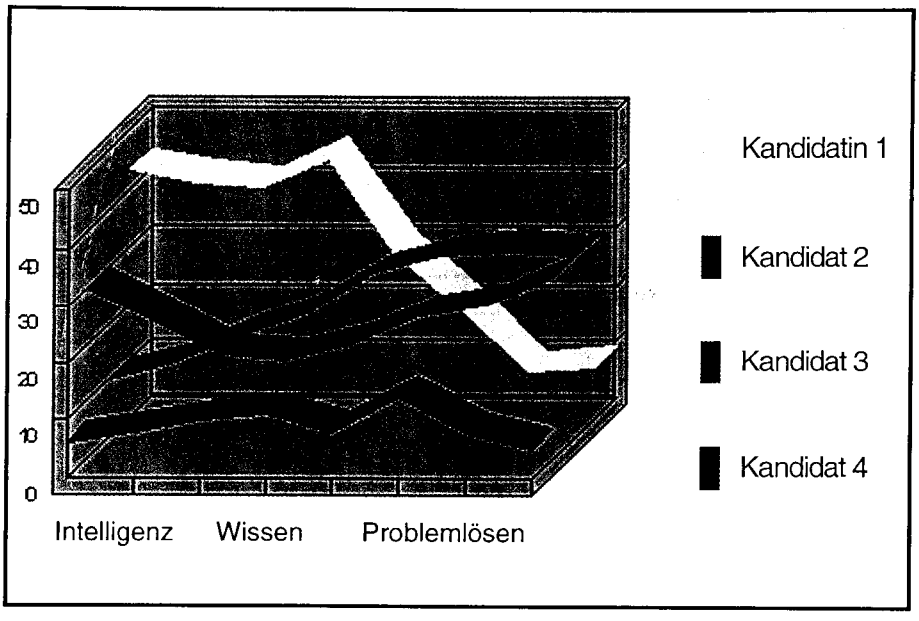
Kersting, M. (1999). Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien? Eine Erörterung und ein empirischer Vergleich der Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests. Göttingen: Hogrefe. [289 Seiten (ISBN 3-8017-1259-1)]

Martin Kersting

Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien?

Zur Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests



Hogrefe

Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien?

Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien?

Zur Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests

von
Martin Kersting

Hogrefe · Verlag für Psychologie
Göttingen · Bern · Toronto · Seattle

Dr. Martin Kersting, geb. 1964. 1985-1991 Studium der Psychologie an der Freien Universität Berlin. Seit 1991 Mitarbeiter der Deutschen Gesellschaft für Personalwesen e.V. *Arbeitsschwerpunkte:* Differentielle und Diagnostische Psychologie, Testkonstruktion, Personalauswahl, -beurteilung und -entwicklung

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Kersting, Martin:
Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien?: Zur Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests/ Martin Kersting.- Göttingen; Bern; Toronto; Seattle: Hogrefe, Verl. für Psychologie, 1999
ISBN 3-8017-1259-1

Diese Arbeit wurde 1998 vom Fachbereich Psychologie der Justus-Liebig-Universität Giessen unter dem Titel *Diagnostik und Personalauswahl mit computergestützten Problemlöseszenarien? Eine Erörterung und ein empirischer Vergleich der Kriteriumsvalidität von Problemlöseszenarien und Intelligenztests* als psychologische Dissertation angenommen.

Druck: Dieterichsche Universitätsbuchdruckerei
W. Fr. Kaestner GmbH & Co. KG, D-37124 Rosdorf / Göttingen
Printed in Germany

Auf säurefreiem Papier gedruckt

ISBN 3-8017-1259-1

Inhaltsverzeichnis

Danksagung

Die vorliegende Arbeit basiert auf einer Dissertationsschrift, die an der Justus-Liebig Universität Giessen angenommen wurde. Zunächst möchte ich den Betreuern und Gutachtern dieser Arbeit, Herrn Prof. Dr. *E. Todt* und Herrn Prof. Dr. *U. Glowalla* Dank sagen.

Die Realisierung der Arbeit verdankt sich darüber hinaus einer Vielzahl weiterer Personen, die mir in mehrfacher Hinsicht immer wieder zur Seite standen und die Bewältigung des - neben der Berufstätigkeit realisierten Projekts - maßgeblich unterstützt haben. Nur einige kann ich hier erwähnen. Beginnen möchte ich mit der Unterstützung aus dem wissenschaftlichen Bereich. Externen, d.h. universitär oder institutsintern nicht-integrierten Doktorand(inn)en ermangelt es häufig der Gelegenheit, die eigene Arbeit mit Fachkolleg(inn)en kritisch und konstruktiv zu diskutieren. Daß ich von diesem notwendigen und bedeutsamen Feedback nicht abgeschnitten war, verdanke ich u.a. den folgenden Personen, die mir u.a. zu einzelnen Kapiteln oder der Gesamtarbeit eine wertvolle Rückmeldung gegeben haben und mir Diskussions- und Ansprechpartner(in) waren: Dipl.-Psych. *Uwe Funke*, Prof. Dr. *Adolf Otto Jäger*, Dr. *Klaus Oberauer*, Dr. *Joachim Stöber*, Dr. *Stefan Strohschneider*, PD Dr. *Heinz-Martin Süß* sowie insbesondere Dr. *André Beauducel* und Dr. *Anja Leppin*. Ohne diese Personen wären die Arbeit in dieser Form nicht möglich gewesen. Herrn *Uwe Funke* möchte ich darüber hinaus für die Bereitstellung des Programms „DISKo" sowie für zahlreiche Anregungen und Hilfestellungen herzlich danken.

Ein wesentlicher Freund und Helfer dieser Arbeit war die Polizei der Länder Nordrhein-Westfalen, Niedersachsen und Schleswig-Holstein. Auf administrativer Ebene möchte ich hier namentlich Herrn *Kripgans*, Herrn *Kunkel,* Herrn *Loh,* Herrn Dr. *Minnier* und Herrn *Zantow* danken. Vor allem verdient aber die Bereitschaft und das Engagement der 104 Polizist(inn)en Dank, die allein der Prädiktorerhebung insgesamt fast 1000 Arbeitsstunden widmeten und größtenteils zu einem späteren Zeitpunkt für eine Nachbefragung zur Verfügung standen. Bedanken möchte ich mich außerdem bei den weiteren Polizeiangehörigen, die -überwiegend als Vorgesetzte- mit der Bearbeitung des Beurteilungsbogens die Datengrundlage für die prädiktive Validierung geschaffen haben.

Das Dissertationsvorhaben wurde dankenswerterweise von der *Deutschen Gesellschaft für Personalwesen e.V. (DGP)* gefördert. Namentlich zu nennen sind hier vor allem drei Personen. Die Zustimmung zu dem Projekt sowie zahlreiche Hilfestellungen verdanke ich dem ehemaligen leitenden Psychologen der *DGP*, Herrn Dr. *Klaus Althoff*. Bei der Untersuchungsleitung sowie in EDV-Fragen wurde ich von Herrn *Thomas Königer* bestens unterstützt, dem ich ebenso Dank sagen möchte wie Herrn *Siegfried Nagel,* der die umfangreiche Datenablochung überaus sorgfältig und zuverlässig erledigt hat.

Ein letzter Dank gebührt dem Hogrefe-Verlag für die Veröffentlichung dieses Bandes.

1. Einleitung

Arbeiten zum komplexen Problemlösen folgen häufig einem Eröffnungsritual, in dem die wachsende Komplexität der Moderne und die damit einhergehenden neuen Anforderungen an ein systemisches Denken beschwörend den überschaubaren alten Zeiten gegenübergestellt wird. Spätestens bei der notwendigen Definition von „Komplexität" treten allerdings Komplikationen auf. Eine konsensfähige Definition dessen, was „Komplexität" psychologisch bedeutet, liegt bislang nicht vor, nicht einmal der Komplexitätsvergleich zweier computergestützter Problemlöseszenarien mit durchaus überschaubarer Anzahl an Programmzeilen gelingt. Diese Einsicht streift aber selten den Eingangstopos von der wachsenden Komplexität der Welt.

Unstrittig ist, daß sich mit der Publikation der frühen deutschsprachigen Arbeiten zum Verhalten bei Problemlöseaufgaben, die von den Autoren[1] als komplex und realitätsorientiert bezeichnet wurden, (Dörner, 1981; Dörner & Reither, 1978, Putz-Osterloh, 1981; Putz-Osterloh & Lüer, 1981) und spätestens mit der Veröffentlichung von „Lohhausen" (Dörner, Kreuzig, Reither & Stäudel, 1983) Problemlöseszenarien insbesondere in der deutschen Psychologie etabliert haben. Dabei steht die Evaluation gegenüber der Konzeption immer neuer Problemlöseszenarien und Szenarienvarianten zurück. Strauß und Kleinmann (1995a, S.291 f.) listeten im Anhang ihres Sammelbandes bereits 1995 siebzig *ausgewählte* Verfahren und Verfahrensvarianten auf. Auf diese Art und Weise sind zahlreiche Instrumenten-Unikate erzeugt worden, aber kaum replizierte Forschungsergebnisse. Die computergestützten Problemlöseszenarien rekurrieren auf Aufgaben, die im Alltag von Verantwortungs- und Entscheidungsträgern vorkommen können. Es geht beispielsweise um die Aufgaben eines Bürgermeisters einer Stadt oder der Regierung eines Kleinstaates, um die Aufgaben von Führungskräften verschiedener Management-Domänen, um die Tätigkeit eines Entwicklungshelfers (einen ausführlichen Einblick in die Vielfalt der Szenarien gibt Funke, 1988, 1991b).

Neben der Zunahme der Anzahl an Szenarien expandierte auch das Einsatzgebiet dieser relativ neuen Instrumente, die nur anfänglich auf eine streng kognitiv orientierte Problemlöseforschung begrenzt waren und mittlerweile in verschiedensten Forschungskontexten Anwendung finden. Beispiele reichen von der Emotionspsychologie (Stäudel, 1987) über die Neuropsychologie (Fritz & Funke, 1988) bis zur kulturvergleichenden Psychologie (Strohschneider, 1994, 1996c). Mittlerweile

[1] Hier wie im folgenden ist die weibliche Form stets mit gemeint.

wurden die computergestützten Problemlöseszenarien auch als neue Generation diagnostischer Instrumente für praktische Fragestellungen ausgerufen. Die neue Testgeneration fand ein beachtliches Interesse, welches nicht nur darauf zurückzuführen ist, daß viele Menschen das Attribut „neu" im Etikett der Instrumente als qualitativen Begriff interpretieren. Vielmehr besteht eine spezifische Unzufriedenheit vieler Anwender mit den herkömmlichen diagnostischen Instrumenten. In einer Untersuchung zum Stand der psychologischen Diagnostik äußerten knapp 59% der befragten praktisch tätigen Psychologen den Wunsch nach neuen bzw. verbesserten Testverfahren (Schorr, 1995, S.11 f.). Die computergestützten Problemlöseszenarien schienen genau diese spezifische Nachfrage zu treffen, da die Befragten der zitierten Studie sich vor allem neue Verfahren zur Intelligenzmessung und zur Erfassung der Berufseignung bei Arbeitnehmern, Auszubildenden und Führungskräften wünschten. Gerade hinsichtlich dieser Anwendungsfelder wurde für computergestützte Problemlöseszenarien ein konzeptioneller diagnostischer Fortschritt postuliert, die Schwächen der als „realitätsfremd" empfundenen Intelligenztests - die sich angeblich seit Jahrzehnten nicht weiterentwickelt haben (z.B. Sternberg & Kaufman, 1996) - sollten mit den neuen Instrumenten überwunden werden.

Es ist erstrebenswert, neuere Erkenntnisse grundwissenschaftlich orientierter Kognitionsforschung für die Testentwicklung zu nutzen (Caroll & Horn, 1981). Neuentwicklungen bedürfen aber der Evaluation; die zu ermittelnde Validität markiert die Grenze der diagnostischen Aussage. Hinsichtlich der diagnostischen Tauglichkeit von Problemlöseszenarien blieb die empirische Überprüfung aber entweder aus oder lieferte widersprüchliche Befunde. Ein empirischer Vergleich mit bereits vorhanden Verfahren, mit dem die *Nützlichkeit* (Lienert, 1967, S.19) der neuen Instrumente überhaupt erst beurteilt werden kann, wurde bislang nicht unternommen.

Die vorliegende Arbeit thematisiert den Anspruch, mit Hilfe computergestützter Problemlöseszenarien professionelle *Diagnostik* betreiben zu können und geht dabei insbesondere auf die Anwendungsmöglichkeiten und Grenzen im Rahmen der Personalauswahl und der sogenannten „Managementdiagnostik" ein. Ziel der Arbeit ist es, handlungsrelevantes Hintergrundwissen aufzubauen, welches für die Entscheidung, ob und in welcher Form Problemlöseszenarien zur Fähigkeitsdiagnostik genutzt werden können, nützlich ist. Die Gültigkeit der Ausführungen ist aber nicht auf die diagnostische Praxis begrenzt: Gerade die (differentiell-psychologische) Forschung setzt voraus, daß durch die eingesetzten Instrumente Kennwerte für inter- und intraindividuelle Merkmalsunterschiede mit zumindest befriedigender Güte gewonnen werden. Wenn mit Kastner (1978, S. 333) die Entwicklung der Intelligenztests in den letzten fünfzig Jahren die Tatsache verdeutlicht, *„daß in der psychologischen Forschung des öfteren die differentiell-psychologische Theorienbildung im Nachhinein erfolgte"*, da nämlich erst *„nach der pragmatischen Suche nach Auf-*

gaben, die zwischen guten und schlechten Schülern trennen, wie bei Binet und Simon, oder nach der Analyse systematischer Datensätze, wie bei Spearman (...) theoretische Konzeptionen, etwa das Generalfaktormodell" entwickelt wurden, so dürfte dies auch für die differentiell-psychologische Problemlöseforschung gelten. Eine empirisch fundierte theoretische Konzeption der (unidimensionalen und/oder multiplen, hierarchisch und/oder in Facetten strukturierten?) Problemlösefähigkeit und deren Abgrenzung zu etablierten Fähigkeitskonstrukten wie der Intelligenz steht nach wie vor aus. Aus der Perspektive der angewandten Diagnostik wäre es günstig gewesen, wenn die Etablierung von Konstrukten der rein pragmatischen Verwendung von Problemlöseszenarien in der Diagnostik vorausgegangen wäre, da somit ein theoretischer Hintergrund für die Interpretation der diagnostischen Informationen zur Verfügung gestanden hätte. Falls nun im „Nachhinein" - auf Grundlage der Analyse vorhandener Datensätze – eine theoretische Konzeption des Problemlösens entwickelt werden soll, so setzt dies wiederum die in der vorliegenden Arbeit thematisierte psychometrische Qualität der Problemlöseszenarien voraus. Nach Gigerenzer (1981) begründet das Mißverhältnis zwischen der Vernachlässigung des Meßproblems einerseits und der intensiven Analyse und Interpretation der möglicherweise ungültigen Daten andererseits ein ernstes Problem psychologischer Forschung (ex falso quod libet). Die Arbeit berührt daher sowohl Aspekte der praktischen Diagnostik als auch der differentiell-psychologischen Forschung.

