren Quellen sowie die Existenz standardisierter (inzwischen auch deutschsprachiger) Beurteilungsinstrumente. Erste empirische Untersuchungen zeigen vielversprechende Befunde zur Effektivität des Verfahrens; ob sie allerdings über die Effekte traditionellen Feedbacks hinausgehen und den betriebenen Aufwand rechtfertigen, muß sich noch zeigen (vgl. zusammenfassend Dunnette, 1993, sowie die anderen Beiträge im Themenheft von *Human Resource Management, 32 (2 & 3)*; praktische Hinweise zur Einführung von 360-Grad-Beurteilungen gibt Hunt, 1995)

360-Grad-Feedback ist ein Ansatz zur Integration mehrerer Urteilsquellen.

5 Beurteilungsverfahren

Messungen erfordern Meßinstrumente; im Falle der beruflichen Leistung geht es dabei darum, eine Abbildungsvorschrift für eines (sog. s*ummarische* Beurteilung) oder mehrere (*analytische* Beurteilung) Leistungskriterien zu konstruieren[3]. Über Jahrzehnte war ein Großteil der psychologischen Forschung zur Leistungsbeurteilung auf die Entwicklung und Verbesserung formaler Beurteilungsverfahren gerichtet. Wie wir später sehen werden, haben diese Bemühungen nicht immer die erhofften Früchte getragen. Dieser Abschnitt beschränkt sich weitgehend auf die *Beschreibung* der wichtigsten Verfahren. Deren Bewertung erfolgt größtenteils im nächsten Abschnitt, nachdem einige wichtige Evaluationskriterien eingeführt wurden (eine sehr ausführliche Diskussion der meisten hier vorgestellten Beurteilungsverfah-

Beurteilungsverfahren sind Meßinstrumente, die eines oder mehrere Leistungskriterien abbilden.

[3] Summarische Beurteilungen sind in der Praxis selten; bei der Analytik wird oft übertrieben, so etwa, wenn 20 oder mehr Merkmale eingeschätzt werden sollen.

ren findet sich bei Bernardin & Beatty, 1984; knapper bei Schuler, 1991; zu technischen Fragen der Skalierung allgemein vgl. Borg & Staufenbiel, 1993).

Neben formalen Skalierungsverfahren können auch freie Eindrucksschilderungen, verschiedene Eignungsdiagnostika und Arbeitsanalyseverfahren der Leistungsbeurteilung dienen.

Außerdem beschränkt sich die Darstellung und nachfolgende Diskussion auf *formale* und *originär* der Leistungsbeurteilung zuzurechnende Verfahren. Formlose, *freie Eindrucksschilderungen* dürften allerdings in kleineren Unternehmen und für außertariflich bezahlte Mitarbeiter (i.d.R. höhere Führungskräfte) nach wie vor die dominierende Beurteilungsform darstellen. Ein aus Sicht der meisten Beurteiler schätzenswerter Vorzug freier Eindrucksschilderungen ist der fehlende Zwang, sich vorgegebenen Schemata anzupassen; die damit verbundene Flexibilität macht sie zur geeigneten Grundlage für Feedback und Verhaltenssteuerung. Vergleichende Beurteilungen, die psychometrischen Mindestansprüchen genügen sollen, erfordern dagegen ein gewisses Maß an Standardisierung. *Arbeitsproben*, *Assessment Center* oder die hierzulande noch wenig beachteten *Kenntnistests* können als Methoden v.a. der Potentialbeurteilung dienen, sind aber prinzipiell eher der Eignungsdiagnostik zuzurechnen und werden dementsprechend an anderer Stelle erörtert (vgl. Kapitel 6). Ihr Hauptproblem im Kontext der Leistungsbeurteilung besteht darin, daß damit eher *maximale* als *typische* Leistung erfaßt wird. Gelegentlich wird auch die *Methode der Kritischen Ereignisse* (Critical Incident Technique [CIT], Flanagan, 1954) als eigenständiges Beurteilungsverfahren dargestellt, die inzwischen aber vorwiegend als Instrument der Anforderungsanalyse eingesetzt wird (vgl. Kapitel 3). Als solches spielt sie eine wichtige Rolle für die Konstruktion einiger Beurteilungsskalen.

5.1 Einstufungsverfahren

Der weitaus größte Teil der in der Praxis gebräuchlichen Skalenformate ist der Gruppe der Einstufungsverfahren zuzurechnen. Ihr Prinzip besteht darin, zunächst die *Ausprägung von Merkmalen*, die auf allen Ebenen der Beurteilung (siehe Abschnitt 2.1) angesiedelt sein können, auf *mehrstufigen Skalen* einschätzen zu lassen. Der Vergleich zwischen Personen findet erst in einem zweiten Schritt statt. Für die Abstände zwischen den Skalenpunkten wird meist – oft ungeprüft und daher mit unbekannter Berechtigung – ein metrisches Skalenniveau unterstellt, so daß der größte Teil statistischer Prozeduren angewendet werden kann. Die Zahl der *Merkmalsdimensionen* – nicht der Items! – sollte das in Abschnitt 2.2 angedeutete Maß nicht überschreiten; bei der Anzahl der *Skalenstufen* liegt ein psychometrisches Optimum etwa im Bereich von 5 bis 7, maximal 9 Punkten. Die verbale Verankerung zumindest der Endpunkte der Skala ist der Angleichung von Urteilsmaßstäben dienlich und bei den elaborierteren Verfahren ohnehin Teil der Konstruktionsvorschrift.

Bei Einstufungsverfahren werden Personen zunächst unabhängig voneinander auf metrischen Skalen eingeschätzt und erst anschließend verglichen. Dieser Verfahrenstyp ist in der Praxis am stärksten verbreitet.

Graphische Einstufungsskala

Im einfachsten Fall, bei der sog. *Graphischen Einstufungsskala*, findet die Umsetzung als relevant erachteter Merkmale in Skalen ohne ein formales Skalierungsverfahren statt (siehe Abbildung 4 zu Beispielen für dieses und andere Einstufungsverfahren). Die Angemessenheit der im Beispiel erfolgten Skalenverankerung läßt sich im Vorfeld nicht und post hoc nur schwer überprüfen. Dies erfordert einen wesentlich komplizierteren Konstruktionsvorgang, wie er für die folgenden Verfahren typisch ist.

Verhaltensverankerte Einstufungsskala

Verhaltensverankerte Einstufungsskala (Behaviorally Anchored Rating Scales [BARS], Smith & Kendall, 1963) werden nach einem komplexen mehrstufigen, inzwischen mehrfach revidierten Schema konstruiert, dessen Grundprinzip der aus der Einstellungsmessung bekannten Thurstone-Skalierung (genauer: Thurstones „Methode gleich erscheinender Intervalle“) entlehnt ist. Wir wollen den Konstruktionsprozeß etwas ausführlicher schildern. Zunächst definiert eine Gruppe von Beurteilern relevante Leistungsdimensionen und konkretisiert diese in Umschreibungen für jeweils gute, mittlere und schwache Leistung, die später als Skalenverankerung dienen (linke, umrandete Statements des Beispiels in Abbildung 4). Eine zweite Beurteilerstichprobe formuliert dann konkrete Verhaltensbeispiele für jede Dimension

(nach Art der CIT; dieser Schritt wird bei einer revidierten Konstruktionsvorschrift vorgezogen), die anschließend von einer dritten unabhängigen Gruppe in Dimensionen „rückübersetzt“ werden. Verhaltensbeispiele, die diesen Prozeß überstanden haben (Kriterium ist eine hohe Übereinstimmung bei der Zuordnung), werden von einer weiteren Stichprobe den einzelnen Skalenstufen zugeordnet. Übrig bleiben Verhaltensbeispiele, die sich, bei möglichst geringer Streuung der Einschätzungen, ungleichmäßig über die gesamte Skalenlänge verteilen (rechte Seite des Bsp. in Abbildung 4). In früheren Versionen waren diese Beispiele so formuliert, daß der Beurteiler seine Einschätzung des vom Beurteilten zukünftig zu erwartenden Verhaltens abzugeben hatte (daher der ursprüngliche Name Verhaltenserwartungsskala); inzwischen soll dagegen beobachtetes Verhalten registriert werden. Ziel der BARS-Methode ist eine verhaltensnahe Beurteilung und die Hinlenkung der Beurteiler auf ein gemeinsames, arbeitsanalytisch fundiertes Beobachtungsschema. Ursprünglich wurde auch den Letztbeurteilern dazu einiger Aufwand abverlangt. Die BARS ist wohl das am besten erforschte Skalenformat.

Verhaltensbeobachtungsskala

Nicht ganz so komplex, aber gleichfalls CIT-basiert und teststatistisch abgesichert ist die Konstruktion von *Verhaltensbeobachtungsskalen* (Behavior Observation Scales [BOS], Latham & Wexley, 1977). Das Ergebnis entspricht formal einer großen Anzahl von Likert-Items (siehe Abbildung 4; vollständig ergeben sich etwa 50 Items), bei denen Häufigkeitseinschätzun-

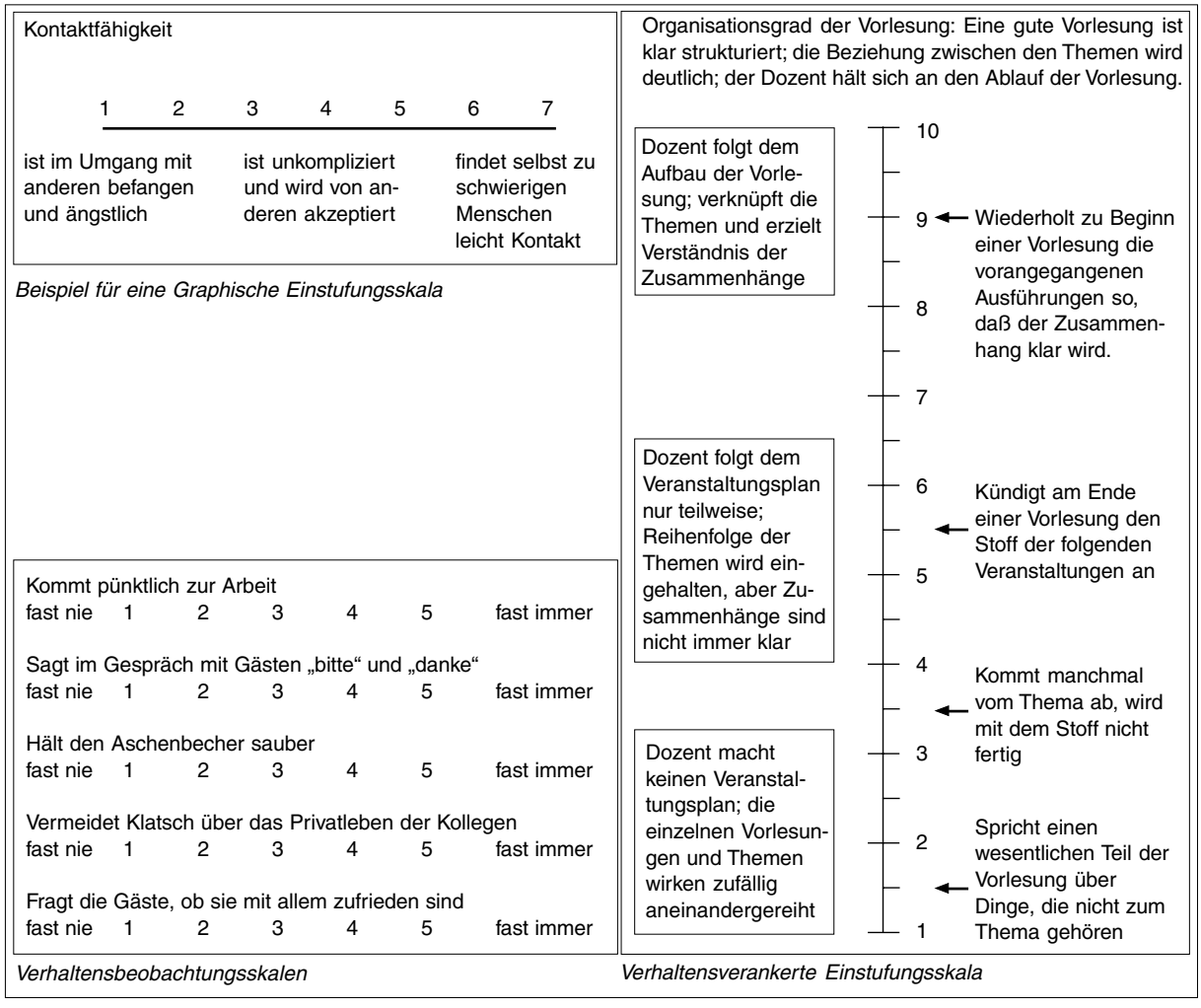


Abbildung 4:
Beispiele für Einstufungsverfahren (aus Schuler, 1991)

gen ausschließlich *beobachtbarer Verhaltensweisen* abgegeben werden sollen, die anschließend zu mehreren Dimensionen zusammengefaßt werden können. Man erhofft sich von diesen Häufigkeitsschätzungen verhaltensnähere, weniger subjektiv-bewertende Urteile und damit auch eine geeignete Grundlage gezielter Verhaltenssteuerung.

Mixed Standard Scale

Die auf den Finnen Blanz zurückgehende *„Mixed Standard Scale"* (MSS, Blanz & Ghiselli, 1972) durchläuft zur Generierung und Verankerung der Dimensionen den gleichen iterativen Prozeß wie die BARS, endet aber bei einem gänzlich anderen Format, das dem Prinzip der Guttman-Skala entspricht; d.h. die Verhaltensaussagen sind *innerhalb einer Dimension* hierarchisch („kumulativ") geordnet, was auf dem Formular durch eine gemischte (daher „mixed") Gruppierung verschleiert wird. Aufgabe des Beurteilers ist es lediglich, anzugeben, ob der Beurteilte typischerweise besser, schlechter oder genauso gut agiert, wie in dem betreffenden Item formuliert. Die Guttman-Skalierung erlaubt eine Überprüfung des Urteils auf logische Inkonsistenzen, wenn bspw. ein Beurteilter als „genauso gut" wie in einer als „schwach" skalierten Aussage, aber „besser" als in einem „mittleren" Statement eingeschätzt wird. Die Generierung reliabler Guttman-Skalen ist allerdings äußerst schwierig.

Modell der Verteilungsmessung

Unter den neueren und noch wenig erforschten und eingesetzten Entwicklungen auf dem Gebiet der Einstufungsverfahren (siehe Borman, 1991, für eine Übersicht) beruhen die vielleicht interessantesten auf dem *Modell der Verteilungsmessung* (Kane, 1986), das neben einem Modalwert der Verteilung auch deren Streuung berücksichtigt und somit eine Aussage über die *Konsistenz* des erfaßten Verhaltens erlaubt. Dahinter steht die Überlegung, daß interindividuelle Leistungsunterschiede sich nicht nur in der Höhe eines typischen Leistungs*niveaus* ausdrücken, sondern auch in unterschiedlich starken *Schwankungen* der Leistung über die Zeit. Zur Itemgenerierung dienen hier jeweils wiederum nach Art der CIT gewonnene Verhaltensbeispiele. Bei *Verhaltensunterscheidungsskalen* (Behavior Discrimination Scales [BDS]) werden diese Items dann in einem relativ komplexen Prozeß daraufhin untersucht, wie *oft* ein Stelleninhaber *Gelegenheit* hat, das beschriebene Verhalten zu zeigen, wie gut dieses Verhalten zwischen befriedigender und unbefriedigender Leistung *diskriminiert* und welchem *Effektivitätsgrad* es entspricht. Der Beurteiler muß letztlich einschätzen, wie oft der Beurteilte Gelegenheit zu dem jeweiligen Verhalten hatte und wie oft er es tatsächlich zeigte. Aus den so errechneten Prozentsätzen lassen sich verschiedene Kennwerte für die Effektivität einzelner Verhaltensweisen und deren Häufigkeit bilden. Die einzelnen Werte werden in nach ihrer Bedeutung gewichteten Verhaltensclustern zusammengefaßt, die weiter zu Dimensionen aufsummiert werden können. *Leistungsverteilungsbeurteilungen* (Performance Distribution Assessment [PDA]) sind eine Variante der BDS, die für Fälle entwickelt wurde, bei denen nicht genügend Positionsinhaber für den komplexen Konstruktionsprozeß zur Verfügung stehen. Die Bestimmung und Gewichtung der Arbeitselemente erfolgt hier in einer traditionellen Arbeitsanalyse; die Konstruktion schließt die Ermittlung *möglicher* Leistungsbereiche unter wechselnden Gegebenheiten ein und erlaubt so eine quantitative Eliminierung situativer Einflüsse. Die beträchtliche Komplexität beider Verfahrensvarianten dürfte allerdings ihrem praktischen Einsatz nicht eben förderlich sein.

5.2 Rangordnungsverfahren

Bei Rangordnungsverfahren werden Mitglieder einer Gruppe direkt miteinander verglichen. Das Ausmaß der Leistungsunterschiede wird nicht erfaßt.

Prinzip der Rangordungsverfahren ist i.a. der *direkte Vergleich* zwischen den Mitgliedern einer Beurteiltengruppe durch einen oder mehrere Beurteiler. Die Beurteilten können dabei auch anhand mehrerer Merkmale („analytisch") verglichen werden. Da die Daten lediglich auf Ordinalskalenniveau erhoben werden, treffen Rangordnungsverfahren keine Aussage über das *Ausmaß* der Leistung beim einzelnen Beurteilten und erlauben somit keine Vergleiche über die jeweils einbezogene Gruppe hinaus. Innerhalb dieser

Gruppe werden die Beurteiler zu Differenzierungen *gezwungen*, unabhängig davon, ob diese gerechtfertigt oder erwünscht sind.

Direkte Rangreihe

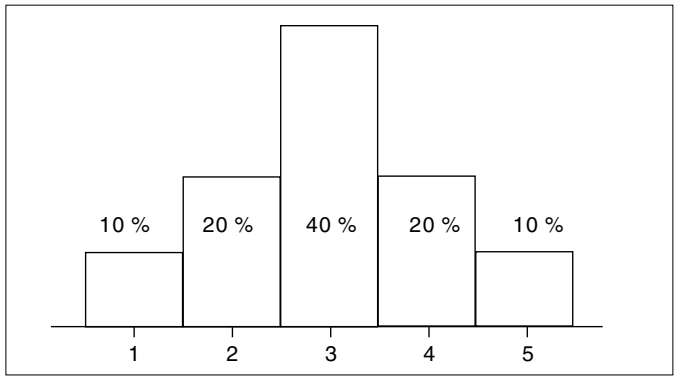
Im methodisch einfachsten Fall werden die Beurteiler aufgefordert, eine *direkte Rangreihe* von Personen (allgemein: *Objekten*) anhand globaler oder differenzierter Kriterien aufzustellen. Mit zunehmender Zahl von Beurteilungsobjekten stößt diese Vorgehensweise schnell (spätestens bei ca. 10 bis 11 Beurteilten) an kognitive Grenzen seitens der Beurteiler, was sich v.a. in Reliabilitätsmängeln niederschlägt.

Quotenvorgabe

Eine Variante der einfachen Rangreihenbildung ist die sog. *Quotenvorgabe*, bei der Rangplätze – in unterschiedlicher Häufigkeit – mehrfach zu vergeben sind. Am verbreitetsten ist dabei die Orientierung an einer Normalverteilung, so daß die Beurteiler aufgefordert sind, extreme Rangplätze zahlenmäßig schwächer zu besetzen als den mittleren Bereich. Bei der Quotenvorgabe tritt der Nullsummencharakter der Rangordnungsverfahren (der Aufstieg eines Beurteilten in der Rangreihe führt automatisch zum Abstieg eines anderen) besonders deutlich zu Tage.

Direkte Rangordung, Gesamtbewertung	Direkte Rangordnung wissenschaftl. Ruf	Paarvergleich Gesamtbewertung
1. Universität A	1. Universität B	A > B
2. Universität B	2. Universität A	I > K
.	.	
.	.	Paarvergleich,
.	.	wiss. Ruf
9. Universität I	9. Universität K	B > A
10. Universität K	10. Universität I	K > I

Direkte Rangreihenbildung und Paarvergleich



Quotenvorgabe

Rangreihe Person	Rangreihe Arbeitsplatz
Fachkenntnisse	Führung
Organisation	Organisation
Ausdauer	Innovation
.	.
.	.
.	.
Innovation	Ausdauer
Führung	Fachkenntnisse

Verhaltensrangprofil

Über gute Fachkenntnisse verfügt:
Wer umfangreiches Wissen auf seinem Gebiet besitzt; praktische Erfahrungen hat; seine Kenntnisse erfolgreich anwenden kann.

- Erfüllt, was seine Fachkenntnisse betrifft, nicht die Anforderungen, die an ihn gerichtet sind
- Seine Fachkenntnisse liegen über den Mindestanforderungen (trifft auf etwa 90 % der Mitarbeiter an vergleichbaren Arbeitsplätzen zu)
- Hat mindestens ebenso gute Fachkenntnisse wie 50 % der anderen Mitarbeiter an vergleichbaren Arbeitsplätzen
- Hat mindestens ebenso gute Fachkenntnisse wie 25 % der anderen Mitarbeiter an vergleichbaren Arbeitsplätzen
- Hat mindestens ebenso gute Fachkenntnisse wie nur 10 % der anderen Mitarbeiter an vergleichbaren Arbeitsplätzen
- Hat mindestens ebenso gute Fachkenntnisse wie nur 5 % der anderen Mitarbeiter an vergleichbaren Arbeitsplätzen

Sequentielle Prozentrangskala

Abbildung 5:
Beispiele für Rangordnungsverfahren (aus Schuler, 1991)

Paarvergleiche

Reliabler als die einfache Rangreihenbildung, weil einfacher zu bewältigen, sind *Paarvergleiche* zwischen jeweils zwei Beurteilten, bei denen die Rangreihe erst im nachhinein rechnerisch ermittelt wird und die zusätzlich – über die Ermittlung logischer Inkonsistenzen (z. B. Verletzung der Transitivität: A>B, B>C, aber C>A) – eine Aussage über die Urteilsqualität erlauben. Zudem ist bei Paarvergleichen eine metrische Skalierung möglich. Allerdings steigt der *Aufwand* mit zunehmender Zahl der Beurteilten steil an[4], so daß hier zwar kaum mit kognitiven, dafür aber mit motivationalen Obergrenzen gerechnet werden muß.

Verhaltensrangprofil

Eine gänzlich andere Intention wird mit dem *Verhaltensrangprofil* (Brandstätter & Schuler, 1974) verfolgt. Hier werden nicht Personen, sondern *Merkmale einer einzelnen Person* in eine Rangreihe gebracht und auf Übereinstimmung mit der – arbeitsanalytisch ermittelten – Rangfolge von Arbeitsplatzanforderungen verglichen. Ziel ist dementsprechend nicht der interpersonale Vergleich, sondern die Überprüfung, inwieweit Position und (aktueller oder potentieller) Inhaber zueinander passen. Dies dient als Grundlage von Personalentwicklung und Arbeitsplatzgestaltung. Über eine Kontrollreihe ist es dabei möglich, einen Qualitätsindikator für jede einzelne Beurteilung zu ermitteln. Der Vergleich zwischen *Beurteilten* erfordert zusätzlich die Anwendung eines Einstufungsverfahrens, wofür von Brandstätter und Schuler (1974) eine *Sequentielle Prozentrangskala* vorgeschlagen wurde, die, zur Vermeidung von Mildetendenzen, im oberen Leistungsbereich stärker differenziert. Abbildung 5 zeigt Beispiele für Rangordnungsverfahren.

5.3 Auswahl- und Kennzeichnungsverfahren

Auswahl- und Kennzeichnungsverfahren verlangen lediglich die Zustimmung oder Ablehnung bestimmter Aussagen, deren Wertigkeit zuvor bestimmt wurde und den Beurteilern selbst unbekannt ist.

Die Verfahren dieser Kategorie erfordern seitens der Beurteiler keine relativen Einstufungen oder direkten Vergleiche, sondern lediglich die *Zustimmung* bzw. *Ablehnung* vorgegebener Aussagen oder deren *Auswahl* aus einer Liste nach dem Multiple-Choice-Prinzip. Die Günstigkeit der damit abgegebenen Urteile wird im Rahmen der Konstruktion bestimmt und ist den Beurteilern selbst i.d.R. unbekannt, womit verschiedenen Urteilstendenzen (siehe Abschnitt 6) entgegengewirkt werden soll. Kennzeichnungs- und Auswahlverfahren zählen zu den ältesten formalen Beurteilungsskalen.