Die Ausführungen stehen unter der Leitfrage, inwieweit computergestützte Problemlöseszenarien diagnostische Standards erfüllen. Der Theorieteil wird mit einer Definition von Problemlöseaufgaben und einer Beschreibung der Interaktionsmöglichkeiten zwischen Situations- und Personmerkmalen eröffnet. In diesem Kapitel (zwei) wird auch eine Abgrenzung von Problemlöseszenarien gegenüber den in der vorliegenden Arbeit nicht thematisierten Planspielen, Planungsaufgaben und quasiexperimentellen Simulationen vorgenommen. Das dritte Kapitel referiert kritisch die Argumente, die bislang für den diagnostischen Einsatz von computergestützten Problemlöseszenarien vorgebracht wurden. Argumentiert wurde vor allem mit der mutmaßlichen Korrespondenz zwischen den beruflichen Anforderungen einerseits und den Anforderungen bei der Bearbeitung der Szenarien andererseits sowie mit der vermeintlich hohen Akzeptanz der neuen Verfahren. Eine dritte Argumentationslinie stützt sich auf die Hoffnung, mit computergestützten Problemlöseszenarien die mit Intelligenztests geleisteten Diagnosen erweitern zu können. Mit dem Simulationsargument beschäftigt sich das vierte Kapitel, wobei sowohl der Anspruch auf Simulation bestimmter Realitätsbereiche als auch der Anspruch auf ökologische Validität im Sinne der Simulation von Anforderungen dargestellt und überprüft wird. Anschließend werden die auf der Annahme der Realitätsnähe logisch aufbauenden Experten-Novizen Vergleiche referiert. Dörner (1992) verbindet mit computerge-

stützten Problemlöseszenarien eine „Philosophie der Verwendung von Mikrowelten", die sich als Kritik der und Alternative zur Experimentalmethodik versteht. Kapitel fünf zeigt auf, daß die im Zusammenhang dieser „Philosophie" formulierten Regeln zur Verwendung von Problemlöseszenarien teilweise nicht auf die Verwendung dieser Instrumente in der praktischen Diagnostik angewendet werden können. Davon, daß sich aus der Szenarienbearbeitung eine Vielzahl verschiedener Problemlösegütemaße ableiten lassen und davon, ob und in welchem Ausmaß die Problemlösegütemaße den Anforderungen an eine Messung genügen, handelt das sechste Kapitel. Im siebten Kapitel wird die Aufgabenschwierigkeit, respektive die Steuerbarkeit der Systeme in ihren Auswirkungen auf die Diagnostik diskutiert. Kapitel acht gibt einen Überblick über die Reliabilität von Steuerungsleistungen. Das neunte Kapitel über die Validität stellt zunächst die Generalität der Problemlöseleistungen dar und setzt sich anschließend mit der Konstruktvalidität auseinander. Diesbezüglich wird insbesondere versucht, die Problemlösefähigkeit in den Kontext der Konstrukte Intelligenz und Wissen einzuordnen. Schließlich werden die bisherigen Befunde zur Kriteriumsvalidierung referiert. Im zehnten und letzten Kapitel des Theorieteils werden die „Fairneß", „Verfälschbarkeit", „Normierung" und „Ökonomie" als besondere Gesichtspunkte beim Einsatz computergestützter Problemlöseszenarien als psychodiagnostische Verfahren aufgegriffen.

Mit der im zweiten, empirischen Teil der Arbeit vorgestellten Untersuchung wird versucht, zur bislang weitgehend fehlenden empirischen Fundamentierung der Diagnostik mit computergestützten Problemlöseszenarien beizutragen. Eine Gruppe von 104 Polizisten bearbeitete einen Intelligenztest, zwei Problemlöseszenarien sowie einen allgemeinen und einen systemspezifischen Wissenstest zu den eingesetzten Szenarien. Vor der Bearbeitung der Problemlöseszenarien wurde die Erfahrung im Umgang mit Computern und die Einstellung gegenüber der Arbeit mit Computern erfaßt. Im Anschluß an die Untersuchung beurteilten die Teilnehmer die Intelligenztests einerseits und die computergestützte Problemlöseszenarien andererseits unter verschiedenen Akzeptanzgesichtspunkten. Hauptanliegen der Studie war eine vergleichende Kriteriumsvalidierung der Intelligenztests und Problemlöseszenarien. Neben der Erhebung verschiedener retrograder und konkurrenter Kriterienmaßen wurde zu Zwecken der prädiktiven Kriteriumsvalidierung im Durchschnitt ca. 1½ Jahre nach der Erhebung der Prädiktoren eine Vorgesetzten- und im Durchschnitt ca. ein Jahr nach der Prädiktorenerhebung eine Teilnehmerbefragung zu den im Berufsalltag gezeigten intellektuellen Leistungen und Problemlöseleistungen durchgeführt. Bei dieser prognostischen Kriteriumsvalidierung wurden auch Daten zur diskriminanten Validierung erhoben. Den Abschluß bildet in Kapitel 18 eine Diskussion und Bewertung der Befunde.

2. Probleme, Problemlöseaufgaben und Problemlösen: Definition und Klassifikation

Die Frage, wie Menschen Probleme lösen, gehört zu den traditionellen Forschungsschwerpunkten der (Denk-)Psychologie. Übersichten finden sich beispielsweise bei Hussy (1984) oder Klauer (1995). Für die vorliegende Fragestellung sind insbesondere solche klassifikatorischen Gesichtspunkte von Problemlöseaufgaben und Problemlöseprozessen interessant, die die Ableitung des diagnostischen Anspruchs oder diagnostisch relevante Aufgabenmerkmale betreffen.

2.1 Probleme und Problemtypen

Eine „erste Phase" der theoretischen Auseinandersetzung führte zu allgemeinen Explikationen des Begriffs *Problemlösen* und zur Klassifikationen von Problemen (z.B. Dörner, 1976; Klix, 1971; Krause, 1982a, 1982b; Newell & Simon 1972). Als Problemlösen wird der Prozeß bezeichnet, der notwendig ist, um von einer gegebenen Ausgangsbedingung zu einem Ziel zu gelangen, welches mit den gegebenen Mitteln (Operationen) nicht unmittelbar erreichbar ist. Ein Problem ist somit durch die drei Komponenten eines (1) unerwünschten Ausgangszustandes, (2) eines erwünschten Endzustandes sowie einer (3) Barriere, die die unmittelbare Transformation des Ausgangszustandes in den Endzustand verhindert, bedingt (Dörner, 1976, S. 10). Diese Beschreibungen gehen auf die klassischen gestaltpsychologischen Ansätze – etwa auf die Definition von Duncker (1935, S. 1) – zurück: *„Ein ›Problem‹ entsteht z.B. dann, wenn ein Lebewesen ein Ziel hat und nicht ›weiß‹, wie es dieses Ziel erreichen soll. Wo immer der gegebene Zustand sich nicht durch bloßes Handeln (Ausführen selbstverständlicher Operationen) in den erstrebten Zustand überführen läßt, wird das Denken auf den Plan gerufen."* Ähnlich beschrieb Wertheimer (1945, S. 224) das produktive Denken als einen Übergang, eine Verwandlung von einer Situation vor einer Problemlösung in eine *vervollständigte* Situation nach der Problemlösung, wobei eine strukturelle Unklarheit, eine Lücke ausgefüllt wird. Diese „Lücke" wird in den gegenwärtigen Ansätze häufig mit einem Terminus von Lewin (1963) als „Barriere" bezeichnet. Die „Barriere" macht das Problemlösen im Sinne eines „Umwegs" nach Köhler (1921) notwendig.

Unterschiedliche Barrieretypen - hier in Abweichung von Dörner in einem weiteren Sinne als *Problemtypen* interpretiert - gewinnt man laut Dörner (1976) durch die Differenzierung des Zielzustandes einerseits und der Bekanntheit der Mittel andererseits (siehe Tabelle 1). Besteht - wie zum Beispiel in der Regel bei Intelligenzaufgaben - Klarheit über das angestrebte Ziel und die zur Verfügung stehenden Mittel, so handelt es sich um ein klar definiertes/strukturiertes/geschlossenes (*Interpolations-*) *Problem* (grau hinterlegtes Feld in Tabelle 1). Weniger gut definierte/unstrukturierte/offene Probleme liegen hingegen vor bei (1) klaren Zielvorgaben und Ungewißheit hinsichtlich der zur Verfügung stehenden Mittel (*synthetisches* Problem), (2) bei unklaren Zielvorgaben und Kenntnis der Mittel (*dialektisches* Problem) sowie (3) bei Unklarheit über Ziele und Mittel.

Tab. 1: Problemtypen
(Tabelle nach Dörner, 1976, S. 14; Bezeichnungen modifiziert)

		Klarheit der Ziele	
		hoch	niedrig
Bekanntheitsgrad der Mittel	hoch	*Interpolationsproblem*	*dialektisches Problem*
	niedrig	*synthetisches Problem*	*dialekt. u. synth. Probl.*

Unterschieden wird also zwischen Interpolationsproblemen einerseits und dialektischen und/oder synthetischen Problemen andererseits. Die traditionelle Problemlöseforschung widmete sich überwiegend Interpolationsproblemen im weitesten Sinne wie beispielsweise den Wasserumfüll-Problemen in der Tradition von Luchins (z.B. Atwood & Polson, 1976). Die bekannteste Aufgabe der traditionellen Problemlöseforschung ist „der Turm von Hanoi" (siehe Klauer, 1993; Klix, 1971), ein Alltagsbeispiel das chinesische Puzzle „Tangram". Auch solche „move problems" oder „puzzle problems" haben, in resultatorientierter Form, Eingang in die praktische Diagnostik gefunden - siehe etwa den Formlegetest (Lienert, 1958) oder die Subskalen „Mosaik-Test" und „Figurenlegen" im Hamburg-Wechsler-Intelligenztest, (Tewes, 1991). Diese Aufgabentypen sind aber nicht Gegenstand der vorliegenden Arbeit zur Diagnostik mit computergestützten Problemlöseszenarien. Zu den Vorzügen der Interpolationsprobleme gehört die einfache formale Beschreibbarkeit des „objektiven" Problemraums (siehe unten: Abschnitt 2.3.2.3) und die Möglichkeit, die optimale Problemlösegüte zu operationalisieren (siehe Kapitel 6).

Entscheidend für den „Boom" der Problemlöseforschung in der Folge von „Lohhausen" und entscheidend für den diagnostischen Anspruch der Problemlöseszenarien war ein Schwerpunktwechsel weg von Interpolationsproblemen hin zu synthetischen und/oder dialektischen Problemen. Als Begründung für diesen Schwer-

punktwechsel wurde u.a. angeführt, daß es den Interpolationsaufgaben an „Alltagsnähe" mangele. Die klassischen (Interpolations-)Problemlöseaufgaben erschienen vielen Forschern als zu stilisiert, als zu ›akademisch‹. Insbesondere die weitgehende (Vor-)Wissensunabhängigkeit wurde als Limitation empfunden (z.B. Chi, Glaser & Rees, 1982, S.11). Demgegenüber wurde die Phase des Wissenserwerbs als charakteristisch für komplexe Probleme bezeichnet (siehe z.B. Spies & Hesse, 1987, S. 372). Interpolationsprobleme werden aus den genannten Gründen auch als *künstliche* Probleme bezeichnet. Kurzum, die Anforderungen der klassischen Problemlöseaufgaben vernachlässigten nach Ansicht von Dörner (1976, S. 141) die für das Alltagsleben relevanten Aspekte des Problemlösens. Dieser Rekurs auf das Alltagsleben begründet - wie in Kapiteln 3 und 4 noch gezeigt werden wird - wesentlich den später erhobenen diagnostischen Anspruch der computergestützten Problemlöseszenarien.

2.2 Attribute komplexer Probleme

Den künstlichen Problemen oder Interpolationsproblemen wurden die dialektischen und/oder synthetischen Probleme als *komplexe Probleme* gegenüber gestellt. Komplexe Probleme sollen Problemen des Alltags ähnlicher sein als künstliche Probleme, sie sind durch die Attribute *Komplexität*, *(Eigen-)Dynamik*, *Vernetztheit* und *Intransparenz* intentional definiert (z.B. Dörner, 1976, 1989b). Ein Problem ist *komplex* im engeren Sinne, wenn eine große Zahl von Aspekten zu beachten ist. Es ist *dynamisch,* wenn es sich aufgrund vorhergehender Problembearbeitungen oder - im Falle der *Eigendynamik* - auch ohne eine Aktion des Problemlösers verändert, und es ist *vernetzt*, wenn eine Veränderung einzelner Aspekte sich in Form von direkten und indirekten Wirkungen auf zahlreiche andere Elemente (sowie eventuell rekursiv auf sich selbst) wie in einem Netzwerk höherer Ordnung ausbreitet. Ein Problem ist *intransparent*, wenn einzelne Elemente, Relationen oder Relationstypen unbekannt sind. Komplexe Probleme können außerdem *polytelisch* sein, indem mehrere - ggf. kontradiktorische - Ziele zu verfolgen sind. Ist das Ziel bzw. sind die Ziele nur vage formuliert, handelt es sich um ein *offenes* oder *unbestimmtes* Problem. Die Attribute *polytelisch* und *offen/unbestimmt* werden nicht in allen Begriffsexplikationen aufgezählt. Die vorliegende Arbeit beschränkt sich auf die zuerst genannten vier Attribute. *Polytelie* läßt sich als *Komplexität* und *Vernetztheit* auf der Ebene der Ziele (siehe unten) definieren, *Offenheit/Unbestimmtheit* läßt sich näherungsweise als (nicht-aufhebbare) *Intransparenz* auf der Zielebene beschreiben.