Gemischte Aussagenliste mit freier Wahl

Bei der *Gemischten Aussagenliste mit freier Wahl* (Checklist, Knauft, 1948) wird zunächst eine große Zahl arbeitsrelevanter Aussagen generiert, die anschließend, wiederum nach dem Prinzip der Thurstone-Skalierung, von einer Expertengruppe daraufhin eingeschätzt werden, welchen Grad an Effektivität der Arbeitsleistung sie jeweils repräsentieren. Am Ende verbleibt eine Aussagenliste, die nach Möglichkeit den gesamten Leistungsbereich gleichmäßig abdecken sollte. Die *Streuung* der Expertenschätzungen über die Effektivität dient als Ausscheidungskriterium für die Items (nur gering streuende Items verbleiben in der Liste); ihr jeweiliger *Mittelwert* oder Median dient zur Gewichtung. Die Beurteilung erfolgt schließlich durch einfache Zustimmung oder Ablehnung; die Summe oder der Median der Effektivitätsgewichte für die Aussagen, denen zugestimmt wurde, bildet den Leistungswert des Beurteilten.

Wahlzwangverfahren

Anders als einfache Checklisten fordert das – ursprünglich von dem Pionier der kognitivistischen Leistungsbeurteilungsforschung Robert Wherry angeregte – *Wahlzwangverfahren* (Forced Choice, zuerst beschrieben bei Sisson, 1948) vom Beurteiler, aus einer Liste von jeweils zwei oder mehreren Aussagen diejenige auszuwählen, die ihm am treffendsten erscheint. Die Aussagen werden dabei derart gruppiert, daß jeweils zusammengehörige Statements *gleich günstig* (sozial erwünscht) erscheinen, aber möglichst *unterschiedlich effektives* Verhalten beschreiben. Die Überprüfung dieser beiden Voraussetzungen erfordert in einem aufwendigen Konstruktionsprozeß

[4] Die Rangreihenbildung von n Urteilsobjekten erfordert bei vollständigem Vergleich $n(n-1)/2$ Paarvergleiche; also z. B. 10 Vergleiche bei 5 Personen, aber bereits 105 bei 15 Beurteilten; bei mehreren Merkmalen tritt ein multiplikativer Faktor hinzu.

(näheres bei Bernardin & Beatty, 1984) die Bestimmung eines *Bevorzugungsindex* (gleiche Erwünschtheit) und eines *Diskriminationsindex* (unterschiedliche Effektivität) für jede einzelne Verhaltensaussage. Mit diesem Skalenformat sollen die Manipulationsmöglichkeiten der Beurteiler minimiert werden (siehe Abbildung 6 für Beispiele der beiden beschriebenen Skalen).

	trifft zu	trifft nicht zu	
Die Schaufenster sind attraktiv gestaltet	☐	☐	(8,5)
Manchmal werden Brötchen verkauft, die auf den Boden gefallen sind	☐	☐	(1,4)
Läßt sich die Arbeit von seinen Bäckern rückdelegieren	☐	☐	(3,1)
Ermutigt seine Mitarbeiter zur Initiative	☐	☐	(8,1)
Hat brauchbare Formulare entwickelt	☐	☐	(6,4)

Gemischte Aussagenliste mit freier Wahl (in Klammern stehen die dem Beurteiler unbekannten Aussagengewichte)

Aussage	Diskriminations-index	Bevorzugungs-index
1. Hat Geduld mit langsam Lernenden	1,15	2,82
2. Zeigt beim Unterrichten Selbstvertrauen	0,54	2,75
3. Weckt Interesse und Aufmerksamkeit bei den Schülern	1,39	2,89
4. Teilt die Lernziele der nächsten Stunde im Voraus mit	0,79	2,85

Gruppierte Aussagenliste mit Wahlzwang (die Werte sind den Beurteilern nicht bekannt)

Abbildung 6:
Beispiele für Kennzeichnungs- und Auswahlverfahren (aus Schuler, 1991)

5.4 An Zielerreichungsgraden orientierte Verfahren

Die Intention von an Zielerreichungsgraden orientierten Verfahren geht weit über die Leistungsbeurteilung im engeren Sinne hinaus.

Wenn in der Einleitung zu diesem Abschnitt einige im Zusammenhang mit Leistungsbeurteilung diskutierte Verfahren wegen ihrer Randständigkeit sehr knapp dargestellt wurden, so hätten wir dieser Aufzählung ebenso gut die zielorientierten Verfahren anfügen können. Die Intention dieser Verfahren, soweit sie sich überhaupt als solche kennzeichnen lassen, geht i.d.R. weit über Leistungs*beurteilung* im engeren Sinne hinaus; sie sind entweder, wie das Konzept der *Zielsetzung* (Locke & Latham, 1990), eher eine Motivationstheorie denn ein Beurteilungsverfahren und stehen damit direkten Interventionen näher; oder sie stellen, wie *Management by Objectives* (MbO, Drucker, 1954) und dessen Derivate, einen umfassenden Managementansatz dar, der Leistungsbeurteilung nur als eines seiner Elemente einschließt. Der Grund, warum hier zumindest letzterer Ansatz gesondert vorgestellt wird,

ist, daß dieser gelegentlich als das Nonplusultra der Leistungsbeurteilung dargestellt wird (z. B. Crisand & Stephan, 1994) und wir es als angezeigt erachten, diese Einschätzung etwas zu relativieren.

MbO ist der Stammvater und noch immer bei weitem bedeutendste Vertreter einer ganzen Reihe von „Management by...“-Ansätzen, deren Spektrum von seriösen oder zumindest diskussionsfähigen Führungskonzepten bis zu Modeerscheinungen mit eher humoristischem Charakter reicht (vgl. z. B. Macharzina, 1993). Grundgedanke des MbO ist die Steuerung der Mitarbeiter, v.a. mittlerer Hierarchieebenen, durch *Ziele*, die, ausgehend von globalen Unternehmenszielen, in einem mehrstufigen Prozeß bis auf die Ebene des individuellen Mitarbeiters heruntergebrochen werden. Die Ziele können *direktiv* verordnet oder *partizipativ* vereinbart werden; bei der Umsetzung soll den Mitarbeitern größtmöglicher Freiraum gelassen werden. Daß die Leistungsbeurteilung schließlich nach dem Grad der Zielerreichung erfolgt (durch Soll-Ist-Vergleich, wobei der Sollwert während der Periode wechselnden Rahmenbedingungen angepaßt werden kann), ist eine logische *Folge* dieses Vorgehens. Umgekehrt *erfordert* eine zielorientierte Beurteilung zunächst eine Ziel*setzung*. Von deren Qualität (Sind die Ziele weder zu hoch noch zu niedrig; sind sie vergleichbar, operationalisiert etc.?) hängt die Qualität der Beurteilung ganz entscheidend ab. Da in der Praxis *Ergebnisziele* dominieren, wirkt sich deren spezielle Problematik auch im Rahmen des MbO voll aus. Zusätzlich zu den in Abschnitt 2.1 angesprochenen Schwierigkeiten ist darauf hinzuweisen, daß die Erkenntnisse der Zielsetzungsforschung (vgl. Kapitel 13) für die Erfüllung der Motivations- und Verhaltenssteuerungsfunktion *verhaltensorientierte* Ziele nahelegen.

Bei zielorientierten Verfahren wird der Erreichungsgrad zuvor vereinbarter Ziele beurteilt.

6 Urteilsqualität

Leistungen werden, wie wir gesehen haben, mit unterschiedlichen Kriterien, Quellen und Verfahren beurteilt. Die Evaluation dieser Vorgehensweisen erfordert ihrerseits wieder eine Beurteilung anhand von „Kriterien für Kriterien“, die sich letztlich, wie noch zu zeigen sein wird, noch einmal aneinander messen lassen. Eine Vielzahl solcher Maßstäbe der Urteilsqualität wird in der Literatur diskutiert. Wir werden uns zunächst dem Bereich der „harten“, eher technischen Kriterien zuwenden, die in wissenschaftlichen Publikationen traditionell den weitaus breitesten Raum einnehmen, um anschließend noch auf einige eher „weiche“, für die Praxis aber nichtsdestoweniger bedeutsame Urteilsmaßstäbe einzugehen. Anhand dieser Kriterien erfolgt dann eine knappe Evaluation der in Abschnitt 5 dargestellten Verfahren. Abschließend werden einige Ansatzpunkte dargestellt, die Ursachen mangelnder Urteilsqualität zu erforschen und nach Möglichkeit zu beheben.

6.1 „Technische“ Gütekriterien

6.1.1 Reliabilität und Validität

Ein naheliegendes Vorgehen zur Evaluation von Leistungsbeurteilungen ist der Rückgriff auf die Gütekriterien der Klassischen Testtheorie, deren Grundzüge hier als bekannt vorausgesetzt werden (vgl. auch Kapitel 22). Die Übertragung erfordert aber schon für den einfacheren Bereich der *Reliabilität* einige inhaltliche Anpassungen. Maße der *internen Konsistenz* (Homogenität) sind bei einem heterogenen Konstrukt wie globaler Berufsleistung nicht sinnvoll anwendbar, und ihre Messung bezüglich einzelner homogener Dimensionen scheitert häufig daran, daß diese in der Praxis meist nur mit wenigen Items, oft sogar nur einem einzigen erfaßt werden. Noch seltener liegen Parallelformen von Beurteilungsverfahren zur Schätzung der *Paral-*

Testtheoretische Gütekriterien lassen sich nur unter einigen Abwandlungen auf die LB übertragen.

leltestreliabilität (Äquivalenz) vor. Als Surrogat für die *Retest-Reliabilität* (Stabilität) wird dagegen vglw. häufig die Übereinstimmung der Einschätzungen der gleichen Beurteilten durch denselben Beurteiler zu verschiedenen Zeitpunkten (*intraindividuelle Urteilskonkordanz*) ermittelt, wobei sich aber zufällige Änderungen (Fehlervarianz) kaum von echten Leistungsschwankungen (Merkmalsvarianz) trennen lassen[5]. Verbreitet sind auch Erhebungen der Übereinstimmung verschiedener Beurteiler (Objektivität oder *interindividuelle Urteilskonkordanz*), was allerdings eine Mehrzahl von Beurteilern desselben Gegenstands erfordert. Viswesvaran, Ones & Schmidt (1996) schätzen die Stabilität von Gesamtbeurteilungen von Vorgesetzten metaanalytisch im Mittel auf r_{tt} = .81, ihre interne Konsistenz (a) auf .85. Wesentlich niedriger fallen die Schätzungen für die Objektivität aus, nämlich .52 für Vorgesetzten- und sogar .42 für Kollegenbeurteilungen. Diese interindividuelle Urteilskonkordanz wird jedoch manchmal bereits als Indikator der Validität interpretiert.

Besonders schwierig ist die Übertragung des Validitätskonzepts auf die Leistungsbeurteilung.

Noch schwieriger als bei der Reliabilität gestaltet sich die Übertragung von Aspekten der *Validität* auf die Leistungsbeurteilung. Auf die Problematik der Bestimmung von *Kriteriumsvaliditäten* durch Validierung eines Leistungskriteriums an einem anderen (z. B. subjektive vs. objektive Kriterien, Fremd- vs. Selbstbeurteilung etc.) wurde bereits hingewiesen; sie führt letztlich in einen unendlichen Regreß. Nicht weniger problematisch ist der Umkehrschluß aus dem Befund, daß Prädiktoren – gemessen an Leistungsbeurteilungen – Validität besitzen, darauf, daß die Beurteilungen ihrerseits valide seien. Eine Vielzahl unterschiedlichster Befunde wurde im Zusammenhang mit dem vielschichtigen Konzept der *Konstruktvalidität* interpretiert. Ohne auf die Schwierigkeiten der daraus gezogenen Schlüsse im einzelnen eingehen zu können (vgl. Murphy & Cleveland, 1995), läßt sich dagegen einwenden, daß solche Interpretationen oft einem Stochern im Nebel gleichen, solange das Konstrukt beruflicher Leistung nicht hinreichend definiert ist (auf diesbezügliche Fortschritte wurde oben bereits hingewiesen).

6.1.2 Genauigkeit (accuracy)

Akkuratheit ist ein Konzept, das – im Gegensatz zu Reliabilität und Validität – auf Leistungsbeurteilungen zugeschnitten ist. Definiert wird die Akkuratheit eines Urteils im Grunde von ihrem negativen Pol (Ungenauigkeit), nämlich als *Abweichung von einem wahren Leistungswert*. Zur Ermittlung dieser wahren Leistung wurden Techniken entwickelt, die unter Laborbedingungen sehr exakte Schätzungen erlauben (siehe z. B. Borman, 1991). Hierin besteht jedoch auch der Haupteinwand gegen Akkuratheitsmaße als Kriterium: Ihre Anwendung beschränkt sich bislang weitgehend auf Laborexperimente. Weitere Kritik richtet sich auf die unklare Operationalisierung von Genauigkeit (Balzer & Sulsky, 1990).

Akkuratheit bezeichnet die Abweichung eines Urteils von einem wahren Leistungswert, wobei vier einzelne Komponenten unterschieden werden müssen.

Cronbach (1955) konnte in einem klassischen Aufsatz zeigen, daß die Operationalisierung von Akkuratheit über globale Abweichungsmaße (Differenzsumme oder euklidische Distanz) zur Konfundierung mehrerer, unterschiedlich zu interpretierender Komponenten führt, wenn mehr als ein Merkmal beurteilt wird. Bei mehreren Beurteilten und mehreren Merkmalen ergeben sich vier Elemente der Akkuratheit, die untereinander nur schwach korrelieren (Balzer & Sulsky, 1990). Alle vier Maße kennzeichnen aber jeweils die Urteilsqualität *eines* Beurteilers (siehe Cronbach, 1955, oder Murphy & Balzer, 1989, für eine mathematische Definition der Komponenten):

---

[5] Wegen dieser Problematik wurde die Leistungsbeurteilung als Anwendungsfeld für das umfassendere Reliabilitätskonzept der *Generalisierbarkeitstheorie* (für eine knappe Darstellung vgl. Nußbaum, 1987) vorgeschlagen, die eine Zerlegung der „Fehlervarianz" in einzelne Komponenten erlaubt (z. B. Borman, 1991). Eine praktische Umsetzung dieses Vorschlags steht u.W. allerdings noch aus.

1. *Elevation* (Abhebung) ist ein *globales* Maß für den durchschnittlichen Abstand zwischen den Urteilen eines Beurteilers und den wahren Werten, über alle Dimensionen und Beurteilten gemittelt.
2. *Differential Elevation* beschreibt die Genauigkeit des Urteils je Beurteiltem, global über alle Dimensionen gemittelt, und trifft somit eine Aussage über *beurteiltenspezifische* Ungenauigkeiten.
3. *Stereotype Accuracy* bezieht sich auf die Genauigkeit, mit der die wahren Mittelwerte je Dimension getroffen werden, gemittelt über die Beurteilten, und mißt also *merkmalsspezifische* Ungenauigkeiten.
4. *Differential Accuracy* schließlich mißt die Korrektheit, mit der individuelle Leistungsunterschiede je Dimension (Stärken und Schwächen) erkannt werden, wobei globale Person- und Merkmalseffekte eliminiert werden.

Welcher der vier Komponenten insgesamt die größte Bedeutung zukommt, ist umstritten. Wichtig ist auf jeden Fall, sie zur Vorbeugung von Fehlinterpretationen auseinanderzuhalten.

6.1.3 Urteilstendenzen

Mit Urteilstendenzen sind Abweichungen der Leistungsbeurteilung von einer theoretisch erwarteten Verteilung bezüglich Mittelwert, Streuung und Interkorrelation der Urteilsdimensionen gemeint.

Die Forschung zur Leistungsbeurteilung beschäftigt sich seit mehr als 50 Jahren intensiv mit gewissen Verteilungsanomalien oder Abweichungen der tatsächlichen Urteilswerte von aufgrund theoretischer Überlegungen erwarteten Verteilungen bzw. Kovarianzen. Solche sog. Urteilstendenzen wurden in der Vergangenheit als Indikator für mangelnde Urteilsqualität, v.a. als Surrogat für Akkuratheit, interpretiert. Neuere Forschung legt, wie noch zu zeigen sein wird, eine erhebliche Relativierung dieser Einschätzung nahe. Brandstätter (1970) faßt die wichtigsten Urteilstendenzen unter den an basalen statistischen Kennwerten orientierten Kategorien der *Mittelwerts-*, *Streuungs-* und *Korrelationstendenzen* zusammen. Zur Illustration soll der jeweils bedeutendste Vertreter dieser Bereiche kurz erläutert werden.

Mittelwertstendenzen

Mittelwertstendenzen beziehen sich auf Abweichungen der *Lage der zentralen Tendenz* von einem angestrebten Wert, der i.d.R. in der Skalenmitte zu suchen ist. Obwohl bei einzelnen Beurteilern oder Beurteiler/Beurteilten-Dyaden auch Abweichungen nach unten (Strengetendenz oder severity) auftreten, besteht die *generelle* Tendenz bei Leistungsbeurteilungen in der Praxis eindeutig in einer Verlagerung zum oberen Skalenende (*Mildetendenz* oder leniency), i.d.R. interpretiert als eine Beschönigung der Urteile. Spätestens seit Einführung der Offenlegungspflicht durch das Betriebsverfassungsgesetz von 1972 scheint es nicht ungewöhnlich zu sein, daß ein Großteil der Beurteilungen im oberen Drittel der Skala liegt, eine deutliche Mehrheit also als „überdurchschnittlich“ beurteilt wird. Derartiges Beurteilerverhalten ist nicht unbedingt als Irrtum oder Fehler zu verstehen. Murphy und Cleveland (1995, Kap. 9) diskutieren ausführlich Gründe, warum sich Beurteiler durchaus rational – im Sinne der Verfolgung eigener Ziele – verhalten, wenn sie ihre Urteile beschönigen (siehe auch die Ausführungen zu mikropolitischem Verhalten oben).

Streuungstendenzen

Streuungstendenzen beziehen sich auf eine erwartungskonträre Ausnutzung der Skalenlänge, in aller Regel in Form einer *Streuungseinschränkung* (range restriction); die Werte klumpen an bestimmten Skalenpunkten. Die in diesem Zusammenhang häufig erwähnte *Tendenz zur Mitte* (der Skala) oder Zentraltendenz[6] ist ein insofern ungeschickt gewähltes Beispiel, als Streuungseinschränkungen im Kontext der Leistungsbeurteilung oft mit Mildetendenzen einhergehen, die Häufung also am oberen Skalenende lokalisiert ist. Geringe Urteilsstreuung wird oft ungeprüft als Fehler interpretiert, kann aber auch tatsächliche Leistungshomogenität widerspiegeln, eine Leistungs-

[6] Dieser Begriff ist wegen der geringeren Verwechslungsgefahr mit der Mittelwertstendenz vorzuziehen.

verteilung also, die – ähnlich wie bei der Mildetendenz – von den Organisationen durchaus angestrebt wird. Dies würde allerdings der Forschung zur Variabilität individueller Leistungen widersprechen (siehe Abschnitt 6.2).

Korrelationstendenzen

Die meistdiskutierte – und -beklagte – Urteilstendenz bezieht sich auf die Korrelation als unabhängig erachteter Leistungsdimensionen: die *Überstrahlung* (Halo) dieser Dimensionen durch ein Globalurteil. Wie wir weiter oben gesehen haben, ist aber eine gewisse Interkorrelation der Urteilsaspekte durchaus realistisch (sog. „true halo"). Als Fehler interpretierbar ist also höchstens eine darüber hinausgehende Kovarianz („illusory halo"); diese kann im Einzelfall durchaus unterhalb des tatsächlichen Wertes liegen, also ein negatives Vorzeichen annehmen. Laborexperimente (vgl. Ilgen, Barnes-Farrell & McKellin, 1993, für eine Übersicht) zeigen einen der Regression zur Mitte vergleichbaren Effekt: „True halo" wirkt sich auf die Korrelation der Urteile aus, aber nur abgeschwächt; sehr niedrige tatsächliche Interkorrelationen werden eher überschätzt, sehr hohe Zusammenhänge unterschätzt.

Der Grund, warum wir hier glauben, auf nähere Ausführungen und die Darstellung weiterer Urteilstendenzen[7] verzichten zu können, liegt in der erst in jüngerer Zeit gewonnenen Erkenntnis, daß diese Tendenzen zwar häufig auftreten, ihre Einstufung als Maßstab der *Urteilsqualität* aber verfehlt wäre. Murphy und Balzer (1989) fanden in einer Metaanalyse aller Studien, in denen Urteilstendenzen und Maße der Akkuratheit erhoben wurden, daß die Beziehungen zwischen beiden Indikatorgruppen i.d.R. minimal sind (im Mittel bei r = -.05). Die einzigen substantiellen Beziehungen fanden sich zwischen Operationalisierungen von Halo und dem vierten Aspekt (differential accuracy) von Akkuratheit und waren *negativ*. Dies bedeutet, daß sich Halo *günstig* auf die korrekte Analyse der Stärken und Schwächen einzelner Mitarbeiter auswirkt. Die Befunde widersprechen diametral der konventionellen Lehrmeinung zu Urteilstendenzen als Indikator der Urteilsgenauigkeit.

Als Indikator für die Urteilsgenauigkeit sind Urteilstendenzen nicht geeignet.

6.2 „Praktische" Gütekriterien

Wissenschaftler, zumal wenn sie empirisch arbeiten, scheuen traditionell davor zurück, von ihnen entwickelte Instrumente an qualitativen, schwer in statistischen Kennzahlen faßbaren Kriterien zu messen. Gleichwohl können Beurteilungssysteme, die als nicht praktikabel oder akzeptabel empfunden werden, in der Praxis schlicht scheitern. Das gleiche gilt, wenn sich herausstellen sollte, daß der Schaden ihres Einsatzes den Nutzen überwiegt oder die angestrebten Ziele nicht adäquat erreicht werden können.

„Weiche" Qualitätsmaßstäbe sind für die Praxis der LB oft wichtiger als technische Gütekriterien.

Nutzenschätzungen für personalpsychologische Instrumente wurden vor fast fünfzig Jahren mit dem Ziel initiiert, den Wert dieser Verfahren in einem vergleichbaren und Betriebswirten (selbst)verständlichen Maßstab (Geld) ausdrücken zu können (vgl. Kapitel 22). Sie verbinden i.d.R. sozialwissenschaftliche Statistik mit betriebswirtschaftlicher Investitionsrechnung. Ein für die Leistungsbeurteilung relevanter Nebeneffekt dieser Forschung besteht darin, daß nunmehr verläßliche Schätzungen über die interindividuelle Variabilität der Produktivität von Berufstätigen vorliegen. Diese fallen i.d.R. beträchtlich aus (Standardabweichungen von DM 20.000.- pro Jahr bilden eine ungefähre Untergrenze; in höheren Positionen ist ein Mehrfaches dessen typisch; vgl. z. B. Schuler, Funke, Moser & Donat, 1995) und widerlegen die v.a. von Anhängern des *Total Quality Management* (TQM, Deming, 1986) gegen Leistungsbeurteilungen erhobene Grundsatzkritik, sie entbehrten schon wegen vernachlässigbarer individueller Leistungsunterschiede jeder Grundlage. Als mittleren *Nutzeneffekt* von Leistungsbeurteilungen fand die vergleichende Studie von Guzzo, Jette und Katzell (1985) eine Steige-

Nutzenschätzungen rechnen psychometrische Kennwerte in betriebswirtschaftliche Renditeerwartungen um. Gute LB-Systeme können einen erheblichen Nutzeneffekt haben.

[7] bspw. die Überbewertung erst kürzlich erbrachter Leistungen (recency-Effekt) und daraus resultierende Verhaltensänderungen der Mitarbeiter („Nikolauseffekt", eigentlich keine *Urteils*tendenz) oder *beurteiltenspezifische* Tendenzen („Klebereffekt", „Hierarchieeffekt", Stereotype, Sympathie etc.)

rung der individuellen Leistung um etwa eine halbe Standardabweichung, wobei aber die Wirkung „reiner" Beurteilung kaum von der des Feedback zu trennen ist. Murphy und Cleveland (1995, S. 301) demonstrieren allerdings an einem hypothetischen Beispiel, daß solche Nutzenschätzungen unter Einbeziehung und hoher Gewichtung schädlicher Konsequenzen (hier: empfundene Ungerechtigkeit) auch negativ ausfallen können. Empirische Anhaltspunkte dafür liegen u.W. zur Zeit nicht vor. Bei geeigneter Modellierung erlauben Nutzenschätzungen die vergleichende Beurteilung mehrerer Systeme.