Die Beschreibung *komplexer Probleme* mit Hilfe der genannten Attribute ist in mehrerer Hinsicht mißverständlich. Zunächst ist es unglücklich, daß der Begriff

Komplexität sowohl als genus proximum als auch als differentia specifica benutzt wird. Auch ein Problem mit einer vergleichsweise kleinen Anzahl an Aspekten (wie z.B. das Sechs-Variablen-Problem „Kühlhaus", siehe etwa Reichert & Dörner, 1988) wird im übergeordneten Wortsinn als *komplexes Problem* bezeichnet, da eines oder mehrere der anderen Attribute dieser Problemklasse zutreffend sind. (Das genannte „Kühlhaus"-Problem zeichnet sich durch seine Eigendynamik aus). In der vorliegenden Arbeit wird - sofern notwendig - die *Komplexität im weiteren Sinne* (genus proximum) von der *Komplexität im engeren Sinne* (als Attribut oder differentia specifica) unterschieden. Ein weiteres Kommunikationsproblem besteht auf der Ebene der Oberbegriffe: Die Labels *komplex* versus *künstlich* sind nicht unmittelbar verständlich, wenn z.B. ein recht künstlich anmutendes Problem wie das Szenario „Sinus" (Funke & Müller, 1988) als *komplexes Problem* künstlichen Problemen taxonomisch gegenüber gestellt werden muß. Schließlich treffen zumindest die Attribute *Komplexität* und *Vernetztheit* im Prinzip auch auf einige klassische - und somit künstliche - Probleme zu, es handelt sich nicht um Alleinstellungsmerkmale komplexer Probleme. Die komplexen Probleme zeichnen sich gegenüber den künstlichen Problemen bestenfalls durch *ein besonders hohes Maß* an *Komplexität* und *Vernetztheit* aus. Einige klassische Problemlöseaufgaben - wie z.B. das Schalter-Problem - sind auch intransparent.

Mißverständnisse entstehen auch dadurch, daß die Attribute auf unterschiedliche Ebenen des Problems angewandt werden können, beispielsweise auf die Ebene der einzelnen Elemente des Problemsachverhalts, auf die Ebene der Relationen zwischen den Elementen, auf die Ebene der vorhandenen Aktionsmöglichkeiten und auf die Ebene der vorgegebenen (oder eben nicht vorgegebenen) Ziele. So werden die Attribute weiter spezifiziert, man spricht von *statischer* oder *dynamischer* Komplexität, von *Kontrollkomplexität* sowie von *Variablenintransparenz* und *Strukturintransparenz* usw. (z.B. Dörner et al., 1983b; Funke, 1984, S. 163; Strohschneider & Schaub, 1995, S. 190). Grundsätzlich erscheint eine vollständige Kreuzklassifikation aller vier Komplexitätsattribute mit zumindest den vier Problemebenen *Elemente des Problemsachverhalts*, *Relationen zwischen den Elementen*, *Aktionsmöglichkeiten* und *Ziele* möglich. Tabelle 2 stellt die entsprechende Kreuzklassifikation dar. Spezifikationen, die in der Literatur häufiger getroffen wurden, wurden in die Zellen der Tabelle eingetragen. Aber auch die zur Zeit noch „leeren" Zellen der Tabelle 2 können grundsätzlich ausgefüllt werden.

Welcher der aufgespannten Ebenen definierender Charakter beikommt, ist ungeklärt. Während beispielsweise Putz-Osterloh (1985; Putz-Osterloh & Bott, 1990) die *Komplexität* (im engeren Sinne) eines computergestützten Problemlöseszenarios über die bloße Anzahl an Variablen bestimmte, steigerte Funke (1985b) die Komplexität eines Systems *bei gleichbleibender Variablenzahl* durch eine erhöhte Anzahl an Ver-

knüpfungen. Bezüglich der Verknüpfungen kann sowohl die Anzahl an Verknüpfungen als auch die Art der Verknüpfungen komplexitätsrelevant sein. Strauß (1993, 1995) wiederum sieht in der Anzahl an Variablen keine hinreichende Bedingung für Komplexität, da er zeigen konnte, daß die Steuerungsleistungen nicht von der Anzahl der Variablen abhängen, falls die Anzahl richtiger Lösungen (die man der Ebene der Aktionsmöglichkeiten zuschlagen kann) konstant gehalten wird.

Tab. 2: Problemattribute und Problemebenen

↘ Ebenen	Intransparenz	(Eigen-) Dynamik	Komplexität	Vernetztheit
Aspekte/ Variablen	Variablen- und Zustandsintran.		statische Komplexität	
Relationen	Struktur-intransparenz		dynamische Komplexität	
Aktionen	Handlungs-intransparenz		Kontroll-komplexität	
Ziele	Zieloffenheit		Polytelie	

Die Definition *komplexer Probleme* über die Attribute *Komplexität*, *Vernetztheit*, *Eigendynamik* und *Intransparenz* führt - da die Attribute selbst nicht hinreichend definiert sind - teilweise zu einem definitorischen Regreß.

Komplexes Problemlösen als Forschungsgegenstand hat gemeinsam mit dem Computer Einzug in die wissenschaftliche Psychologie gefunden. Zahlreiche Aufgaben zum komplexen Problemlösen wurden in Form von Computerprogrammen realisiert, ohne daß dieser Realisationsform definierender Charakter für die Aufgabenstellung beikommt.

Problemlösen findet statt in einem *Bereich*, welcher durch eine Menge von Zuständen/Situationen/Objekten einerseits und durch eine Menge von Operatoren andererseits gekennzeichnet ist (Dörner, 1974, S. 25). In jüngeren Arbeiten wird dieser Bereich als System beschrieben; diese Beschreibung eines Problembereichs als System ist möglich und zumeist sinnvoll, sie ist aber keine notwendige Beschreibungsform. In der vorliegenden Arbeit wird die allgemeinere Bezeichnung des dem Problem zugrundeliegenden *Sachverhaltes* gewählt.

2.3 Für wen ist was welche Art von Problem? Person- und Situationsmerkmale

Ob ein Problem vorliegt und welche Art von Problem vorliegt, ergibt sich trivialerweise aus einer *Interaktion* zwischen *Situationsmerkmalen* einerseits und *Personmerkmalen* andererseits. Die meisten Klassifikationen des Problemlösens - wie z.B. die Taxonomie Hussys (1984) - begnügen sich mit diesen beiden Klassifikationsstufen, wobei Hussy in seiner Taxonomie anstelle des Begriffs *Situationsmerkmale* den Begriff *Problemmerkmale* verwendet. (Hussy zielte im übrigen auf eine Klassifikation der Problem*schwierigkeit*, wobei seiner Ansicht nach die Problemmerkmale nicht unabhängig von den Personmerkmalen betrachtet werden sollten.) Funke (1986, 1990, 1992) führte interessanterweise in seiner - auf den Umgang mit dynamischen Systemen ausgerichteten - Klassifikation einen dritten Gesichtspunkt ein, nämlich die *Aufgabenmerkmale*. Die Aufgabenmerkmale charakterisieren die formale Struktur des Problemsachverhalts und seine semantische Einkleidung. Demgegenüber zählt zu den Situationsmerkmalen nach Funke (1990) die Aufgabenstellung (Zeitdruck, Zielvorgabe) mit der eine Person ein Problem bearbeitet sowie der in der Situation gewährte oder verwehrte Zugang zu Informationen (Transparenzaspekt). Die Trennung zwischen Situations- und Aufgabenmerkmalen entspricht laut Funke (ebd, S. 145) der in der Psychologie gängigen konzeptionellen Trennung zwischen dem Untersuchungsmaterial und der Untersuchungssituation oder der konzeptionellen Trennung zwischen den objektiven Untersuchungsmerkmalen und der *subjektiven internalen Repräsentation* beim Probanden (*Problem space* in der klassischen Problemlöseforschung, vgl. Lüer & Spada, 1990).

Diese beiden taxonomischen Vorgehensweisen - mit entweder zwei (z.B. Hussy) oder drei (z.B. Funke) Klassifikationsgesichtspunkten - stehen recht unverbunden nebeneinander. In der vorliegenden Arbeit wird - wenn auch mit anderen Bezeichnungen - dem Ansatz von Strauß (Strauß, 1993; Strauß & Kleinmann, 1995b) folgend versucht, die beiden Ansätze zu integrieren, indem eine zusätzliche Hierarchieebene eingeführt wird (siehe Abbildung 1). Auf der ersten Ebene werden wie bei Hussy (1984) nur zwei Klassifikationsgesichtspunkte berücksichtigt, nämlich *Person-* und *Situationsmerkmale (im weiteren Sinne)*. Auf der nächsten Ebene werden die *Situationsmerkmale im weiteren Sinne* dann, wie bei Funke (1990) vorgeschlagen, spezifiziert. Dabei werden die *inhaltlichen* und *formalen* Aufgabenmerkmale noch strikter getrennt als bei Funke. Auf der ersten Ebene sind also den *Personmerkmalen* die *Situationsmerkmale im weiteren Sinne* als Pendant beigeordnet. Unterhalb der Ebene der *Situationsmerkmale im weiteren Sinne* finden sich dann die *situativen* sowie die *inhaltlichen* und die *formalen Aufgabenmerkmale* (siehe Abbildung 1).

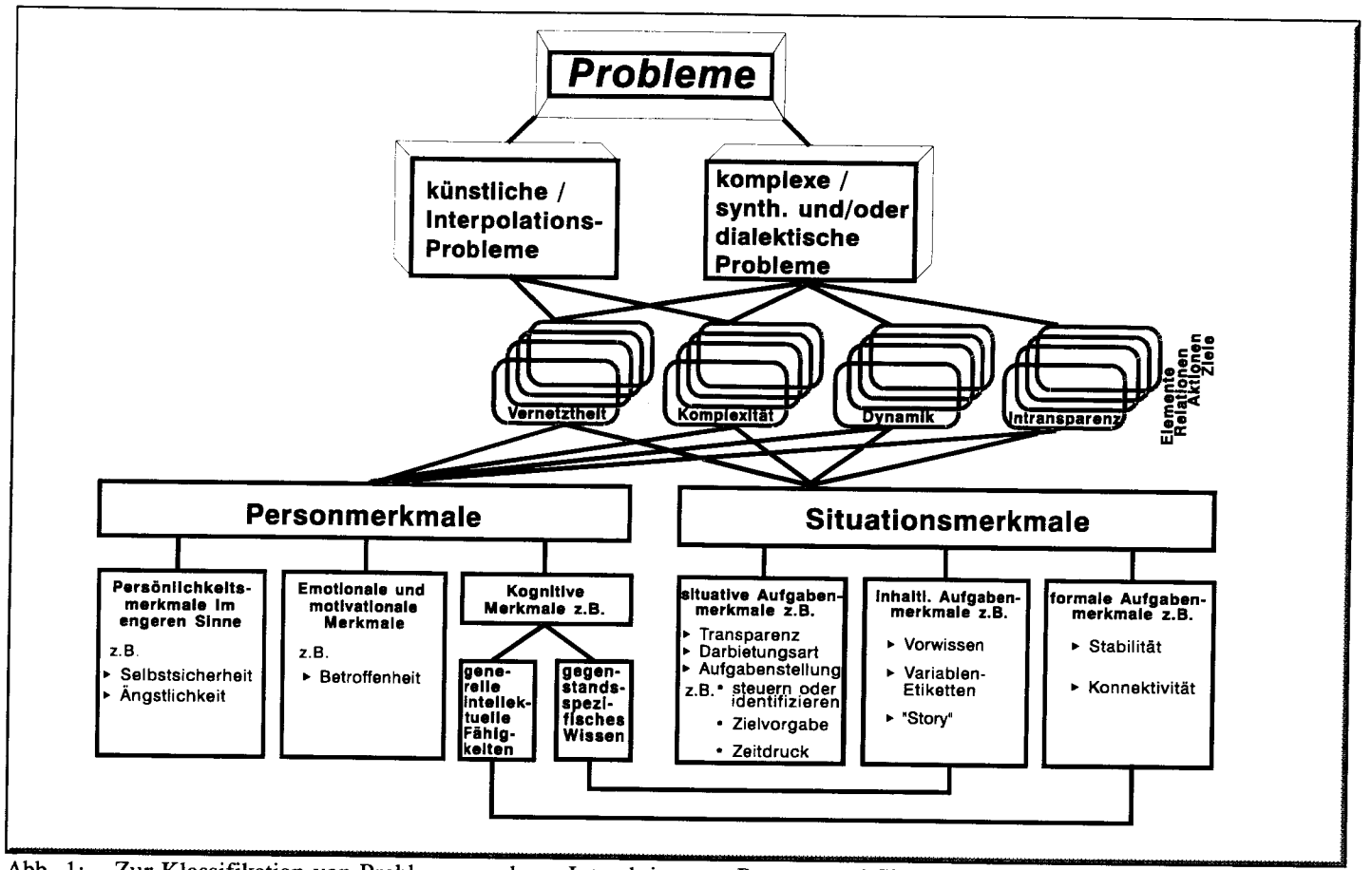


Abb. 1: Zur Klassifikation von Problemen und zur Interaktion von Person- und Situationsmerkmalen

Während die Unterscheidung zwischen Person- und Situationsmerkmalen konsensuabel scheint, wird die auf der nächsten Ebene vorgenommene Trennung zwischen situativen, inhaltlichen und formalen Aufgabenmerkmalen kontrovers diskutiert (Funke, 1991a; Strohschneider, 1991b). Kritiker der Klassifikation weisen darauf hin, daß ein durch formale Merkmale definierter Sachverhalt „als solcher" nicht unabhängig von den situativen Merkmalen existiert. *„Die Kenntnis der reinen mathematischen Struktur wäre mithin für das Verständnis des Systems sinnlos"* (Strohschneider, 1991b, S. 110). Innerhalb dieser Argumentationslinie wird betont, daß die *objektive* Seite eines Problems zur Rekonstruktion des Verhaltens der Problemlöser vernachlässigt werden kann, da nur jene formalen Aspekte Bedeutung gewinnen, welche mit der *Repräsentation* des Problemlösers übereinstimmen (Putz-Osterloh & Bott, 1990, S.283). Unbestritten ist es so, daß das Problemlösen nicht (nur) von Aufgabenmerkmalen, sondern von der Perzeption und Deutung der Aufgabenmerkmale angeleitet wird. Es gilt aber auch, daß - wie Funke (1992) experimentell gezeigt hat - Unterschiede in den Systemmerkmalen auch dann verhaltenssteuernd wirksam werden können, wenn die Versuchspersonen die Unterschiede subjektiv nicht beschreiben können. Der zweifellos wichtige Punkt der *subjektiven Sicht* wird mit der Taxonomie überhaupt nicht in Frage gestellt. Die Möglichkeit der einen Sichtweise (der subjektiven Sicht) kann nicht als Argument gegenüber der Möglichkeit einer anderen Sichtweise (der *objektiven* Sicht) angeführt werden (Funke, 1991a, S. 114 f). Die Tatsache, *„daß ein naiver Betrachter einer bunten Sommerwiese nicht die verschiedenen Arten von pflanzlichen und tierischen Organismen wahrnimmt, die ein Biologe in der Tradition von Lenné identifiziert, tut dem Wert der biologischen Taxonomie doch keinen Abbruch"* (Funke, 1991a, S. 116).