Unverständliche oder schwer handhabbare LB-Systeme scheitern meist in der praktischen Anwendung.

Aspekte der *Praktikabilität* beziehen sich bspw. auf die Verständlichkeit und Einfachheit der Handhabung von Beurteilungssystemen oder auf den dazu erforderlichen Trainingsaufwand. Sie werden in wissenschaftlichen Arbeiten selten explizit berücksichtigt, können in der Praxis – v.a. auf lange Sicht – aber durchaus zum „K.O.-Kriterium" werden, etwa wenn ein gutgemeinter und im Ansatz auch effektiver Vorschlag zur Verbesserung von Leistungsbeurteilungen den Vorgesetzten abverlangt, täglich eine umfangreiche Aufzeichnung von Verhaltensbeobachtungen jedes einzelnen Mitarbeiters durchzuführen (positive Aspekte solcher *Verhaltenstagebücher* betonen z. B. Bernardin & Beatty, 1984).

LB-Systeme, die als unrealistisch oder unfair empfunden werden, können bei den Beteiligten zu ernsten negativen Reaktionen führen.

Stärker noch als die Praktikabilität läßt sich die *Akzeptabilität* als zwar nicht hinreichende, aber unbedingt notwendige Bedingung dafür ansehen, daß Leistungsbeurteilungen ihre Funktionen erfüllen können. Akzeptabilität wird gewöhnlich über die Reaktion der Betroffenen (Beurteiler und Beurteilte) operationalisiert, nämlich als das Ausmaß, in dem *Prozeß* und *Ergebnisse* der Beurteilung als gerecht (fair und der Wirklichkeit entsprechend) *empfunden* werden. Unzureichend akzeptierte Beurteilungssysteme können negative Reaktionen provozieren, die von Gleichgültigkeit bis zu offener Reaktanz reichen können und eine wesentliche Ursache der eingangs dieses Kapitels erwähnten grundsätzlichen Vorbehalte gegenüber Leistungsbeurteilungen darstellen. Den bislang eher spärlichen Forschungsergebnissen zu diesem Gebiet (zusammenfassend Dickinson, 1993) zufolge darf mit einer guten Akzeptanz der Leistungsbeurteilung dann gerechnet werden, wenn folgende Bedingungen erfüllt sind: kurze Beurteilungsintervalle; Existenz eines formalen Beurteilungssystems; hohe Sach- und Gesprächsführungskompetenz der Beurteiler (letzteres bedeutet z. B., Kritik nicht gehäuft und nicht eigenschaftsbezogen anzubringen, vgl. Kapitel 16); Gelegenheit für die Beurteilten, Einwände zu äußern; Beurteilungsdimensionen werden als relevant angesehen; auf empfundene Schwächen des Systems wird mit Veränderungen reagiert. Eine Schlüsselrolle für die Zufriedenheit mit Beurteilungssystemen spielt die *Partizipation* der Betroffenen während der Entwicklung und Einführung.

6.3 Vergleich der Beurteilungsverfahren

Die vorwiegend psychometrisch orientierte Forschung zu LB-Verfahren hat nur wenig eindeutige Resultate erbracht.

Landy und Farr (1980) kamen in einem einflußreichen Übersichtsartikel zu dem desillusionierenden Schluß, daß den jahrzehntelangen, ambitionierten Versuchen, Leistungsbeurteilungen durch verfeinerte Skalenformate zu verbessern, kein entscheidender Durchbruch beschieden war. Ein Grund hierfür, dessen Wahrscheinlichkeit durch neuere Befunde immer offensichtlicher wird, mag in der weitgehenden Konzentration dieser Bemühungen auf die Vermeidung von Urteilstendenzen liegen. Wie wir oben gesehen haben, stellen diese „Urteilsfehler" kaum einen geeigneten Indikator der Urteilsqualität dar. Weitgehend vernachlässigt wurde dagegen die vermutlich unterschiedliche Eignung der einzelnen Verfahren für verschiedene Beurteilungszwecke; wir sind hier größtenteils noch auf Plausibilitätsüberlegungen angewiesen. Auf weitere Ursachen für die Unergiebigkeit der bisherigen Forschung wird im nächsten Teilabschnitt noch einzugehen sein.

Tabelle 2 stellt den Versuch dar, vergleichende Untersuchungen – teilweise ergänzt um deduktive Schlußfolgerungen – zur relativen Eignung verschiedener Beurteilungsverfahren in maximaler Weise zu verdichten. Es muß

wohl kaum betont werden, daß die Zusammenfassung zu einem Wert auf einer fünfstufigen Qualitätsskala (+ +, +, O, –, – –) eine grobe Vereinfachung darstellt. Auf die Aufnahme von Urteilstendenzen als Kriterien – und damit auf einen erheblichen Teil der Forschung – wurde aus den geschilderten Gründen verzichtet. Dennoch sollte der Hinweis nicht fehlen, daß bspw. Mittelwerts- und Streuungstendenzen bei Rangordnungsverfahren logisch ausgeschlossen sind und beim Wahlzwangverfahren kaum auftreten können. Außerdem fehlen in der Tabelle die Kriterien Validität und Nutzen; erstere, weil eine sinnvolle Erörterung hier u.a. eine Aufspaltung in einzelne Validitätsfacetten erforderlich macht, die den Rahmen dieses Abschnitts sprengen würde; letzterer, weil er sich nicht für allgemeingültige (wohl aber für einzelfallbezogene) Vergleiche eignet. Bei den Verfahren wurde wegen seines grundsätzlich andersartigen Charakters auf die Aufnahme des MbO verzichtet. Dort, wo zu einzelnen Verfahrens-/Kriterienkombinationen Befunde fehlen, ergeben sich Lücken in der Matrix, was insbesondere für das erst in jüngerer Zeit erschlossene Gebiet der Akkuratheit zutrifft. Ergänzend wurde die Eignung der einzelnen Verfahren für die beiden antipodischen Funktionskomplexe der Verhaltenssteuerung (Verh.strg.) und der administrativen Entscheidungen (adm. Entsch.) aufgenommen.

Tabelle 2:
Evaluation verschiedener Beurteilungsverfahren

	Reliabilität	Akkuratheit	Praktikabilität	Akzeptabilität	Verh.strg.	adm. Entsch.
Graph. Skala	O	–	+	O	O	O
BARS	O	O	O	+	+	O
BOS	O	O	+	O	+	O
MSS	O		+	O	–	O
PDA/BDS	–	O			–	O
direkte Rangreihenbildung	+		O	–	– –	(+)*
Paarvergleich	+ +		+	–	– –	(+)*
Verhaltensrangprofil	+ +		O		+ +	(– –)**
checklist	O		+	–	–	O
Wahlzwangv.	+		–	– –	– –	+

*= nur für Laufbahnentsch., für Entgeltfindung ungeeignet; **= ohne zusätzliche Verwendung einer Einstufungsskala

Generell läßt sich sagen, daß Verfahren, deren innere Logik den Beurteilern verborgen bleibt, für Feedback weniger gut geeignet sind. Ähnliches gilt auch für die Akzeptabilität. Hier zahlt sich die bei der Konstruktion der BARS besonders intensive Partizipation der Beurteiler aus. Insgesamt ist festzuhalten, daß in Tabelle 2 die Unterschiede zwischen den Verfahren, insbesondere bei den „harten“ Kriterien Reliabilität und Akkuratheit, eher etwas überzeichnet dargestellt sind. Im Zweifel sollte sich auch hier die Entscheidung eher an qualitativen und zweckorientierten Kriterien ausrichten. Hier bietet sich ein potentiell fruchtbares, leider aber noch weitgehend brachliegendes Feld für zukünftige Forschung.

Auch bei der Wahl zwischen verschiedenen Skalenformaten gilt es, zunächst nach dem Zweck der Beurteilung zu fragen.

6.4 Der Urteilsprozeß: Modelle, Einflüsse und Eingriffsmöglichkeiten

Als Konsequenz des ernüchternden Fazits ihrer bereits zitierten Literaturübersicht schlugen Landy und Farr (1980) eine stärkere Konzentration der Forschung auf die während der Urteilsbildung ablaufenden Vorgänge und diese beeinflussende Faktoren vor und stellten die ihrer Ansicht nach maßgeblichen Variablen in einem Modell zusammen. Dieses und nachfolgende Modelle des Beurteilerverhaltens (am einflußreichsten: DeNisi, Cafferty & Meglino, 1984; Feldman, 1981) sind vorwiegend kognitivistisch ausgerichtet und orientieren sich, vereinfachend, an der Betrachtung der Beurteilung als *Informationsverarbeitungsprozeß* mit den bekannten Phasen *Wahrnehmung/Informationsaufnahme, -enkodierung, -speicherung* und *-zugriff/-wiedergabe*. Im Verlaufe dieses Prozesses spielen sozialkognitivistische Konstrukte wie Schemata, Kategorien, Prototypen oder Skripten eine zentrale Rolle, auf deren Bedeutung hier nicht im einzelnen eingegangen werden kann (eine knappe Darstellung findet sich bei Murphy & Cleveland, 1995). Die kognitivistische Richtung der Leistungsbeurteilungsforschung hatte in den achtziger Jahren v.a. in den USA einen geradezu dominierenden Einfluß. Obwohl dabei zahlreiche interessante Einsichten zu einzelnen Aspekten zutage gefördert wurden (zusammenfassend Ilgen, Barnes-Farrell & McKellin, 1993), scheint sie jedoch inzwischen ihren Zenit schon wieder überschritten zu haben. Ein Grund hierfür mag – neben dem Mangel an praktischer Umsetzbarkeit der Ergebnisse – in der weitgehenden Nichtberücksichtigung motivationaler, affektiver und situativer Faktoren liegen, obwohl diese in einigen Modellen ansatzweise integriert sind. Der wesentliche Beitrag dieses Forschungszweigs kann darin gesehen werden, die *Grenzen* der Machbarkeit durch technische Verfeinerung der Beurteilungsinstrumente unterstrichen zu haben, so bspw., daß menschliche Beurteiler nur wenige Urteilsdimensionen wirklich unterscheiden können und daß sie sich trotz Verhaltensverankerung an eigenschaftsähnlichen Konzepten orientieren.

Die kognitivistische Richtung der LB-Forschung betrachtet Beurteilungen als Informationsverarbeitungsprozeß.

Ein älteres, gleichwohl umfassenderes, aber weniger geschlossenes Prozeßmodell der Beurteilung wurde von Brandstätter (1969; im folgenden nach der aktualisierten Darstellung bei Schuler, 1989) vorgeschlagen. Darin werden drei „Ebenen" oder Stufen der sozialen Urteilsbildung differenziert, innerhalb derer sich im wesentlichen allgemeinpsychologische Erkenntnisse zu situativen, kognitiven, emotionalen und motivationalen Determinanten der Urteilsbildung integrieren und auf den Spezialfall der betrieblichen Leistungsbeurteilung übertragen lassen:

In dem Modell von Brandstätter (1969) wird bei den Einflüssen auf den Urteilsprozeß zwischen den Ebenen des Verhaltens, des Eindrucks und der Aussage unterschieden.

- Auf der *Ebene des Verhaltens* sind es Charakteristika der *Beurteilten*, der unmittelbaren *Situation* (z. B. das Verhalten anderer Anwesender), aber auch Determinanten aus dem *weiteren Umfeld* (z. B. die familiäre Situation des Beurteilten), deren Kenntnis und individuelle Interpretation seitens des Beurteilers insbesondere dessen Attribution von Verantwortung für das beobachtete Verhalten beeinflußt.
- Auf der *Ebene des Eindrucks* befinden sich Merkmale des Beurteilers, die sich größtenteils verzerrend auf die Informationsverarbeitung auswirken. Dessen *Selbstbild*, seine generellen *Werthaltungen*, *Interessen* oder *Vorinformationen* bestimmen das Urteil ebenso mit wie der *erste Eindruck, implizite Persönlichkeitstheorien* (unpünktlich = unzuverlässig) oder persönliche *Sympathie* und Antipathie, deren Bestätigung später u.U. nach dem Muster der self-fulfilling prophecy eintritt und als prognostische Leistung fehlattribuiert wird.
- Auf der *Ebene der Aussage* wird schließlich das intern gebildete Urteil nach außen transformiert, was nochmals mit erheblichen Änderungen einhergehen kann. Hier spielen, neben der *sprachlichen Kompetenz* des Beurteilers und seinen *Einstellungen zum Beurteilungssystem*, vor allem *Urteilsstrategien* und damit der ganze Komplex mikropolitischen Verhaltens eine Rolle.

Nachdem in den beschriebenen Modellen eine Vielzahl potentieller Störquellen – oder, technisch gesprochen: Quellen der Fehlervarianz – identifiziert wurde, stellt sich die Frage nach Maßnahmen zu deren Eliminierung, oder anders gesagt: zur Standardisierung der Urteile. Neben der Veränderung des Instrumentariums, auf deren Grenzen bereits eingegangen wurde, kommt hier vor allem die *Einwirkung auf die Person des Beurteilers* in Betracht. Ein potentiell fruchtbarer, aber in der Praxis fast nie realisierter Ansatz bestünde darin, *Anreize* für korrekte Beurteilungen zu schaffen. Sehr viel häufiger wird der Weg des *Beurteilertrainings* gewählt. Im wesentlichen lassen sich vier Arten von Beurteilertrainings unterscheiden:

- *Urteilsfehlertraining:* Die älteste Kategorie ist auf die Vermeidung von Urteilstendenzen gerichtet.
- *Leistungsdimensionentraining:* Dieser kognitivistisch beeinflußte Typus versucht, die Urteilsbildung stärker an den Leistungsdimensionen zu orientieren.
- *Verhaltensbeobachtungstraining:* Hiermit wird versucht, bereits bei der Wahrnehmung, also vor der Urteilsbildung, einen gewissen Standardisierungsgrad zu vermitteln.
- *Bezugsrahmentraining:* Das umfassendste Trainingskonzept soll den Beurteilern eine gemeinsame Vorstellung von Leistungsdimensionen und -standards vermitteln, innerhalb deren sie ihre Beobachtung im Sinne eines Bezugsrahmens (frame of reference) einordnen können.

Entgegen früheren Untersuchungen, denen zufolge Urteilsfehlertraining zwar die Urteilstendenzen, aber auch die Urteilsgenauigkeit vermindert, fand eine Metaanalyse von Woehr und Huffcutt (1994) eine positive Wirkung auch auf die Akkuratheit, wenn auch nur mit geringer Effektstärke. Noch mäßiger fiel der Erfolg von Leistungsdimensionentrainings aus. Weitaus wirksamer zeigten sich Verhaltensbeobachtungs- und vor allem Bezugsrahmentrainings. Insgesamt – und besonders für die Beobachtungstrainings – sind die Befunde der Metaanalyse wegen ihrer vglw. geringen Fallzahlen vorsichtig zu interpretieren.

Die Wirkung von Beurteilertrainings ist i.d.R. eher gering, wobei Unterschiede zwischen verschiedenen Trainingsformen bestehen.

7 Praktische Aspekte der Leistungsbeurteilung

7.1 Konstruktion und Einführung von Beurteilungssystemen

Die Einführung formaler Beurteilungen in einem Unternehmen ist ein komplexes, innovatives Unterfangen, in dessen Verlauf personelle und sachliche Ressourcen wechselnder Quantität und Qualität gebunden werden. Es hat sich bewährt, mit der Koordination und Durchführung solcher komplexer Aufgabenstellungen *Projektgruppen* zu betrauen, deren genaue Zusammensetzung und organisatorische Einbindung u.a. davon abhängig gemacht werden sollte, in welcher Phase sich das Projekt jeweils befindet. Dies kann den Wechsel einiger direkt involvierter Mitarbeiter und eventuell die Hinzuziehung externer Berater erfordern. Eine über die betriebsverfassungsrechtlich (siehe unten) gebotene Einbeziehung ihrer Interessenvertreter (v.a. des Betriebsrats) hinausgehende *direkte Partizipation* der Betroffenen zahlt sich langfristig in höherer Akzeptanz und damit einem besseren Funktionieren des Systems aus. Ein spürbares Engagement der Unternehmensleitung unterstreicht die Bedeutung des Vorhabens.

Die Einführung eines LB-Systems ist eine komplexe Aufgabe, für die sich die Bildung einer Projektgruppe und eine umfangreiche Partizipation der Betroffenen empfiehlt.

Schuler (1991) schlägt eine Gliederung des *Konstruktionsprozesses* für ein neues, formales Beurteilungssystem in zehn Schritte vor:

1. *Bestandsaufnahme:* Analyse vorhandener Beurteilungsverfahren und Rahmenbedingungen

2. *Zielformulierung:* Partizipative Festlegung der wichtigsten angestrebten Funktionen
3. *Kosten-/Nutzen-Kalkulation:* Investitionsrechnung auf der Basis der geschätzten Validität und Leistungsvarianz; Abschätzung sozialer Wirkungen
4. *Zielgruppen:* Festlegung der Beurteiler und der zu Beurteilenden: Klärung von Partizipations- und Akzeptanzfragen
5. *Arbeitsanalyse:* Ermittlung der wichtigen Tätigkeiten und ihrer Verhaltensanforderungen
6. *Beurteilungskriterien:* Bestimmung der Ebenen und Maße; Ableitung der wichtigsten Kriterien aus der Arbeitsanalyse
7. *Skalierungsverfahren:* Wahl der Methode(n) entsprechend den Zielsetzungen und Möglichkeiten
8. *Skalenkonstruktion:* Sammlung und Zuordnung von Einzelaussagen zu Beurteilungskriterien; statistische Überprüfung
9. *Probeverwendung:* Erprobung an repräsentativen Gruppen; Auswertung und gegebenenfalls Modifikation
10. *Beurteilertraining:* Training bezüglich der Urteilsprozesse, der Verfahrensanwendung, der Gesprächsführung und Zielsetzung

Sparsamkeit bei der Konstruktion von LB-Verfahren zahlt sich i.d.R. nicht aus.

Es ist angesichts dieses recht aufwendig anmutenden Vorgangs davor zu warnen, einzelne Schritte für verzichtbar zu halten oder zu glauben, – etwa durch Übernahme eines anderswo schon entwickelten Systems – den gesamten Prozeß einsparen zu können. Mehr noch als in Mängeln der rein technischen Funktionalität könnte sich solche „Sparsamkeit" über die Köpfe der Betroffenen hinweg in negativen Einstellungen bis hin zu Reaktanz niederschlagen, was letztlich den Nutzen des gesamten Systems in Frage stellt.

7.2 Handhabung von Beurteilungssystemen

Auch ein sorgfältig konstruiertes Beurteilungssystem kann nach seiner Einführung keineswegs als „Selbstläufer" betrachtet werden. Wie alle komplexen Systeme erfordert es ständige Pflege und gegebenenfalls Revisionen, wenn sich herausstellt, daß die angestrebten Ziele damit nicht adäquat verfolgt werden können. Bevor sich dies gewissermaßen schleichend in einem allgemeinen Gefühl des Unbehagens äußert, empfiehlt es sich, von Zeit zu Zeit eine formale *Evaluation* (vgl. Kapitel 22) durchzuführen, wofür bspw. *Mitarbeiterbefragungen* (vgl. Kapitel 14) eine methodische Option darstellen. Aber auch eine systematische *Auswertung der Beurteilungsergebnisse* kann bereits einigen Aufschluß über Schwachstellen liefern, wie das Beispiel der folgenden Fallstudie zeigt.

LB-Systeme erfordern auch nach ihrer Einführung ständige Pflege, gelegentliche Evaluation und gegebenenfalls Revisionen.

Fallstudie zur Personalbeurteilung

Fallstudie:
Personalbeurteilung bei Mercedes-Benz, Werk Gaggenau (nach Watzka, 1995)

Zum Zeitpunkt der Erhebung verfügte die Mercedes-Benz AG über ein unternehmensweites, positionsübergreifendes Beurteilungssystem, das auch für die ca. 9.500 Beschäftigten des zur Nutzfahrzeugsparte gehörenden Werkes Gaggenau galt. Zu beurteilen war jeweils eine positionsspezifische Auswahl von 8 bis 13 Merkmalen aus insgesamt 17 Dimensionen, die überwiegend deutlichen Eigenschaftscharakter hatten (Initiative, Zuverlässigkeit, Einsatzbereitschaft etc.). Die Beurteiler hatten die Möglichkeit, die vorgegebenen Merkmalskataloge durch Auslassung oder Hinzufügung von maximal zwei Dimensionen individuell zu variieren. Die Skalen entsprachen formal einer 7-stufigen graphischen Einstufungsskala, wobei die einzelnen Skalenstufen durch verhaltensnahe Verankerungen erläutert waren, deren dimensionale Abgrenzung nicht immer als gelungen bezeichnet werden konnte. Da die Leistungsbeurteilung u.a. als

Grundlage der Verteilung einer betrieblichen Leistungszulage mit festgelegter Gesamtsumme diente, war der werksweite Durchschnitt der Beurteilungsergebnisse per Betriebsvereinbarung auf die Skalenmitte, gemittelt über alle Dimensionen, festgesetzt. Das Skalenformat verband also ein Einstufungsverfahren mit einem Element der Quotenvorgabe; die Kriterien vermischten die Eigenschafts- mit der Verhaltensebene. Grundlage der Zulagenbemessung war ein über alle Dimensionen gemittelter Gesamtbeurteilungswert. *Gleichzeitig* sollten die Beurteiler (i.d.R. direkte Vorgesetzte) auf der Rückseite des Beurteilungsformulars in freier Eindrucksschilderung konkrete Verhaltensbeobachtungen aufzeichnen, die der Verhaltenssteuerung und als Basis der Personalentwicklung dienen sollten. Die gesamte Beurteilung wurde den betroffenen Mitarbeitern in einem jährlichen Beurteilungsgespräch eröffnet, im Rahmen dessen sie auch Gelegenheit zur Stellungnahme hatten.

Im Zuge der Übertragung der Prinzipien eines formalen betriebswirtschaftlichen Controlling auf den Personalbereich wurde das Beurteilungssytem in Gaggenau 1992, 18 Jahre nach seiner Einführung (!), einer systematischen Auswertung seiner Ergebnisse unterzogen. Neben einer gewissen Streuungseinschränkung, die sich auch als statistischer Effekt der Mittelung über alle Dimensionen interpretieren läßt, zeigten sich v.a. folgende Resultate: (1) Der einzelne Mitarbeiter wurde im großen und ganzen auf allen Kriterien gleich eingestuft (Halo); (2) Einzelne Mitarbeiter wurden zwar von Periode zu Periode gelegentlich herauf-, aber kaum jemals zurückgestuft; der festgelegte Mittelwert wurde offensichtlich durch schlechtere Beurteilungen für neue Mitarbeiter eingehalten; (3) Die freien Eindrucksschilderungen waren alles andere als konkret; sie gaben i.d.R. keinerlei Verhaltenshinweise und auch keinen Aufschluß über individuelle Stärken und Schwächen. Aufschlußreich ist hier insbesondere der letztgenannte Befund. Offensichtlich ist es nicht gelungen, die Ziele Gehaltszuteilung, Verhaltenssteuerung und Personalentwicklung mit *einem* Beurteilungsinstrument zu erfüllen. Die – hier tarifvertraglich vorgeschriebene und per Betriebsvereinbarung spezifizierte – Funktion der Verteilung von Leistungszulagen dominierte so eindeutig, daß die Führungsfunktion (der Beurteilung, nicht der Zulage) und in der Folge der diagnostische und prognostische Wert der Beurteilung weitgehend verdrängt wurden. Aber auch für die Gehaltszuteilung scheint sich neben das eigentlich intendierte Leistungsprinzip ein Element des Senioritätsprinzips gesellt zu haben.

Bei der Handhabung von LB-Systemen sollten zumindest die Ebenen des day-to-day-Feedback, der Regelbeurteilung und der Potentialbeurteilung auseinandergehalten werden.