Die *subjektive* Sicht entspricht dem *mentalen Modell* nach Kluwe und Haider (1990, S. 175), während die *objektive Sicht* dem *objektiven Modell* entspricht. Neben diesem *objektiven* und *mentalen Model* unterscheiden Kluwe und Haider noch das *psychologische Modell des mentalen Modells*, welches eine Beschreibung des hypothetischen Wissens eines Individuums über ein System darstellt und das *Design- und Instruktionsmodell*, das der Konstruktion und Auslegung sowie der Ausbildung und Instruktion dient.

In der vorliegenden Arbeit werden situative Merkmale sowie inhaltliche und formale Aufgabenmerkmale taxonomisch getrennt (siehe Abbildung 1), womit der Ansicht von Funke (1991a, S. 117) gefolgt wird, daß es angemessen ist, verschiedene Perspektiven einzunehmen. Die folgenden Abschnitte erläutern kurz die einzelnen Merkmalsklassen und geben in einzelnen Fällen ausgewählte Literaturhinweise auf beispielhafte Arbeiten, die sich spezifisch mit der Bedeutung der jeweils besprochenen Merkmale für das Problemlöseverhalten auseinandergesetzt haben.

2.3.1 Personmerkmale

Dörner (1979, S. 14) gibt zur Veranschaulichung der Bedeutung von Personmerkmalen folgendes Beispiel: Angenommen jemand sieht sich mit der Aufgabe konfrontiert, Ammoniak herzustellen. Wenn jemand von Chemie nichts versteht, so stellt diese Anforderung ein synthetisches Problem dar. Für einen Chemiker handelt es sich hingegen um ein Interpolationsproblem. Auch die Intransparenz eines Problems kann in Abhängigkeit von den Personmerkmalen variieren, nämlich z.B. in Abhängigkeit vom Wissen der Person über das Problem. So ist ein Automotor für einen Laien in einem größeren Ausmaß intransparent als für einen Mechaniker.

Unter die Kategorie Personmerkmale fallen laut Funke (1990, S. 146) alle Eigenschaften, Fähigkeiten und Kenntnisse, die eine Person in die Problemsituation mitbringt oder während der Situation erwirbt. Funke nennt im einzelnen *kognitive Merkmale* (z.B. deklaratives und prozedurales Wissen, mentale Modelle, kognitive Stile und Intelligenz), *emotionale und motivationale Merkmale* sowie *Persönlichkeitsmerkmale im engeren Sinne* (z.B. Selbstsicherheit und Ängstlichkeit usw.). Süß, Kersting und Oberauer (1991, S. 337) haben die kognitiven Voraussetzungen für komplexes Problemlösen darüber hinaus weiter in den Bereich der *generellen intellektuellen Fähigkeiten* einerseits und den Bereich des *gegenstandsspezifischen Wissens* andererseits subsumiert, eine Einteilung, die hier aufgegriffen werden soll.

Zum Zusammenhang verschiedener Arten von *Personmerkmalen* und Problemlösen wurden zahlreiche empirische Studien durchgeführt, die im Kapitel über die Konstruktvalidität (Kapitel 9) thematisiert werden.

2.3.2 Situationsmerkmale im weiteren Sinne

Zu den Situationsmerkmalen (im weiteren Sinne) zählen die situativen Aufgabenmerkmale, die inhaltlichen und die formalen Aufgabenmerkmale.

2.3.2.1 Situative Aufgabenmerkmale

Situative Aufgabenmerkmale beschreiben den Kontext innerhalb dessen ein gegebenes Problem realisiert wird. Nach Funke (1990) stellen die *Transparenz* eines Problems und die der Problembearbeitung zugrundegelegte *Aufgabenstellung* situatives Aufgabenmerkmale dar. Hier soll ergänzend auch die *Darbietungsart* als situatives Aufgabenmerkmal Berücksichtigung finden. Die *Transparenz* eines Problems kann dadurch variiert werden, daß Informationen vorenthalten oder (zusätzlich) gegeben werden. Ein und dasgleiche Problem kann mit anderen *Aufgabenstellungen*

präsentiert werden. So kann die Aufgabe etwa lauten, eine bestimmten Zielvariable zu maximieren oder auf einen bestimmten Referenz-Wert zu fixieren (Sollwertsteuerung); oder die Probanden sollen den zugrundeliegenden Sachverhalt nicht steuern, sondern identifizieren usw. Wird den Probanden gar keine konkrete Aufgabe gestellt, ist die Situation wieder eine andere. Die Aufgabenstellung ist ein situatives Aufgabenmerkmal, das sich vor allem auf den *Problemtyp*, die *Polytelie* und die *(Ziel-)Offenheit* auswirkt. Problemlöseaufgaben werden oft mit Hilfe des Computers realisiert. Dabei unterscheiden sich die Untersuchungen auch darin, ob die Probanden unmittelbar oder mittelbar (über einen Versuchsleiter) mit dem Computer interagieren. Diese Variation soll als *Darbietungsart* bezeichnet werden.

Die Wirkung des Situationsmerkmals *Transparenz* demonstrierten zum Beispiel Putz-Osterloh und Lüer (1981), die das Szenario „Schneiderwerkstatt" von einer Gruppe unter Transparenzbedingungen und von einer weiteren Gruppe unter Intransparenzbedingungen steuern ließen. Die Probanden, die das System unter der Transparenzbedingung steuerten waren durchschnittlich erfolgreicher und nur unter dieser Bedingung konnte das Verhalten durch die Leistungen in einem Intelligenztest (Raven-Matrizen) vorhergesagt werden. Ein Beispiel für die gleichzeitige Variation von mehreren situativen Aufgabenmerkmalen - nämlich eine Variation sowohl hinsichtlich des Transparenzaspekts als auch hinsichtlich des Aspekts der Aufgabenstellung - gibt die Arbeit von Gardner und Berry (1995). Hier wurde einzelnen experimentellen Gruppen bei der Problemlösung mehr oder weniger Informationen und Hilfen gewährleistetet (Transparenzaspekt). Außerdem wurde die Aufgabenstellung abgewandelt, indem die Teilnehmer dazu aufgefordert wurden, sich bei ihrer Systemsteuerung in einem je Gruppe unterschiedlichen Ausmaß an die Hilfen und Vorgaben zu halten. Ein anderes Beispiel ist die Arbeit von Funke und Müller (1988), bei der die situativen Aufgabenmerkmale unter anderem dadurch variiert wurden, daß für eine Subgruppe eine zusätzliche Prognoseanforderung realisiert wurde (Apekt der Aufgabenstellung).

Putz-Osterloh und Bott (1990) ließen eine Gruppe der Probanden das „Moro"-Problem über einen Versuchsleiter steuern, während die Probanden der anderen Gruppe den Rechner selbst steuerten. Dadurch wurde die *Darbietungsart* variiert, auch wenn die Autorinnen in der speziellen Studie über die Variation der Darbietungsart auf den Transparenzaspekt abzielten. Innerhalb der Studien, in denen die Versuchspersonen unmittelbar mit dem Rechner interagieren, kommt auch der Gestaltung der Systemoberfläche eine Bedeutung zu, die ebenfalls den Situationsmerkmalen zugeschlagen werden kann.

2.3.2.2 Inhaltliche Aufgabenmerkmale

Die inhaltlichen und formalen Aufgabenmerkmale repräsentieren das Problem selbst. Zu den *inhaltlichen* Aufgabenmerkmalen eines Problems zählt nach Funke (1990) die *semantische Einbettung* (Variablen-Etiketten, Rahmengeschichte, Instruktion). Formal identische Probleme können sich hinsichtlich ihrer inhaltlichen Aspekte, z.B. hinsichtlich ihrer semantischen Einkleidung, voneinander unterscheiden. Solche formal identischen Probleme mit unterschiedlichen semantischen Einkleidungen werden *isomorphe Probleme* genannt. Beispielsweise sind der „Turm-von-Hanoi" und das „Monster-Problem" isomorph (siehe Klauer, 1993). Isomorphe Probleme sind für Untersuchungsteilnehmer im allgemeinen in Abhängigkeit von der unterschiedlichen semantischen Einbettung *leichter oder schwerer* zu lösen. Die semantische Einkleidung des formal identischen *move-problem* als „Turm-von-Hanoi" (*„Auf drei Haken verteilt liegen drei Scheiben. Es gibt Scheiben in drei Größen...."*) bedingt durchschnittlich andere Problemlösungen als die semantische Einkleidung als „Monster-Problem" (*„Drei fünfhändige Monster mit Namen Oskar, Anton und Erich halten drei Kristallkugeln. Wegen der quantenmechanischen Besonderheiten ihrer Umgebung gibt es Kugeln in genau drei Größen: klein, mittel und groß..."*): die „Monster-Version" des Problems ist deutlich schwerer zu lösen (Klauer, 1993). Bereits Kotovsky, Hayes und Simon (1985) konnten für das „künstliche" Problem des „Turm-von-Hanoi" feststellen, daß unterschiedliche Einbettungen (neben semantischen auch figürliche) des Problemraums die Schwierigkeit der Aufgabe drastisch veränderten. Dieser Effekt ist nicht auf *künstliche* Probleme beschränkt, sondern gilt auch für *komplexe* Probleme. Besonders schwer zu steuern sind vorwissensinkompatible Systeme. Funke (1992, S. 120 ff.) variierte in einem Experiment mit 80 Studenten u.a. die Vorwissensverträglichkeit eines Systems. Während das eingesetzte Szenario zum Gegenstandsbereich „Umweltverschmutzung durch Altöl" in der einen experimentellen Bedingung dem in einer Voruntersuchung ermittelten Vorwissen der Probanden entsprach, wurde in der anderen experimentellen Bedingung die Wirkungsweise einzelner Variablen vorwissensinkompatibel gestaltet. Das Szenario, welches dem Vorwissen der Probanden nicht entsprach, wurde unter allen untersuchten Leistungsaspekten schlechter bearbeitet. Bei konstanten semantischen Problemeinbettungen variiert die Problemschwierigkeit u.a. in Abhängigkeit vom Vorwissen der Versuchsperson.

Die semantische Einkleidung bestimmt *gemeinsam* mit dem Vorwissen der Versuchsperson, ob die Informationen so encodiert werden können, daß sie in bestehende Strukturen passen (Newell & Simon, 1972). Ein Problem wird, einer weit verbreiteten Ansicht zufolge (z.B. Anderson, 1981, 1983) je leichter, desto eher die Struktur der Vorwissenselemente der aktuellen Struktur des Problems entspricht.

Bedeutungsrepräsentationen eröffnen wahrscheinlich Handlungspläne, wie die Untersuchungen von Hesse zur Variation des semantischen Kontextes nahelegen. Hesse (1982, 1985) gab einer Gruppe von Probanden das „Dori" Problem unter der üblichen Einkleidung vor - d.h. die Aufgabe lautete, das Leben der „Doris", einem Nomadenstamm in der Sahelzone, zu verbessern- , während eine andere Gruppe eine *nichtsemantische* Version des Systems steuerte, bei der die einzelnen Variablen lediglich mit Buchstaben bezeichnet wurden. Die Probanden mit dem semantisch sinnvoll eingekleideten System erzielten bessere Steuerungsleistungen als diejenigen, welche dasselbe Problem in der abstrakten Buchstabenvariante bekamen. Dies galt im wesentlichen auch dann, wenn alle Probanden eine Vernetzungsgraphik des Systems bekamen.

Dem mit den Inhaltsmerkmalen verbundenen *Vorwissenseffekt* (der auch in anderen Bereichen, beispielsweise der „Sprachpsychologie" erforscht wurde, siehe etwa Beyer, Artz & Guthke, 1990; Gerrards, 1988) kommt für eine potentielle Diagnostik mit computergestützten Problemlöseszenarien eine herausragende Bedeutung bei. Je nachdem, ob das durch die Inhaltsmerkmale aktivierte Wissen für die diagnostische Fragestellung relevant oder irrelevant ist, und je nachdem, ob die Diagnostikanden hinsichtlich des Vorwissens eine homogene oder eine heterogene Gruppe darstellen, kann sich dieser Aspekt als eine Stärke oder als eine Schwäche der Diagnostik mit computergestützten Problemlöseszenarien auswirken.

2.3.2.3 Formale Aufgabenmerkmale

Die formalen Aspekte dienen zur Beschreibung der *objektiven* Seite eines Problems. Wie komplex (objektive Komplexität im Sinne Dörner et al., 1983b) der Sachverhalt ist, d.h. wieviele Aspekte bei einem Problem zu beachten sind und welche Aspekte wie miteinander verbunden sind, wieviele Lösungsalternativen es gibt (Problemumfang; siehe Hussy 1985), wie groß der Anteil richtiger Lösungen an der Gesamtzahl aller möglichen Lösungen ist (Strauß, 1993, 1995) - diese und weitere Punkte bestimmen ganz wesentlich die Schwierigkeit eines Problems, die möglichen Einflußfaktoren auf die Steuerungsleistung und somit schließlich das Problemlöseverhalten. Die Analyse und Beschreibung der formalen Aspekte kann, je nach Problem, auch zur Formulierung einer *optimalen Problemlösung* - falls es eine gibt - oder zu Näherungswerten für Optimallösungen führen. Eine optimale Problemlösung ist ein wichtiger Vergleichsmaßstab zur Bewertung einzelner Lösungsversuche (siehe Kapitel 6). Aber auch die formalen Aspekte von Problemen ohne Optimallösungen müssen hinreichend analysiert und bei der Bewertung der Problemlöseleistung berücksichtigt werden. Eine solche Analyse kann zur wichtigen Isolation *lösungsrelevanter* Aspekte führen. Vor allem besteht bei der Unkenntnis formaler Aspekte und

ihrer Bedeutung für die Problemlösung die Gefahr, daß der Aussagenzusammenhang zwischen der abhängigen und der unabhängigen Variablen und somit die *interne Validität* (Campbell, 1957) der jeweiligen Untersuchung bzw. die Gültigkeit der diagnostischen Aussage gefährdet ist. Es ist nicht auszuschließen, daß es zahlreichen Untersuchungen im Rahmen der Problemlöseforschung an *interner Validität* mangelt (siehe beispielsweise Abschnitt 7.3). Dies ist nicht zuletzt auf die Unkenntnis der formalen Aspekte des jeweils verwendeten Problems zurückzuführen. Ein Grund dafür liegt in der Komplexität (im weiteren Sinne) der Problemsituationen selbst. Die Szenarien sind nicht nur für die Untersuchungsteilnehmer, sondern auch für die Wissenschaftler selbst komplex, und offensichtlich oft auch intransparent. Die Gefahr mangelnder interner Validität nimmt mit der Komplexität (im weiteren Sinne) der Problemsituationen - und somit mit der „Alltagsorientierung" - zu. Bei der Arbeit mit komplexen Problemen ist daher eine gründliche Analyse der *Aufgabenmerkmale* des Problemsachverhalts dringend geboten. Dies kann in Form einer *Aufgabenanalyse* (z.B. Resnick, 1976; Resnick & Ford, 1981; Schott, 1984) erfolgen. Schon an dieser Stelle kann ausgeführt werden: Falls computergestützte Problemlöseszenarien überhaupt als diagnostische Meßinstrumente verwandt werden sollen, kommen nur solche Szenarien in Frage, für die gründliche Aufgabenanalysen durchgeführt wurden. Denn nur die genaue Kenntnis der Aufgabenmerkmale ermöglicht es, direkt oder über theoretische Konstrukte vermittelt, auf das Verhalten der Diagnostikanden zu schließen. Als Ergebnis der Aufgabenanalyse erhält man im allgemeinen ein (Status- oder Prozeß-) *Modell* des Sachverhalts, der dem Problem zugrundeliegt. (Oft ist dieses Modell nicht erst Ergebnis der Aufgabenanalyse, sondern bereits Ausgangspunkt der Aufgabenkonstruktion.) Es existieren verschiedene aufgabenanalytische Techniken (die zu verschiedenen Modellarten führen) und verschiedene Definitionen der zu analysierenden Kennwerte. Über die „richtige" Art der Beschreibung „wichtiger" formaler Aspekte des Zusammenwirkens besteht keine Einigkeit. Die formalen Aufgabenmerkmale können mit unterschiedlichen Zielsetzungen und mit unterschiedlichen Techniken beschrieben werden. Einige Techniken werden im folgenden kurz vorgestellt.

Zur Aufgabenanalyse der formalen Aspekte komplexer *computergestützter* Probleme wurden häufig sogenannte „sentenziöse" Darstellungen verwandt, d. h. der Algorithmus des Szenarios wird in einer der Programmsprache nahen algorithmischen Darstellung abgebildet. Eine andere, vereinfachende Herangehensweise stellt die Erstellung einer Vernetzungsgraphik dar. Graphische Darstellungen sind zwar mit dem Nachteil eines geringen Auflösungsgrades behaftet, bieten gegenüber der sentenziösen Darstellung aber den Vorzug einer Informationsbündelung (Larkin & Simon, 1987). In diesem Zusammenhang ist auf die von Hübner (1988) ermittelten Befunde zur vergleichenden Wirkung von numerischen und grafischen

Systempräsentationen hinzuweisen. Hübner konnte zeigen, daß eine analoge Darstellung zu einer besseren Systemsteuerung in kürzerer Zeit führte. Man sollte meinen, daß man sich den Hinweis auf die Notwendigkeit *inhaltlich zutreffender* Darstellungen ersparen kann. Aufgrund der schon angesprochenen Komplexität der Sachverhalte ist die Anfertigung zutreffender Darstellungen aber nicht trivial, an diesem Punkt ist schon die interne Validität mancher Untersuchung gescheitert. So finden sich in einigen Veröffentlichungen zu Studien mit dem Problemlöseszenario „Schneiderwerkstatt" beispielsweise Vernetzungsgraphiken, die in mehreren Punkten nicht dem Programm entsprechen, welches sie vorgeblich abbilden[2] (z.B. Putz-Osterloh & Lüer, 1981, S. 314). Es versteht sich von selbst, daß Untersuchungen, die eine solche fehlerhafte Darstellung zur Realisierung von experimentellen *Transparenzbedingungen* (z.B. Putz-Osterloh & Lüer, 1981) oder zur Wissensdiagnose einsetzen (z.B. Hörmann & Thomas, 1989), keine interne Validität beanspruchen können. Diese Beispiele unterstreichen ein weiteres Mal die Bedeutung sorgfältiger Aufgabenanalysen.

Neben sentenziösen und diagrammatischen Darstellungen kann schließlich zur Analyse der formalen Aufgabenmerkmale auch auf systemtheoretisch orientierte Verfahren zurückgegriffen werden, wie sie zum Beispiel von Hübner (1987, 1988, 1989a, 1989b) oder Funke (1985, 1990, 1992) vorgeschlagen wurden. Strauß und Kleinmann (1995b) geben einen Überblick über die verschiedenen mathematischen Modellierungen, die auf der Systemtheorie (Hübner) oder auf der Theorie der multivariaten autoregressiven Prozesse (Funke, 1990, 1992) bzw. - für den Fall diskreter Systemvariablen - auf der kybernetischen Theorie finiter Automaten (Buchner & Funke, 1993; Funke & Buchner, 1992) beruhen. Für die in der Eignungsdiagnostik verwendeten Szenarien liegen keine mathematisch präzisen Beschreibungen der jeweiligen formalen Systeme vor.

Zahlreiche Studien widmeten sich der Analyse des Zusammmenhangs zwischen Problemlöseverhalten und formalen Aufgabenmerkmalen. Erforscht wurden zum Beispiel die Schwierigkeiten, ein System (das Szenario „Kühlhaus") mit zeitverzögert reagierenden Variablen zu steuern (Reichert & Dörner, 1988) oder der Zusammenhang von Eigendynamik, Nebenwirkungen und Steuerbarkeit einerseits und der Steuerungsleistung sowie des Kausalwissens andererseits (z.B. Funke 1992). Gediga, Schöttke und Tücke (1983) prüften, ob dynamische Ausgangszustände gegenüber statischen Ausgangszuständen die Problemlöseleistung erschweren, indem sie

[2] Zwei Beispiele für Fehler in der von Putz-Osterloh und Lüer (1981, S. 314) verwandten Vernetzungsgraphik: die im Programm implementierte „direkte" Verbindung der Variablen „Anzahl an Maschinen" und „Produktion" fehlt in der Graphik, die in der Graphik eingetragene direkte Verbindung zwischen den Variablen „Anzahl an Maschinen" und „Anzahl an Arbeitern" ist im Programm nicht implementiert.

das System „Summaria“ unter zwei Versuchsbedingungen - einmal mit und einmal ohne Eigendynamik - vorgaben.

Eine sorgfältige Aufgabenanalyse sowie die einheitliche Bestimmung von Systemkennwerten (z.B. *Steuerbarkeit*, *Beobachtbarkeit* und *Stabilität*, siehe Hübner, 1989) oder formaler Charakteristika (siehe Funke 1990) sind nach Ansicht zahlreicher Autoren zum Standard für Untersuchungen mit computergestützten Problemlöseszenarien zu erheben (Funke, 1990, 1992; Kluwe, 1990a; Kluwe, Misiak & Haider, 1990, 1991a; Strauß, 1993). Die Vernachlässigung der Analyse formaler Systemmerkmale bedingt u.a. die folgenden gravierenden Nachteile:

- Der Verzicht auf eine Analyse und Beschreibung der formalen Aufgabenmerkmale von Problemlöseszenarien bedeutet, daß nicht nur die Problemlöser, sondern auch die verantwortlichen Wissenschaftler und Diagnostiker unter *Intransparenzbedingungen* arbeiten. Die Ergebnisse von Studien, in denen Szenarien eingesetzt wurden, über deren formale Merkmale keine ausreichenden Informationen zur Verfügung stehen, können nicht mit den Ergebnissen anderer Studien mit anderen Szenarien verglichen werden. Dies ist ein Grund dafür, daß sich trotz der jahrzehntelanger Forschungen im Bereich „komplexes Problemlösen“ kein wesentlicher kumulativer Erkenntnisgewinn eingestellt hat. Diese - bereits in frühen Phasen der Problemlöseforschung erkannte - unglückliche Situation hat sich durch die anhaltende Produktion „neuer“ Szenarien und die über die Jahre hinweg vorgenommene Modifikation bereits existierender Szenarien noch verschlechtert. Die permanente Modifikation bereits vorhandener Szenarien führt dazu, daß auch Untersuchungen mit vom Namen her gleichen Szenarien teilweise auf formal unterschiedlichen Aufgaben beruhen.

- Die mit computergestützten Problemlöseszenarien erhobenen Verhaltensdaten können - sofern für das System keine gründlichen Aufgabenanalysen vorliegen - nicht eindeutig der Person, sondern ebensogut zum Teil dem System zugeschrieben werden, ein für diagnostische Zielsetzungen unhaltbarer Zustand (siehe auch Abschnitt 6.3.2.3).

- Ohne eine gründliche Aufgabenanalyse kann der potentielle Verhaltensspielraum der Diagnostikanden nicht bestimmt werden, beobachtete Verhaltensweisen können ohne diesen Bezugsrahmen nur bedingt bewertet werden, eine in der Leistungsdiagnostik wünschenswerte Normierung beschränkt sich auf den in den einzelnen Normierungsgruppen mehr oder minder zufällig aufgetretenen Verhaltensbereich (zur Normierung siehe Abschnitt 10.3).

2.4 Probleme, Problemlöseaufgaben und Problemlösen: Eingrenzungen des Themas

Die vorliegende Arbeit setzt sich am Beispiel der Personalauswahl mit dem Anspruch auseinander, mit computergestützten Problemlöseszenarien *kognitive Fähigkeitsdiagnostik* betreiben zu können. Im nachfolgenden dritten Kapitel wird ausführlich dargestellt, daß die für computergestützte Problemlöseszenarien erhobenen diagnostischen Ansprüche insbesondere auf diese *kognitive Fähigkeitsdiagnostik* zielen. Nicht-kognitive Einflüsse auf die Bewältigung komplexer Handlungs- und Entscheidungssituationen werden durch die Perspektive der vorliegenden Arbeit nicht in Abrede gestellt. Es wird lediglich dafür plädiert, differenzierbare Personmerkmale auseinanderzuhalten. Theoretisch ist die Einbeziehung von Kontextmerkmalen in Fähigkeitstheorien mit Jäger, Süß und Beauducel (1997, S. 8) als wenig zweckmäßig zu kennzeichnen, eine separate, technisch-unabhängige Messung und Berücksichtigung der interessierenden Merkmale ist hier einer ausufernden Ausweitung der Fähigkeitskonstrukte vorzuziehen (siehe Eysenck, 1988). In der Problemlöseforschung galt die Aufmerksamkeit vorwiegend der Extraktion einer großen Anzahl an immer neuen, unterschiedlichen (kognitiven und nicht-kognitiven) Informationen aus der Szenarienbearbeitung (siehe Kapitel 6). Demgegenüber trat die theoretische Elaboration von Konstrukten in den Hintergrund. Ob es sich bei den jeweiligen Informationen nun um Indikatoren für Motivations-, Temperaments- oder Fähigkeitskonstrukte handelt oder ob all diese Konstrukte im Laufe des Problemlöseprozesses eine Rolle spielen, wurde nicht hinreichend thematisiert bzw. untersucht. Die vorliegende Arbeit will u.a. einen Baustein zur Klärung der Konstruktfrage beitragen und konzentriert sich dabei explizit auf die Ebene der Fähigkeitskonstrukte. Diese Ausrichtung der Arbeit auf die Ebene der Fähigkeitskonstrukte bedeutet nicht, daß potentielle Vorteile eines Meßinstruments, welches verschiedene Persönlichkeitsmerkmale anspricht, unberücksichtigt bleiben. Sofern es solche Vorteile gibt, finden diese bei der Betrachtung der Kriteriumsvalidität ihre Berücksichtigung. Falls nämlich nicht-kognitive Faktoren sowohl für die Szenarienbearbeitung als auch für das Kriteriumsverhalten relevant sind und mit Hilfe von computergestützten Problemlöseszenarien valide gemessen werden, müßte sich dies positiv in der Güte der so erstellten Diagnosen widerspiegeln. Schließlich messen auch die Protagonisten der Problemlöseforschung den Erfolg von Intelligenztests an deren Vorhersagekraft und lassen den Einwand, daß die vorhergesagten Leistungen auch noch durch andere Faktoren bestimmt sind, nicht gelten (Dörner, 1986, S. 291).

Unberücksichtigt bleiben hingegen andere Verwendungsmöglichkeiten der gleichen Instrumentklasse sowie gleiche Verwendungen mit anderen Instrumentklassen. So werden computergestützte Problemlöseszenarien beispielsweise zu

Trainingszwecken verwandt (siehe z.B. die Überblicksdarstellungen bei U. Funke, 1995a, 1995b) oder zur Generierung von plausiblem Reizmaterial für situative Gruppen oder Einzelübungen (u.a. im Rahmen von Assessment Centern) empfohlen (Hasselmann & Strauß, 1993a, S.43). Das computergestützte Problemlöseszenario wird in diesem Fall nur als Medium, z.B. *„um Gruppendiskussionen in einem Assessment-Center auf einer motivierenden und realistischen Basis stattfinden zu lassen"*, (Geilhardt & Mühlbradt, 1995, S. 12) verwendet und stellt somit eher eine Alternative zu Diskussions- oder Rollenspielthemen als eine Alternative zu Fähigkeitstests dar. (Zu dieser Art der Verwendung siehe z.B. Sonnenberg, 1993, S. 148f.). Diese Aspekte werden in der vorliegenden Arbeit ebensowenig aufgegriffen wie die kognitive Fähigkeitsdiagnostik mit Hilfe von möglicherweise verwandten, aber eben nicht typgleichen Aufgaben. Während die Abgrenzung computergestützter Problemlöseszenarien gegenüber Rollenspielen und (computergestützten) Fallbearbeitungen trotz partieller Überlappungen zumeist noch offensichtlich ist, soll die Abgrenzung gegenüber Planspielen und Aufgaben zur Planungsdiagnostik sowie gegenüber quasi-experimentellen Simulationsaufgaben kurz erläutert werden.