Die Lehren aus der obigen Fallstudie unterstreichen eine Erkenntnis, die schon weiter oben immer wieder hervorgehoben wurde: Unvereinbare Ziele erfordern unabhängige Vorgehensweisen. Schuler (1991) hat vorgeschlagen, zumindest drei *Ebenen der Beurteilung* zu unterscheiden und in der Handhabung auseinanderzuhalten, die zusammen ein vollständiges System der Leistungsbeurteilung bilden:

Tabelle 3:
Die drei Ebenen der Beurteilung (nach Schuler, 1991)

Ebene	Funktion	Verfahrensweise
1. Ebene Day-to-day-Feedback	Verhaltenssteuerung Lernen	Gespräch Unterstützung
2. Ebene Regelbeurteilung	Leistungseinschätzung Zielsetzung	Systematische Beurteilung Beurteilungs-gespräch
3. Ebene Potentialbeurteilung	Fähigkeitseinschätzung Prognose	Eignungsdiagnose Assessment Center

Auf der *1. Ebene* geht es nicht um systematische Vergleichbarkeit und Standardisierung, sondern um ein verhaltensnahes, an konkreten Beobachtungen orientiertes Gespräch, das zweckmäßigerweise nicht erst nach Ablauf einer Regelperiode erfolgen sollte („day-to-day"!) und in dem Kritik nicht mit unmittelbaren Konsequenzen für Gehalt oder Laufbahn verknüpft ist. Erforderlich ist aber auch hier die Kenntnis der Ziele und erfolgskritischen Verhaltensweisen sowie die Fähigkeit des Beurteilers, diese in konstruktiver Weise zu vermitteln. Sehr viel höher sind die Ansprüche an Vergleichbarkeit und Systematik auf der *2. Ebene*, auf der die Zielerreichung und -vereinbarung sowie daran geknüpfte Konsequenzen im Mittelpunkt stehen. Der Beurteilte sollte auch hier ausführlich Gelegenheit erhalten, seine Sicht der Dinge zu schildern. Eine reine Ergebnisdiskussion ohne die Erarbeitung zielführender Verhaltensänderungen ist wenig fruchtbar; zu vermeiden ist auf dieser Ebene jedoch die Einbeziehung stabiler Dispositionen wie Eigenschaften und Fähigkeiten. Diese sind Gegenstand der *3. Ebene*, wo es um die langfristige Prognose des Potentials mit dem Ziel der Laufbahnplanung geht. Die methodischen Ansprüche sind hier eher noch höher als auf Ebene 2, so daß hier als Ergänzung zum Rückgriff auf Erfahrungswerte aus der unmittelbaren Zusammenarbeit auch eigenschafts- und simulationsorientierte Verfahren der Eignungsdiagnostik wie Tests oder Assessment Center in Betracht kommen (vgl. Kapitel 5 und 6). Daß die Gesprächsführung bei dieser für den einzelnen existentiellen Thematik ein besonders ausgeprägtes Fingerspitzengefühl erfordert, versteht sich von selbst (vgl. Kapitel 16).

Auf jeder Ebene werden unterschiedliche Funktionen mit unterschiedlichen Mitteln erfüllt, wobei der methodische Anspruch von Ebene zu Ebene ansteigt.

7.3 Rechtliche Aspekte

Als gesetzliche Grundlage der rechtlichen Regelung von Leistungsbeurteilungen ist in Deutschland in erster Linie das *Betriebsverfassungsgesetz* von 1972 (BetrVG; vgl. Jedzig, 1991a, b) einschlägig. Daneben sind in einzelnen Branchen (z. B. Metallindustrie) *Tarifverträge* zu beachten. Die nähere, betriebsindividuelle Ausgestaltung erfolgt i.d.R. im Rahmen einer *Betriebsvereinbarung* (§ 77 BetrVG) zwischen Betriebsrat – eventuell auch Gesamtbetriebsrat – und Arbeitgeber.

Wichtigste Einzelnorm ist § 94 Abs.2 BetrVG (i.V.m. § 77 BetrVG). Danach obliegt die *Entscheidung* über die Einführung allgemeiner Beurteilungsgrundsätze ausschließlich dem Arbeitgeber. Ein Recht auf erzwingbare Initiative besteht insoweit für den Betriebsrat nicht. Dieser hat allerdings ein Mitbestimmungsrecht (Zustimmung) bei der *inhaltlichen Ausgestaltung* bezüglich der Beurteilungsmerkmale, -verfahren und Bewertungsstufen. Dient die Beurteilung auch zur Ermittlung von Entgeltbestandteilen (Leistungszulagen), so verdrängt die inhaltlich weitergehende Mitbestimmung aus § 87 Abs.1 Nr. 10 BetrVG (Initiativrecht) den § 94 als Gesetzesgrundlage, sofern die fraglichen Grundsätze nicht durch Tarifvertrag abschließend und zwingend geregelt werden (sog. Tarifvorrang). Ein Mitbestimmungsrecht besteht auch bezüglich der *technischen Erhebung* und *Speicherung* von Leistungsdaten (§ 87 Abs.1 Nr.6 BetrVG). Für die *Durchführung* von Beurteilungen ist wiederum allein der Arbeitgeber verantwortlich, der dabei aber nicht gegen Betriebsvereinbarungen verstoßen darf (Kontroll- und Einspruchsrechte des Betriebsrats). Für *Potentialbeurteilungen* gelten z.T. andere Regelungen (§§ 94 Abs.1, 95 BetrVG), die dem eignungsdiagnostischen Charakter dieser Verfahren Rechnung tragen (vgl. Jedzig, 1996).

Vor allem bei der inhaltlichen Ausgestaltung von LB-Systemen sind Mitbestimmungsrechte des Betriebsrats zu beachten.

Neben diesen kollektiven Mitbestimmungsrechten stehen dem einzelnen Arbeitnehmer *individuelle Rechte* aus §§ 81 ff. BetrVG zu. Neben dem Recht auf Einsichtnahme in die Personalakten (§ 83 BetrVG) besteht ein Anspruch auf Erörterung der Beurteilung und beruflichen Entwicklungsmöglichkeiten (§ 82 Abs.2 BetrVG) sowie auf eine Beschwerdemöglichkeit gegen die Ergebnisse der Beurteilung (§ 84 BetrVG). Der Arbeitnehmer kann dabei jeweils die Hinzuziehung eines Betriebsratsmitglieds verlangen. In der Praxis wird diesen Ansprüchen regelmäßig durch ein obligatorisches Beurteilungsgespräch Rechnung getragen.

Neben der kollektiven Mitbestimmung bestehen auch individuelle Arbeitnehmerrechte auf Einsichtnahme, Erörterung und Beschwerde.

Zusammenfassung

Zusammenfassung

Die Lektüre dieses Kapitels sollte gezeigt haben, daß betriebliche Leistungsbeurteilungen ein komplexes Gebiet sind, auf dem durch Pauschalurteile oder Patentrezepte wenig Erkenntnisgewinn zu erzielen ist. Die Forschung zum Konstrukt der Berufsleistung und der Urteilsprozesse, die zu deren Einschätzung durch menschliche Beurteiler führen, deutet zu ihrem gegenwärtigen Stand darauf hin, daß Leistung zwar eine mehrdimensionale Struktur aufweist, eine begrenzte Anzahl von Inhalten aber in unterschiedlichen Positionen wiederkehrt und es wenig Sinn hat, über diese Inhalte hinaus noch sehr stark differenzieren zu wollen. Die Forderung nach anforderungsbezogenen reinen Verhaltensurteilen erscheint vor diesem Hintergrund weder einlösbar noch sinnvoll – zumindest nicht im Rahmen jährlicher Regelbeurteilungen. Ein differenziertes Feedback mit dem Ziel der Verhaltenssteuerung, das dem jeweiligen Einzelfall in all seinen Facetten gerecht wird, läßt sich wesentlich besser unmittelbar nach Verhaltensbeobachtung in einem formlosen Gespräch übermitteln und diskutieren.

Bevor Detailentscheidungen im Zusammenhang mit der Einführung eines Beurteilungssystems getroffen werden können, ist es unbedingt erforderlich, sich über die damit verfolgten Ziele klar zu werden und dabei zumindest die unvereinbaren Funktionen der Unterstützung administrativer Entscheidungen über Laufbahn oder Entgelt und persönliche Verhaltenssteuerungs- und Entwicklungsziele zu trennen. Dabei kann es hilfreich sein, je nach Zweck der Beurteilung, neben dem direkten Vorgesetzten auch andere Quellen wie Kollegen, Mitarbeiter oder den Beurteilten selbst, möglicherweise auch Außenstehende und objektive Daten einzubeziehen. Auch bezüglich der Angemessenheit verschiedener Skalenformate bestehen Unterschiede zwischen den Funktionen. Der Aufwand zur Konstruktion einer formalen Beurteilungsskala kann beträchtlich sein, zahlt sich aber u.U. durch eine erhöhte Akzeptanz des Systems aus, da in ihrem Verlauf vielfach die Notwendigkeit zur Partizipation der Betroffenen besteht. Die Möglichkeiten zur Verbesserung der psychometrischen Eigenschaften durch stärkere Annäherung an beobachtetes Verhalten sind dagegen durch die Natur des menschlichen Urteilsprozesses beschränkt. Die Rolle sog. Urteilstendenzen wie Halo oder Mildetendenz in diesem Zusammenhang wurde in der Vergangenheit vermutlich überschätzt.

Die Einführung eines formalen Beurteilungssystems ist ein organisatorisch komplexes Vorhaben, bei dem sich vermeintliche Einsparungen bei einzelnen Konstruktionsschritten, dem Einsatz von Mitarbeiterressourcen und der Partizipation aller Betroffenen langfristig eher kontraproduktiv auswirken dürften. Zudem ist die Leistungsbeurteilung in Deutschland mitbestimmungspflichtig, und es sind individuelle Arbeitnehmerrechte zu beachten. Ein erheblicher Teil potentieller Konflikte und Schwierigkeiten kann vermieden werden, wenn in der Handhabung zwischen den Ebenen der unmittelbaren Verhaltensrückmeldung, der Regel- und der Potentialbeurteilung getrennt wird. Ein umfassendes Beurteilungssystem, bei dem die aufgeführten Grundprinzipien beachtet werden, kann ein hocheffizientes personalpolitisches Instrument zur Führung und Entscheidungsunterstützung sein.

Weiterführende Literatur

Weiterführende Literatur

Bernardin, H.J. & Beatty, R.W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Borman, W.C. (1991). Job behavior, performance, and effectiveness. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology*, (2nd. ed., Vol. 2, pp. 271-326). Palo Alto: Consulting Psychologists Press.

Weiterführende Literatur

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives.* Thousand Oaks: Sage.

Schuler, H. (1989). Leistungsbeurteilung. In E. Roth (Hrsg.), *Organisationspsychologie. Enzyklopädie der Psychologie D/III/3* (S. 399-430). Göttingen: Hogrefe.

Schuler, H. (Hrsg.). (1991). *Beurteilung und Förderung beruflicher Leistung.* Göttingen: Hogrefe/Verlag für Angewandte Psychologie.

Literatur

Literatur

Aiello, J.R. & Kolb, K.J. (1995). Electronic performance monitoring and social context: Impact on productivity and stress. *Journal of Applied Psychology, 80*, 339-353.

Balzer, W.K. & Sulsky, L.M. (1990). Performance appraisal effectiveness. In K.R. Murphy & F. Saal (Eds.), *Psychology in organizations: Integrating science and practice.* Hillsdale: Lawrence Erlbaum.

Becker, F.G. (1994). *Grundlagen betrieblicher Leistungsbeurteilungen* (2., durchges. Aufl.). Stuttgart: Schäffer-Poeschel.

Bernardin, H.J. & Beatty, R.W. (1984). *Performance appraisal: Assessing human behavior at work.* Boston: Kent.

Bernardin, H.J., Dahmus, S.A. & Redmon, G. (1993). Attitudes of first-line supervisors toward subordinate appraisals [special issue]. *Human Resource Management, 32*, 315-324.

Blanz, F. & Ghiselli, E.E. (1972). The mixed standard scale: A new rating system. *Personnel Psychology, 25*, 185-199.

Bommer, W.H., Johnson, J.L., Rich, G.A., Podsakoff, P.M. & MacKenzie, S.B. (1995). On the interchangeabilty of objective and subjective measures of employee performance: A meta-analysis. *Personnel Psychology, 48*, 587-605.

Borg, I. & Staufenbiel, T. (1993). *Theorien und Modelle der Skalierung* (2. rev. Auflage). Bern: Huber.

Borman, W.C. (1987). Personal constructs, performance schemata, and „folk theories" of subordinate effectiveness: Explorations in an army officer sample. *Organizational Behavior and Human Decision Processes, 40*, 307-322.