2.4.1 Computergestützte Problemlöseszenarien und Planspiele

Konfusion besteht hinsichtlich der Begriffe „Problemlöseaufgaben" und „Planspiel". Dieser Konfusion wird mit dem Sammelband von Geilhardt und Mühlbradt (1995) Vorschub geleistet, da dort computergestützte Problemlöseaufgaben und -szenarien den Planspielen untergeordnet werden. Die zugrundegelegte Definition des Planspiels ist entsprechend weit gefaßt: *„Ein Planspiel ist eine konstruierte Situation, in der sich eine /mehrere Person(en) in/an einem diskreten Modell nach vorgegebenen Regeln verhalten, wobei das gezeigte Verhalten systematisch festgehalten und nach einem explizierbaren Kalkül bewertet werden kann"* (Geilhard, 1995, S. 49). Mit dieser Konvention können nicht nur Problemlöseaufgaben, sondern auch Postkorbaufgaben, realitätsorientierte Rollenspiele und tausende von (Computer)-Spielen den Planspielen zugeordnet werden. Unter dem gemeinsamen Begriffsdach „Planspiel" werden dabei unterschiedliche Traditionen zusammengefaßt, ohne daß die Gemeinsamkeiten und Unterschiede dieser Traditionen bislang ausreichend herausgearbeitet worden wären (siehe aber U. Funke, 1991, S.110; 1992, S. I-22; Kreuzig, 1995a, S.97). Während Problemlöseaufgaben und -szenarien – wie gezeigt – der denkpsychologischen Tradition entstammen, sind etliche „Planspiele" der wirtschafts- und betriebswirtschaftlichen Tradition zuzuordnen. Es handelt sich um „Unternehmensplanspiele", die Bereiche und Vorgänge eines Industriebetriebes in einer Art dynamischer Fallstudie simulieren (Rohn, 1995b, S. 69). Erst später wurden

auch andere Domänen, wie Ökologie und Entwicklungspolitik in Planspielen berücksichtigt. Den Ursprung für Planspiele sieht Rohn (1995a) in Kampfspielen wie „Schach" oder „Chaturango", von dort führt eine weitere Traditionslinie - die im folgenden nicht weiter berücksichtigt wird - zu „Kriegsspielen" (vgl. Melter, 1995).

Ein wesentlicher Unterschied zwischen Planspielen in der Unternehmenstradition und Problemlöseaufgaben und -szenarien in der denkpsychologischen Tradition besteht darin, daß Unternehmensplanspiele im Gegensatz zu Problemlöseaufgaben überwiegend dem *Unterricht* und dem *Training* fachlicher, methodischer und sozialer (mitarbeiterbezogener) Kompetenzen dienen und häufig spezifische Kenntnisse voraussetzen. Unternehmensplanspiele zielen eher auf eine Fortbildung berufserfahrener Personen (und werden daher häufig auch als „business-**teaching** games" bezeichnet) denn auf eine *Diagnose* der Leistungen von Bewerbern. Es wurden nur wenige Versuche unternommen, den Beitrag der Leistungen in Unternehmensspielen zur Berufserfolgsprognose überhaupt zu untersuchen. Die Ergebnisse waren entweder schwer zu interpretieren (Wolfe & Roberts, 1986) und/oder sprachen eher dafür, daß die Unternehmensplanspiele sich nicht als diagnostisches Instrument eignen (Norris & Snyder, 1982). Ein weiteres Unterscheidungsmerkmal ist darin zu sehen, daß in Planspielen - im Gegensatz zu den klassischen Problemlöseaufgaben - wesentliche Sachverhaltsbereiche häufig durch die Spielpartner (oder Gegner) realisiert werden (vgl. Leutner, 1995, S. 108).

Unternehmens-Planspiele und denkpsychologische komplexe Problemlöseaufgaben verfolgen zum Teil unterschiedliche Zielsetzungen, arbeiten mit unterschiedlichen Methoden und werden entsprechend nach unterschiedlichen Kriterien erstellt und ausgewertet. Eine systematische Explikation dieser Unterschiede sollte einer Zusammenschau beider Bereiche vorgeordnet sein. Nach einer solchen systematischen Aufarbeitung wird man möglicherweise feststellen, daß einige Unternehmensplanspiele gute Problemlöseaufgaben abgeben und vice versa. Von einer bereichsübergreifenden Systematik und Klassifikation ist man zur Zeit - da ja selbst für die komplexen Problemlöseaufgaben der denkpsychologischen Tradition keine konsensfähige Systematisierung existiert - weit entfernt. Eine Ausdehnung der Interessen über den ursprünglich psychologischen Bereich hinaus erscheint unangemessen, solange noch nicht einmal das bisherige „hauseigene" Terrain vermessen ist. Mit der zufriedenstellenden Bearbeitung der fachspezifischen Thematik kann der Grundstein gelegt werden, auf dem ein *späteres* interdisziplinären Vorgehen zur Systematisierung des allgemeinen Planspielansatzes aufbauen kann.

2.4.2 *Problemlösen und Planen, computergestützte Problemlöseszenarien und Aufgaben zur Planungsdiagnostik*

Die im vorherigen Abschnitt erläuterte Konfusion hinsichtlich der Begriffe „Problemlöseaufgaben" und „Planspiel" besteht auch zwischen den Begriffen „Planen" und „Problemlösen" sowie auf der Ebene der Meßinstrumente zwischen „computergestützten Problemlöseszenarien" und Aufgaben, die zur sogenannten „Planungsdiagnostik" eingesetzt werden, wie z.B. Postkorbaufgaben (siehe z.B. Funke, 1993) oder Konstruktionsübungen (z.B. Fay & Heilmann, 1995). In Beiträgen zum Thema „Planen" wird immer wieder auch auf Aspekte und Ergebnisse der Problemlöseforschung rekurriert. (Siehe beispielsweise das Kapitel „Planen" bei Dörner (1989) sowie Beiträge zu den Sammelbänden von Funke & Fritz (1995) sowie Strohschneider & von der Weth (1993)). Dabei ist die begriffliche und empirische Trennung der Begriffe nicht immer nachvollziehbar. Die vorliegende Arbeit folgt einer Definition von Funke und Glodowski (1990, S. 32), derzufolge Problemlösen als das *Resultat* vorangegangener Planungsprozesse betrachtet wird. Funke und Fritz (1995, S. 32) führen aus, daß dementsprechend Problemlösen stets Planungsprozesse involviert, Planen aber auch ohne Problemlösen erfolgen kann. Nach Fritz und Funke (1995, S. 72f.) bieten computergestützte Problemlöseszenarien Stimulusmaterial, welches hervorragend zur Provokation von Planungsprozessen geeignet ist. Die Autoren beklagen aber, daß bei den Szenarien keine geeignete Auswertung dieser Planungsprozesse vorgesehen sei.

Die vorliegende Arbeit beschränkt sich explizit auf die Ebene des bislang theoretisch und empirisch unzureichend aufgeklärten Konzepts „Problemlösen". Das Konzept „Planen" wird nur insofern thematisiert, wie Planungsprozesse im Problemlösen zum Ausdruck kommen. Andere Aufgabentypen zur sogenannten Planungsdiagnostik (z.B. Postkörbe, Konstruktionsaufgaben) werden nicht berücksichtigt. Gleichwohl sind die in der vorliegenden Arbeit versammelten theoretischen Ausführungen und empirischen Belege auch für Leser von Interesse, die am Konzept „Planen" interessiert sind. Wenn Problemlösen stets Planungsprozesse involviert, ist es auch für das Konzept „Planen" interessant, etwas über das Problemlösen und über die Kriteriumsvalidität von Problemlöseszenarien zu erfahren. Der Ansatz, die bei Problemlöseszenarien gezeigten Leistungen näher zu analysieren, kann sogar weiterführen als das diagnostische Interesse über die schon ungeklärte Problemlösefähigkeit hinaus noch auf eine „höhere" Ebene auszurichten und somit noch mehr Unbekannte in seine Gleichung aufzunehmen. Auch die Forschung zum Planungskonzept wäre nicht schlecht beraten, wenn sie eine theoretische Einbettung des Konzepts in die Konstrukte Intelligenz und Wissen leisten würde und prüfen würde, ob die Aufgaben zur Planungsdiagnostik irgendeinen Anteil an kriteriumsrelevanter,

systematischer Varianz erschließen, der nicht bereits durch vorhandene Intelligenzaufgaben und Wissensfragen abgedeckt wird. Gerade die theoretischen Ausführungen zum Konzept Planen legen ein solches Vorgehen nahe. So kennzeichnet beispielsweise Dörner (1989, S. 243) in seinen Erläuterungen zum Begriff des Planens den Analogieschluß als das vielleicht wichtigste Verfahren der Suchraumerweiterung. Diese Ausführungen legen es nahe zu prüfen, ob die zur herkömmlichen Intelligenzdiagnostik eingesetzten klassischen Analogieaufgaben nicht auch zur Planungsdiagnostik taugen, bzw. ob die Intelligenzdiagnostik nicht bereits eine Planungsdiagnostik beinhaltet.

2.4.3 Computergestützte Problemlöseszenarien und quasi-experimentelle Simulationen

Die von der Arbeitsgruppe um Streufert vorgestellten „quasi-experimentellen Simulationen" (Streufert, Pogash & Piasecki, 1988) zur Erfassung „kognitiver Komplexität" (Streufert & Swezey, 1986) unterscheiden sich von den computergestützten Problemlöseszenarien vor allem darin, daß die Aufgabenstellung (das Problem) von den Probanden überhaupt nicht gelöst werden kann. Die zentralen Ereignisse der komplexen, dynamischen Szenarien dieser Kategorie - die in Deutschland unter der Bezeichnung „Strategische Management Simulationen" kommerziell vertrieben werden (Streufert, Breuer & Michalik, 1995) - sind nämlich vorprogrammiert und nicht durch den oder die Teilnehmer beeinflußbar. Diese Aufgaben können daher lediglich zu einer Verhaltensbeurteilung im weiteren Sinne, nicht aber zu einer Beurteilung der Steuerungsleistung genutzt werden (siehe die Ausführungen zu Problemlösegütemaßen in Kapitel 6 und Abbildung 2). Bei herkömmlichen computergestützten Problemlöseszenarien hängt die Wahrscheinlichkeit für eine bestimmte Reaktion nicht nur von der Personfähigkeit und einer fixen Aufgabencharakteristik, sondern wesentlich auch von den Reaktionen der Person auf den vorhergehenden Sachverhalt ab (siehe Abschnitt 6.4). Durch den Verzicht auf Interaktivität sollen in quasi-experimentellen Simulationen die mit den „Item-Interdependenzen" der Problemlöseszenarien einhergehenden Schwierigkeiten der Vergleichbarkeit vermieden werden. Ziel der quasi-experimentellen Simulationen ist es, das unter Streß gezeigte Entscheidungsverhalten der Diagnostikanden in einer standardisierten komplexen Situation zu untersuchen und *„eine differenzierte Beurteilung des vorhandenen Entscheidungspotentials"* (Breuer, 1992, S. 89) zu gewinnen. Es wäre aufschlußreich zu erfahren, ob und in welchem Ausmaß diese Beurteilungen des Entscheidungspotentials mit Steuerungsleistungen bei der Bearbeitung computergestützter Problemlöseszenarien zusammenhängen. Zum Erstellungszeitpunkt der vorliegenden Arbeit

publizierte Studien zu dieser Fragestellung sind dem Autor der vorliegenden Arbeit nicht bekannt. Breuer und Streufert (1995, S. 35) vertreten selbst die Auffassung, daß quasi-experimentelle Simulationen nicht in die Methoden computergestützter Problemlöseszenarien eingereiht werden sollten. Quasi-experimentelle Simulationen unterscheiden sich von computergestützten Problemlöseszenarien. Anders als bei den quasi-experimentelle Simulationen wird bei den Problemlöseszenarien aufgrund der Eingriffe der Problemlöser ein jeweils neuer Problemstatus errechnet. Die zuletzt genannten Verfahren sind Gegenstand der vorliegenden Arbeit, so daß die quasi-experimentellen Simulationen keine Berücksichtigung finden. (Dies unterscheidet die vorliegende Arbeit von Arbeiten mit vergleichbaren Fragestellungen; vgl. U. Funke 1993, 1995a, 1995b.) Die Ausgrenzung quasi-experimentellen Simulationen bedeutet keinen großen Informationsverlust, da die für diese Verfahrensgruppe vorliegenden Berichte nicht den Standard aufweisen, der als Voraussetzung einer wissenschaftlichen Evaluation gegeben sein muß (vgl. U. Funke, 1995c).

2.5 Zusammenfassung, Schlußfolgerungen und Ausblick

Ein unerwünschter Ausgangszustand, ein erwünschter Zielzustand und eine „dazwischenliegende" Barriere: diese drei Merkmale definieren ein Problem. „Komplexe Probleme" werden gegenüber „künstlichen (Interpolations)Problemen" durch die Attribute *Komplexität*, *Vernetztheit*, *(Eigen-)dynamik* und *Intransparenz* abgegrenzt, obgleich einige dieser Attribute auch auf künstliche Probleme zutreffen können. Ob überhaupt ein Problem vorliegt und welche Art von Problem vorliegt, ergibt sich aus einer *Interaktion* zwischen *Situationsmerkmalen im weiteren Sinne* einerseits und *Personmerkmalen* andererseits.

Integriert man einige der in der Literatur dargestellten Klassifikationen (insbesondere die Ausführungen von Dörner (1976, 1989; Dörner et al. 1983b), Funke (1990, 1992) und Strauß (1993) mit den hier erarbeiteten Modifikationsvorschlägen, so ergibt sich die in der Abbildung 1 dargestellte Übersicht über Probleme, Person- und Situationsmerkmale und deren wechselseitige Abhängigkeiten.

Während „künstliche" Probleme seit vielen Jahrzehnten diagnostisch genutzt werden, drängen neuerdings auch computergestützte „komplexe Problemlöseaufgaben" in die diagnostische Praxis. Betrachtet man die Ausführungen im Kapitel 2 vor dem Hintergrund der Frage nach den Möglichkeiten und Grenzen einer Diagnostik mit computergestützten Problemlöseszenarien, so kommt den folgenden beiden Punkten eine besondere Bedeutung bei:

(1) Die Ebene der Personmerkmale und die Ebene der Situationsmerkmale bedingen einander. Eine an individuellem Problemlöseverhalten - und somit an einem Personmerkmal - interessierte Individualdiagnostik muß den Einfluß der Situationsmerkmale kontrollieren. Mindestvoraussetzung für den diagnostischen Einsatz computergestützter Problemlöseszenarien ist eine gründliche Analyse und Dokumentation der formalen *Aufgabenmerkmale* des zugrundeliegenden Problemsachverhalts. Nur unter dieser Voraussetzung kann die Varianz im Problemlösungsverhalten auf interindividuelle Unterschiede zurückgeführt werden. Dieser Aspekt wird im Fortgang der Arbeit insbesondere bei der Ableitung der diagnostisch relevanten Kennwerte des Problemlöseverhaltens - den Problemlösegütemaßen - wieder aufgegriffen (Kapitel 6).

(2) Aber auch die situativen Aufgabenmerkmale und inhaltlichen Aufgabenmerkmale dürfen in ihrer Bedeutung für das Problemlöseverhalten nicht unterschätzt werden. Besondere Beachtung verdient dabei aus diagnostischer Sicht der Vorwissenseffekt. Wird durch die inhaltlichen Aufgabenmerkmale eines computergestützten Problemlöseszenarios Wissen aktiviert, welches für die diagnostische Fragestellung irrelevant und in der Gruppe der Diagnostikanden im unterschiedlichen Ausmaß verbreitet ist, so stellt der Vorwissenseffekt ein diagnostisches Problem dar. Umgekehrt kann es aber auch gerade *für* den diagnostischen Einsatz computergestützter Problemlöseszenarien sprechen, wenn die Problemlösung nachweislich die Anwendung eignungsdiagnostisch relevanter Wissensbestände oder aber eine weitgehend domänenunabhängige Fähigkeit zur Anwendung und zum Erwerb von Wissen erfordert. Der Zusammenhang von Wissen und Problemlösen soll daher weiter unten (Abschnitt 9.1.2) - auch im Kontext der Fairneßfrage (Abschnitt 10.1) - noch vertieft werden.

3. Der diagnostische Anspruch: Begründungen, Erwartungen

„Die wohl faszinierendste Innovation im Bereich der computergestützten Diagnostik hat die Gruppe um Dörner mit dem Paradigma des »komplexen Problemlösens« und der Simulation Lohhausen (...) geschaffen. Durch die wertvollen Untersuchungen der Gruppe wurde die theoretische Basis geschaffen, die es überhaupt ermöglicht, Managementfähigkeiten in komplexen Problemsituationen eignungsdiagnostisch greifbar zu machen.“ (Obermann 1992, S. 302). Läßt man die leichtfertige Benutzung des Paradigmenbegriffs[3] unkommentiert, so verdeutlicht dieses Zitat vor allem den Anspruch, computergestützte Problemlöseszenarien als Instrumente der (Management-)Diagnostik einzusetzen. Nach Kreuzig (1995a, S. 95) galt bei der Erforschung des Problemlöseverhaltens von Anfang an *„das besondere Interesse einer differenzierten Diagnostik, die sich der Computerprotokolle bediente“*, und Dörner, Schaub, Stäudel und Strohschneider (1988, S.219) erklären *„die Erarbeitung einer Diagnostik der Handlungsfähigkeit in komplexen Situationen“* zu einem Nebenziel der Forschungen der Bamberger Gruppe. Während zumindest einzelne Mitglieder der Bamberger Forschungsgruppe den diagnostischen Einsatz computergestützter Problemlöseszenarien inzwischen *„ausgesprochen kritisch“* betrachten (Strohschneider & Schaub, 1995, S. 201), haben mittlerweile zahlreiche andere Autoren - mit mehr oder minder großen Einschränkungen - den Anspruch geltend gemacht, in der ein oder anderen Form computergestützte Problemlöseszenarien als diagnostische Instrumente nutzen zu können (z.B. Badke-Schaub & Tisdale, 1995, S. 52; Birkhan & Reitzig, 1989, S.72; Funke & Rasche, 1992, S. 118; U. Funke, 1991, S. 120,

[3] Bezogen auf die Wissenschaft bedeutet der Begriff *Paradigma*, wie Kuhn (1976, S. 37) ausdrücklich betont, mehr als „anerkanntes Schulbeispiel“ oder „Schema“. Ein wissenschaftliches Paradigma ähnelt laut Kuhn eher der *Entscheidung eines Präzedenzfalls* im Rechtswesen. Es stellt einen Forschungs*konsens* dar, eine Theorie, die besser erscheint als die mit ihr im Wettstreit liegenden. Durch ein Paradigma wird eine überwiegende Mehrheit der Fachleute zu einem Wechsel in der Art der wissenschaftlichen Vorgehensweise bewegt. Als Beispiele nennt Kuhn u.a. Newtons Paradigma für die physikalische Optik und das Franklinsche Paradigma für die Elektrik. Es schmälert Dörners Leistungen nicht, wenn man den wertvollen und folgenreichen Ansatz der Problemlöseforschung im Bereich der „normalen Wissenschaft“ ansiedelt, zumal mit Kuhn (ebd, S. 30) wohl immer noch gilt, daß es offen ist, ob irgendein Teilgebiet der Sozialwissenschaft überhaupt schon Paradigmen-Status erworben hat.

1992a, S. IV-4; Hasselmann, 1991, S. 110; 1993, S. 216; 1995, S. 258; Hasselmann & Strauß, 1993a, S. 1 und S. 7, einschränkend S. 45; Kreuzig, 1995a, S. 89; 1995b, S. 387 und S. 399; Kreuzig & Schlotthauer, 1991, S. 106; Obermann, 1991, S.2, 1995, S. 402; Putz-Osterloh, 1990, S. 198f., 1991a, S. 101, 1993a, S. 289; Putz-Osterloh & Haupts, 1989, S. 47; 1990, S. 153; Putz-Osterloh & Schroiff, 1987, S. 214; Reichert & Stäudel, 1991, S. 105; Strauß & Kleinmann, 1996, S. 81 f.; 1998); entsprechende Instrumente werden bereits kommerziell vertrieben. Einen ausführlichen Überblick über die in der Eignungsdiagnostik verwendeten Szenarien - inklusive einer anschaulichen Kurzbeschreibung - gibt U. Funke (1995a). Für eignungsdiagnostische Zwecke kommen insbesondere die Szenarien „Airport", „Autohaus", „DISKo", „Feuer", „Heizölhandel", „Manage!", „Schoko-Max", „Textilfabrik" und „Utopia" zum Einsatz.

Die Empfehlung, komplexe computergestützte Problemlöseaufgaben für eignungsdiagnostische Zwecke zu verwenden, wird unterschiedlich begründet, wobei sich im wesentlichen drei Argumentationsfiguren herauskristallisieren. Für den Einsatz der neuen Verfahren vor allem in der Managementdiagnostik spricht nach Ansicht vieler Autoren insbesondere (1.) die vermutete Korrespondenz zwischen bestimmten beruflichen Anforderungen einerseits und den Anforderungen bei der Bearbeitung computergestützter Problemlöseszenarien andererseits. Damit einhergehend wird argumentiert, daß die neuen Verfahren (2.) wichtige - bei herkömmlichen Verfahren unberücksichtigt bleibende - Fähigkeiten und Merkmale der „Gesamtpersönlichkeit" erfassen würden, wobei insbesondere der Vorteil einer Prozeßdiagnostik gegenüber der üblichen ergebnisorientierten Diagnostik betont wird. Schließlich wollen die Autoren (3.) mit dem Argument überzeugen, daß die Diagnostik mit computergestützten Problemlöseszenarien bei den Diagnostikanden eine hohe Akzeptanz finden würde.

Allen drei Positionen ist zum einen gemeinsam, daß sie sich bislang auf empirisch unzureichend abgesicherten Voraussetzungen beziehen. Zum anderen zeigen die folgenden Abschnitte, daß für alle drei Positionen sehr oft der theoretische - oder auch nur rhetorische - Vergleich mit Intelligenztests gesucht wird. Der Verfahrenstechnik des „Negativdrucks" vergleichbar, gewinnen die vermeintlichen Vorzüge der computergestützten Problemlöseszenarien vor allem dadurch Kontur, daß als Umgebung Intelligenztests gewählt werden, die dann sehr nachteilig dargestellt werden, während der eigentliche Vorzug des neuen diagnostischen Instruments oft ausgespart bleibt. Diese Beobachtung ist für die vorliegende Fragestellung bedeutsam, da die Arbeit mit dem bislang fehlenden empirischen Vergleich der beiden diagnostischen Instrumente eine Konsequenz aus der theoretischen Verfahrenskonfrontation zieht.

Mit der direkten Gegenüberstellung der Vorteile der computergestützten Problemlöseszenarien mit den vermeintlichen Nachteilen der Intelligenztests und der gleichzeitigen Charakterisierung der prognostischen Validitäten der Intelligenztests als „enttäuschend", wurden indirekt sehr hohe Erwartungen hinsichtlich der prognostischen Validität der neuen Instrumente geweckt. In welchen Dimensionen sich diese Hoffnungen bewegten, zeigt sich daran, daß die herkömmlichen Validitätswerte als unzureichend angesehen wurden. So war Dörner (1986, S. 292) angesichts einer Korrelation von r=.45 zwischen Intelligenztestleistungen und Examensergebnissen überrascht, *„daß **nur** 20 % der Unterschiede im Examen von den intellektuellen Leistungen der Studenten abhängen"* (Hervorhebung hinzugefügt), und Dörner und Kreuzig (1983, S. 185) betonen, daß sogar Validitätskoeffizienten von r=.40 bis zu r=.60 *„nicht befriedigend"* sind. Da die Intelligenzdiagnostik die bis dato höchsten Validitätskoeffizienten in bezug auf den Ausbildungs- und Berufserfolg aufzuweisen hat (siehe z.B. den Überblick über Validitäten verschiedener Verfahren bei Hunter und Hunter, 1984), muß ein Forschungsvorhaben, dem diese Validitäten und mittelstarken Effekte nicht annähernd genügen, Größeres im Sinn haben.

In den folgenden Abschnitten werden die drei Argumentationslinien, die für eine Diagnostik mit Problemlöseszenarien sprechen (erstens Anforderungskorrespondenz, zweitens „ganzheitliche" Prozeßdiagnostik und drittens Akzeptanzerwägungen), ausgeführt und dabei teilweise weiter differenziert. Jeweils im Anschluß an die Erläuterungen findet sich eine kurze, keinesfalls abschließende Stellungnahme.

3.1 Anforderungsbezug

Die Attraktivität eines eignungsdiagnostischen Einsatzes computergestützter Problemlöseszenarien verdankt sich u. a. der Annahme, die Anforderungen der Problemlöseszenarien seien repräsentativ für typische Anforderungen an berufstätige Personen in bestimmten Aufgabenbereichen. Computergestützte Problemlöseszenarien gelten dieser Argumentation zufolge als eine Form der *Arbeitsprobe*, auch wenn dieser Rekurs auf einen charakteristischen Aspekt der Eignungsdiagnostik der zwanziger Jahre selten explizit erfolgt. U. Funke (1993, S. 111 f.; 1995a, S. 175f.) unterscheidet innerhalb der Gruppe derjenigen Autoren, die den diagnostischen Einsatz mit der Korrespondenz zwischen den Anforderungen computergestützter Problemlöseaufgaben und den beruflichen Anforderungen begründen, drei unterschiedliche Positionen. Während die größte Gruppe der Autoren sich hinsichtlich der Anforderungskorrespondenz auf den Augenschein verläßt, setzen einzelne Autoren auf eine „realitätsnahe" Modellierung oder auf Anforderungsanalysen. Diese drei Vari-

anten des Anforderungsbezugs werden in den nächsten Abschnitten dargestellt. Das Argument der Anforderungsnähe und der sich daraus ergebenden „face validity" baut auf dem Simulationsargument auf. Um den Simulationsgedanken ausführlicher darzustellen zu können, wird dieser Gesichtspunkt nicht hier als Unterpunkt, sondern in einem eigenen Kapitel (4) gewürdigt, wodurch sich allerdings einige Überschneidungen zwischen den folgenden Ausführungen und dem Kapitel 4 ergeben.