Borman, W.C. (1991). Job behavior, performance, and effectiveness. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd. ed., Vol. 2, pp. 271-326). Palo Alto: Consulting Psychologists Press.

Borman, W.C. & Motowidlo, S.J. (1993). Expanding the criterion domain to include elements of contextual performance. In N. Schmitt & W.C. Borman (Eds.), *Personnel selection in organizations* (pp. 71-98). San Francisco: Jossey-Bass.

Borman, W.C. & Motowidlo, S.J. (1997). Task performance and contextual performance: The meaning for personnel selection research. *Human Performance, 10*, 99-110.

Brandstätter, H. (1969). *Soziale Urteilsbildung in Organisationen.* Unveröff. Habil., Ludwig-Maximilians-Universität München.

Brandstätter, H. (1970). Die Beurteilung von Mitarbeitern. In A. Mayer & B. Herwig (Hrsg.), *Handbuch der Psychologie, Bd.9: Betriebspsychologie* (S. 668-734). Göttingen: Hogrefe.

Brandstätter, H. & Schuler, H. (1974). *Overcoming halo and leniency: A new method of merit rating.* Vortrag zum 18th International Congress of Applied Psychology, Montreal.

Brief, A.P. & Motowidlo, S.J. (1986). Prosocial organizational behaviors. *Academy of Management Review, 11*, 710-725.

Campbell, J.P. (1990a). An overview of the army selection and classification project (Project A) [special issue]. *Personnel Psychology, 43*, 231-239.

Campbell, J.P. (1990b). Modeling the performance prediction problem in industrial and organizational psychology. In M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd. ed., Vol. 1, pp. 687-732). Palo Alto: Consulting Psychologists Press.

Fortsetzung Literatur

Campbell, J.P., McCloy, R.A., Oppler, S.H. & Sager, C.E. (1993). A theory of performance. In N. Schmitt & W.C. Borman (Eds.), *Personnel selection in organizations* (pp. 35-70). San Francisco: Jossey-Bass.

Campbell, J.P., McHenry, J.J. & Wise, L.L. (1990). Modeling job performance in a population of jobs [special issue]. *Personnel Psychology, 43*, 313-333.

Cleveland, J.N., Murphy, K.R. & Williams, R.E. (1989). Multiple uses of performance appraisal: Prevalence and correlates. *Journal of Applied Psychology, 74*, 130-135.

Conway, J.M. (1996). Additional construct validity evidence for the task/contextual performance distinction. *Human Performance, 9*, 309-329.

Conway, J.M. & Huffcutt, A.I. (1997). Psychometric properties of multisource performance ratings: A meta-analysis of subordinate, supervisor, peer, and self-ratings. *Human Performance, 10*, 331-360.

Crisand, E. & Stephan, P. (1994). *Personalbeurteilungssysteme: Ziele, Instrumente, Gestaltung.* Heidelberg: Sauer.

Cronbach, L.J. (1955). Processes affecting scores on „understanding of others" and „assumed similarity". *Psychological Bulletin, 52*, 177-193.

Deming, W.E. (1986). *Out of the crisis.* Cambridge: MIT Institute for Advanced Engineering Study.

DeNisi, A.S., Cafferty, T. & Meglino, B. (1984). A cognitive view of performance appraisal process: A model and research propositions. *Organizational Behavior and Human Performance, 33*, 360-396.

Dickinson, T.L. (1993). Attitudes about performance appraisal. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 141-161). Hillsdale: Erlbaum.

Domsch, M.E. & Gerpott, T.J. (1992). Personalbeurteilung. In E. Gaugler und W. Weber (Hrsg.), *Handwörterbuch des Personalwesens* (2., neugest. Aufl., Sp. 1631-1641). Stuttgart: Schäffer-Poeschel.

Donat, M. (1991). Selbstbeurteilung. In H. Schuler (Hrsg.), *Beurteilung und Förderung beruflicher Leistung* (S. 135-145). Göttingen: Hogrefe/Verlag für Angewandte Psychologie.

Drucker, P.F. (1954). *The practice of management.* New York: Harper.

Dunnette, M.D. (1993). My hammer or your hammer? [special issue]. *Human Resource Management, 32*, 373-384.

Feldman, J.M. (1981). Beyond attribution theory: Cognitive processes in performance appraisal. *Journal of Applied Psychology, 66*, 127-148.

Flanagan, J.C. (1954). The critical incidents technique. *Psychological Bulletin, 51*, 327-358.

Gerpott, T.J. & Domsch, M.E. (1995). Personalbeurteilung von Führungskräften. In A. Kieser (Hrsg.), *Handwörterbuch der Führung* (2., neugest. Aufl., Sp. 1694-1704). Stuttgart: Schäffer-Poeschel.

Guzzo, R.A., Jette, R.D. & Katzell, R.A. (1985). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel Psychology, 38*, 275-292.

Harris, M.M. & Schaubroeck, J. (1988). A meta-analysis of self-supervisory, self-peer, and peer-supervisor ratings. *Personnel Psychology, 41*, 43-62.

Hattrup, K., O'Connell, M.S. & Wingate, P.H. (1998). Prediction of multidimensional criteria: Distinguishing task and contextual performance. *Human Performance, 11*, 305-319.

Hunt, J.W. (1995). Das 360-Grad-Feedback: Neue Instrumente zur Kaderbeurteilung. *gdi-impuls, o.J. (3)*, 40-53.

Hunt, S.T. (1996). Generic work behavior: An investigation into the dimensions of entry-level, hourly job performance. *Personnel Psychology, 49,* 51-83.

Ilgen, D.R. (1993). Performance appraisal accuracy: An illusive and sometimes misguided goal. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment: Individual and organizational perspectives* (pp. 235-252). Hillsdale: Lawrence Erlbaum.

Ilgen, D.R., Barnes-Farrell, J.L. & McKellin, D.B. (1993). Performance appraisal process research in the 1980s: What has it contributed to appraisals in use? *Organizational Behavior and Human Decision Processes, 54*, 321-368.

Ilgen, D.R. & Hollenbeck, J.R. (1991). The structure of work: Jobs and roles. In

Fortsetzung Literatur

M.D. Dunnette & L.M. Hough (Eds.), *Handbook of industrial and organizational psychology* (2nd. ed., Vol. 2, pp. 165-207). Palo Alto: Consulting Psychologists Press.

Jedzig, J. (1991a). Einführung standardisierter Verfahren zur Leistungsbeurteilung von Arbeitnehmern. *Der Betrieb, 44*, 753-758.

Jedzig, J. (1991b). Mitbestimmung des Betriebsrats bei der Durchführung von Betriebsvereinbarungen über Leistungsbeurteilung von Arbeitnehmern. *Der Betrieb, 44*, 859-864.

Jedzig, J. (1996). Mitbestimmung bei Einführung von Verfahren zur Potentialanalyse von Arbeitnehmern. *Der Betrieb, 49*, 1337-1342.

Jochum, E. (1991). Gleichgestelltenbeurteilung – ein Instrument der Personalführung und Teamentwicklung. In H. Schuler (Hrsg.), *Beurteilung und Förderung beruflicher Leistung* (S. 107-134). Göttingen: Hogrefe/Verlag für Angewandte Psychologie.

Kane, J.S. (1986). Performance distribution assessment. In R. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 237-273). Baltimore: Johns Hopkins University Press.

Kavanagh, M.J. (1971). The content issue in performance appraisal: A review. *Personnel Psychology, 24*, 653-668.

Knauft, E.B. (1948). Construction and use of weighted checklist rating scales for two industrial situations. *Journal of Applied Psychology, 32*, 63-70.

Landy, F.J. & Farr, J.L. (1980). Performance rating. *Psychological Bulletin, 87*, 72-107.

Latham, G.P. & Wexley, K. N. (1977). Behavioral observation scales for performance appraisal purposes. *Personnel Psychology, 30*, 255-268.

Locke, E.A. & Latham, G.P. (1990). *A theory of goal setting and task feedback.* Englewood Cliffs: Prentice Hall.

London, M. & Beatty, R.W. (1993). 360-degree feedback as a competitive advantage [special issue]. *Human Resource Management, 32*, 353-372.

Longenecker, G.O., Sims, H.P. & Gioia, D.A. (1987). Behind the mask: The politics of employee appraisal. *Academy of Management Executive, 1*, 183-193.

Mabe, P.A. & West, S.G. (1982). Validity of self-evaluation of ability: A review and meta-analysis. *Journal of Applied Psychology, 62*, 280-296.

Macharzina, K. (1993). *Unternehmensführung: das internationale Managementwissen.* Wiesbaden: Gabler.

McCloy, R.A., Campbell, J.P. & Cudeck, R. (1994). A confirmatory test of a model of performance determinants. *Journal of Applied Psychology, 79*, 493-505.

Moser, K. & Krauß, S. (1997). *A meta-analysis of self-supervisory ratings*. Manuscript in preparation.

Motowidlo, S.J. & Van Scotter, J.R. (1994). Evidence that task performance should be distinguished from contextual performance. *Journal of Applied Psychology, 79*, 475-480.

Mount, M.K. (1984). Psychometric properties of subordinate ratings of managerial performance. *Personnel Psychology, 37*, 687-702.

Murphy, K.R. & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*, 619-624.

Murphy, K.R. & Cleveland, J.N. (1995). *Understanding performance appraisal: Social, organizational, and goal-based perspectives*. Thousand Oaks: Sage.

Neuberger, O. (1980). Rituelle (Selbst-)Täuschung: Kritik der irrationalen Praxis der Personalbeurteilung. *Die Betriebswirtschaft, 40*, 27-43.

Nußbaum, A. (1987). Das Modell der Generalisierbarkeitstheorie. In K.J. Klauer (Hrsg.), *Kriteriumsorientierte Tests*. Göttingen: Hogrefe.

Organ, D.W. (1988). *Organizational Citizenship Behavior: The good soldier syndrome*. Lexington: Lexington Books.

Schmidt, F.L. & Hunter, J.E. (1992). Development of a causal model of processes determining job performance. *Current Directions in Psychological Science, 1*, 89-92.

Schmidt, F.L. & Kaplan, L.B. (1971). Composite vs. multiple criteria: A review and resolution of the controversy. *Personnel Psychology, 24*, 419-434.

Schuler, H. (1989). Leistungsbeurteilung. In E. Roth (Hrsg.), *Organisationspsychologie. Enzyklopädie der Psychologie D/III/3* (S. 399-430). Göttingen: Hogrefe.

Schuler, H. (1991). Leistungsbeurteilung – Funktionen, Formen und Wirkungen. In H. Schuler (Hrsg.), *Beurteilung und Förderung beruflicher Leistung* (S. 11-40). Göttingen: Hogrefe/Verlag für Angewandte Psychologie.

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

Sisson, E.D. (1948). Forced choice: The new army rating. *Personnel Psychology, 1*, 365-381.

Smith, P.C. & Kendall, L.M. (1963). Retranslation of expectations: An approach to the construction of unambiguous anchors for rating scales. *Journal of Applied Psychology, 47*, 149-155.

Steel, R.P. & Mento, A.J. (1986). Impact of situational constraints on subjective and objective criteria of managerial job performance. *Organizational Behavior and Human Decision Processes, 37*, 254-265.

Thorndike, R.L. (1949). *Personnel Selection: Test and measurement technique*. New York: Wiley.

Varma, A., DeNisi, A.S. & Peters, L.H. (1996). Interpersonal affect and performance appraisal: A field study. *Personnel Psychology, 49*, 341-360.

Viswesvaran, C. (1993). *Modeling job performance: Is there a general factor?* Unpublished doctoral dissertation, University of Iowa, Iowa City.

Viswesvaran, C., Ones, D.S. & Schmidt, F.L. (1996). Comparative analysis of the reliability of job performance ratings. *Journal of Applied Psychology, 81*, 557-574.

Watzka, K. (1995). Controlling der Handhabung von Leistungsbeurteilungen – Ein Beispiel aus dem Werk Gaggenau der Mercedes-Benz AG. In T.J. Gerpott & S.H. Siemers (Hrsg.), *Controlling von Personalprogrammen* (S. 175-209). Stuttgart: Schäffer-Poeschel.

Woehr, D.J. & Huffcutt, A.I. (1994). Rater training of performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology, 67*, 189-205.

Fortsetzung Literatur