3.1.1 Plausibilitätsbedingte Anforderungskorrespondenz

Die Botschaft hör ich wohl, allein mir fehlt der Glaube;
JOHANN WOLFGANG VON GOETHE; Faust I, V. 765; Weimarer Ausgabe

Meistens wird die Übereinstimmung zwischen beruflichen und szenarienspezifschen Anforderungen im Sinne der face-validity (oder treffender: *„faith validity"* nach Cattell, Eber & Tatsuoka, 1970, S. 34) einfach postuliert, Merkmale von Problemlöseszenarien (Komplexität, Dynamik, Intransparenz und Vernetztheit) werden kurzerhand zu Merkmalen von Alltagsproblemen erklärt. *„Viele Alltagsprobleme sowie solche, mit denen Manager konfrontiert werden, sind jedoch typischerweise durch Intransparenz gekennzeichnet. So ist selten eindeutig nachvollziehbar, durch welche Ursachen eine aktuelle Situation entstanden ist, und somit ist nicht eindeutig prognostizierbar, was daraus künftig folgen kann. Die Probleme sind komplex und vernetzt: Entscheidungen haben neben Haupt- auch meist daran geknüpfte Nebenwirkungen, die möglicherweise erst spät sichtbar werden. Und grundsätzlich sind Probleme nicht statisch, sondern dynamisch: Die Außenwelt ›wartet‹ nicht, bis ein Manager in beliebig langer Zeit alle Argumente für oder gegen eine Entscheidung gegeneinander abgewogen hat."* (Putz-Osterloh, 1990, S. 196 f.; vergleiche auch Kreuzig, 1995a, S. 99, der die Attribute komplexen Problemlösens [siehe oben] zur Charakterisierung der Anforderungen an eine Führungskraft benutzt). Hasselmann (1991, S. 110) sieht hinsichtlich der zur Bewältigung komplexer Probleme benötigten „heuristischen Expertise" eine *„deutliche Parallele zu den beruflichen Anforderungen, wie sie sich etwa einer Führungskraft stellen"*, wobei er hinsichtlich der beruflichen Anforderungen im Managementbereich in einer anderen Publikation (1993) eine Literaturübersicht bereit stellt. Der begründete Rekurs auf entsprechende Anforderungsanalysen bleibt aber leider die Ausnahme, zumeist schöpft die Aussage der Anforderungskorrespondenz ihre beschwörende Kraft hauptsächlich aus der Wiederholung: die Gleichsetzung von Management und Problemlösen findet sich zu Hauf, ohne daß die theoretische Fundierung dieser Position dabei wesentlich elaboriert wird. So heißt es etwa bei Birkhan und Reitzig (1989, S. 58): *„Management*

bedeutet im weitesten Sinne ›Probleme lösen‹". Auch Kreuzig und Schlotthauer (1991, S.106) sind überzeugt: „*Management läßt sich im wesentlichen als komplexes Problemlösen verstehen*" und leiten aus dieser ausgemachten Ähnlichkeit dann (ebd.) explizit den diagnostischen Anspruch der computergestützten Problemlöseszenarien ab: „*Will man als Diagnostiker erfolgreiche Kandidaten für Managementtätigkeiten prognostizieren, liegt der Gedanke nahe, ein Verfahren einzusetzen, das die genannten Kriterien widerspiegelt.*" Ebenso führen Reichert und Stäudel (1991, S, 103) aus, daß eine Diagnose der persönlichen Stärken und Schwächen des einzelnen im Umgang mit komplexen Systemen ein Verfahren erfordert, „*das den Anforderungen einer komplexen Realität vergleichbar ist.*" Mit der vermeintlichen Realitätsnähe und/oder der sinnfälligen Korrespondenz zwischen Managementanforderungen und den Anforderungen bei der Steuerung der computergestützten Problemlöseszenarien begründen schließlich auch Hartung und Schneider (1995, S.220), Hasselmann und Strauß (1993a, S. 5) sowie Obermann (1991, S. 2) den diagnostischen Einsatz computergestützter Problemlöseszenarien.

Die Stellungnahme zu dieser Position kann sich auf den trivialen Hinweis beschränken, daß die bloße Plausibilität der Behauptung den Beleg nicht ersetzen kann. Wissenschaftliches Vorgehen zeichnet sich u.a. dadurch aus, daß es sich von den reinen Plausibilitätserwägungen des Alltagsverstandes emanzipiert und den prima-facie Beweis nicht als solchen gelten läßt. Aussagen über die Qualität der Relation zwischen den beruflichen Anforderungen und den Anforderungen der Steuerung computergestützter Problemlöseszenarien setzen zunächst voraus, daß beide Anforderungsbereiche für sich definiert sind. Diese Voraussetzung ist aber nur selten erfüllt (siehe unten). Selbst wenn die Managementanforderungen expliziert wären und die offensichtlich unterstellte Universalität dieser Anforderungen für verschiedene Managementdomänen gelten würde, blieben die Unklarheiten auf Seiten der Anforderungen der Problemlöseszenarien. Weiter oben (Abschnitt 2.3.2.3 „formale Aufgabenmerkmale") wurde bereits ausgeführt, daß es weitgehend unklar ist, welche Anforderungen ein konkretes computergestütztes Problemlöseszenario an die steuernde Person genau stellt; eine hinreichende Aufgabenanalyse fehlt häufig. Gerade die unscharfe Beschreibung der Problemlöseanforderungen sichert den Aussagen zur Anforderungskorrespondenz die Plausibilität. Die Aussage: „*Management bedeutet im weitesten Sinne ›Probleme lösen‹*" (Birkhan und Reitzig, 1989, S. 58) ist deshalb so bezwingend, weil der Satz so allgemeingültig ist, daß Beliebiges ergänzt werden kann. Der Satz „*Management bedeutet im weitesten Sinne ›Probleme lösen‹*" ist sicherlich ebenso sinnvoll wie andere Satzanfänge mit permutierten Tätigkeitsbezeichnungen („Wissenschaft bedeutet ...", „Psychotherapie bedeutet ..."). Der geheime Funktionsmechanismus liegt gerade in der Unbestimmtheit des Wortes „*Problemlösen*". Beschreibt man Denken als Problemlösen (so wie etwa Oerter,

1977, S. 133f.) geht die Gleichung im Sinne der cartesianischen Formel des cogito ergo sum immer auf (*„Mensch sein bedeutet im weitesten Sinne ›Probleme lösen‹"*) und der so begründete Anspruch, mit computergestützten Problemlöseszenarien Managementdiagnostik betreiben zu wollen, muß im Vergleich zu dem rhetorisch ebenso gut begründbaren universellen diagnostischen Geltungsanspruch gar bescheiden wirken.

3.1.2 Realitätsnahe Modellierung

Die Korrespondenzanahme ist an die Qualität der Simulation gebunden. Umso genauer die Abbildungsgenauigkeit der Simulation, umso eher wird man geneigt sein, eine Korrespondenz zwischen den Anforderungen der Simulation und des simulierten beruflichen Aufgabenbereichs anzunehmen. Ein hochwertiger Flugsimulator, bei dem der Simulationsteilnehmer in einem - in Abhängigkeit von den simulierten Steuerausschlägen um die Achsen beweglichen - Original-Cockpit sitzt, dürfte mit der Flugsituation einige überlappende Anforderungen aufweisen - auch hier ist jedoch keinesfalls von einer 1:1 Abbildung auszugehen (zur Frage von Simulationen und ihrer Abbildungsgenauigkeit, zur „Realitätnähe" und zur „ökologischen Validität" siehe Kapitel 4). Die hier thematisierten computergestützten Problemlöseszenarien wollen situative Simulationen sein, die eine geringere Abbildungsgenauigkeit aufweisen als physikalische und prozedurale Simulationen (zu den Begrifflichkeiten siehe z.B. Leutner, 1995, S.107). Erwähnenswert ist aber der Versuch, *„auf der Basis ökonomischen Theorie- und Praxiswissens zum Zweck diagnostischer Untersuchungen, ein computergestütztes ‚Realitätsszenario' (...) zu konstruieren."* (Hasselmann, Strauß & Hasselmann, 1993b, S. 559). Die dem Szenario „Autohaus" (siehe auch Hasselmann, 1995, S. 256 f.) zugrundeliegende Modellierung wurde aufgrund von Expertenbefragungen und Auswertungen sekundärer Informationsquellen vorgenommen; der Modellentwurf wurde von Praktikern überprüft.

Aus dem Bereich der Betriebs- und Wirtschaftswissenschaften stammen die wohl vielversprechendsten Ansätze zur realitätsnahen Modellierung betriebswirtschaftlicher Zusammenhänge. Ausgangspunkt der Modellierung ist hier das wissenschaftliche Know-How über wirtschaftliche Zusammenhänge, z.B. über Preis-Absatz Funktionen und den Produktlebenszyklus. Erste entsprechende Szenarien (z. B. das Szenario „Learn") hat Milling (z.B. 1996) entwickelt. Dabei muß beachtet werden, daß auch die perfekteste Simulation nicht die „Wirklichkeit" simuliert, sondern bestimmte Vorstellungen über die Wirklichkeit, im Beispiel der betriebswirtschaftlichen Simulation also die aktuellen und dominierenden betriebswirtschaftlichen Theorien (siehe Kapitel 4).

Die realitätsnahe Modellierung stellt einen vielversprechenden Ansatz dar, der seine Bewährungsprobe allerdings noch vor sich hat. Ein grundsätzliches Problem der „realitätsnahen Modellierung" besteht darin, daß mit zunehmender Abbildungsgenauigkeit die diagnostische Verwendungsbreite der Szenarien vermutlich abnimmt. Ein realitätsnah modelliertes Szenario eines „Autohauses" läßt sich im besten Fall nur noch für Diagnosen im beruflichen Umfeld eines Autohauses, nicht aber zur „allgemeinen Managementdiagnostik" benutzen (siehe allgemein das „bandwidth-fidelity dilemma", Cronbach & Gleser, 1965). Zur Zeit ist der Mangel an empirischen Studien allerdings derart überwältigend, daß über die Relation zwischen der Spezifität der Realitätsorientierung eines Szenarios und der Generalisierbarkeit der mit diesem Szenario erhobenen Daten keine Aussage gemacht werden kann.

3.1.3 Arbeitsanalysen

Von der naheliegenden Möglichkeit, die beruflichen Anforderungen mit Arbeitsanalysen zu bestimmen und die computergestützten Problemlöseszenarien diesen Analyseergebnissen entsprechend auszuwählen oder zu konstruieren, hat U. Funke (1992) Gebrauch gemacht. Die Ergebnisse der Anforderungsanalyse wurden zur Generierung von arbeitsähnlichen Situationen genutzt, das computergestützte Problemlöseszenario erhält dadurch im Idealfall den Status einer Arbeitsprobe (Schuler, Funke, Moser & Donat, 1995, S. 13). Das in dieser vorbildlichen Arbeit von Schuler und Mitarbeitern ausführlich dargestellte Arbeitsanalyseverfahren für den Bereich „Forschung und Entwicklung" ergab 13 Aufgabengebiete, eins davon wurde als „Problemlösen" gekennzeichnet und bildete die inhaltliche Grundlage für die *Konstruktion* des computergestützten Problemlöseszenarios.

Das Vorgehen von U. Funke zeigt, daß die Annahme einer Anforderungskorrespondenz nicht pauschal, sondern berufs- und szenarienspezifisch validiert werden kann. Es gibt keinen Grund, die „klassischen" Methoden der Arbeitsplatz- und Anforderungsanalysen als mögliche Begründung für oder gegen den diagnostischen Einsatz von computergestützten Problemlöseszenarien zu Gunsten autochthoner Reflexionen über die Anforderungskorrespondenz zu vernachlässigen.

Darüber hinaus weckt das Ergebnis der differenzierten Vorgehensweise von U. Funke (ebd.) erneut die Skepsis gegenüber generellen Aussagen der Art: *„Manager sein heißt, Problemlöser zu sein"*. Dieser Arbeitsanalyse zufolge bedeutet nämlich zumindest „Forschung und Entwicklung" nicht nur rein kognitives Problemlösen, sondern auch Führung, Präsentation, Kundenkontakt, Kooperation mit Vorgesetzten und vieles mehr. Neben der Problemsimulation wurden zur Prognose beruflicher Leistungen im Bereich Forschung und Entwicklung daher auch zahlreiche andere

Arbeitsproben und Testverfahren eingesetzt. Die Konfrontation mit mannigfaltigen Aufgaben dürfte nicht nur diesen Bereich - „Forschung und Entwicklung" - kennzeichnen. So beschreibt Sarges (1994, S. 420 f.) eine Vielzahl von Funktionen, die ein Manager inne hat und betont die kommunikativen und interpersonalen Aktivitäten - Aufgabenbereiche, die bei computergestützten Problemlöseszenarien nicht vorgesehen sind (siehe auch die Aufstellung von Schlüsselqualifikationen von Managern bei Grunwald, 1995). Der Einsatz zusätzlicher diagnostischer Verfahren wird zwar auch in den meisten Handbüchern diagnostisch orientierter computergestützter Problemlöseszenarien empfohlen, dabei bleibt aber zumeist unklar, welches Verfahren für welchen Anforderungsbereich valide diagnostische Aussagen liefern soll, und der ursprüngliche generell formulierte Anspruch der „Management-Diagnostik" mit den Problemlöseszenarien wird selten entsprechend eingeschränkt.

3.2 Erweiterung des diagnostischen Konzepts

Computergestützte Problemlöseszenarien gelten in der Sicht verschiedener Autoren als *„(...) an adequate tool for the assessment of cognitive abilities, that are not measured by classical tests of intelligence"* (Putz-Osterloh, 1993a, S.289). Problemlöseszenarien als diagnostische Instrumente liefern dieser Position zufolge gegenüber den Intelligenztests eine „Erweiterung" sowohl hinsichtlich der gestellten Anforderungen als auch hinsichtlich einer nicht länger auf eine „reine Produkterfassung" begrenzten Auswertung (z.B. Putz-Osterloh & Schroiff, 1987, S. 210). Hinsichtlich der Auswertung sollen u.a. auch *Strategien* Berücksichtigung finden. Anstelle einer Endproduktdiagnostik soll mit computergestützten Problemlöseszenarien eine Prozeßdiagnostik möglich werden (z.B. Putz-Osterloh, 1990, S.194; 1991, S. 97). Durch computergestützte Problemlöseszenarien wird nach Dörner (1986, S. 294) die *„herkömmliche Intelligenzdiagnostik durch eine Diagnostik des operativen Aspekts der Intelligenz"* ergänzt.

Als Beleg für die Position der diagnostischen Erweiterung werden keine Daten zur inkrementellen Validität geliefert, sondern es werden einerseits auf allgemeiner Ebene narrative Vergleiche zwischen den beiden Verfahren angestellt, und es werden andererseits die unzureichenden Korrelationen zwischen Intelligenztests und Problemlöseleistungen angeführt. Als prägnantes Beispiel für die letztgenannte Argumentation kann die englische Zusammenfassung des Dörnerschen Aufsatzes zur „Diagnostik der operativen Intelligenz" angeführt werden. Dort (1986, S. 290) heißt es: *„Considering the reasons for the low correlations between classical measures of intelligence and problem solving it is assumed that intellligence tests are deficient*

with regard to ›operative intelligence‹. " Da die Gültigkeit der mit den Problemlöseszenarien ermittelten Daten nicht in Frage gestellt wird, indiziert der ausbleibende Zusammenhang der beiden Messungen „automatisch" ein deviantes Leistungsverhalten der Intelligenztests. Der Grund dafür, daß Intelligenztestleistungen und Problemlöseleistungen nur so gering miteinander korrelieren, obwohl *„Planungsaufgaben Anforderungen an die Intelligenz der Versuchspersonen stellen"* (1986, S. 297) ist nach Dörners Ansicht die Tatsache, daß die operative Intelligenz in Intelligenztests ungenügend berücksichtigt wird (1986, S. 298, siehe auch 1984, S. 19).

Was Intelligenztests fehlt, beschreiben z.B. Dörner und Kreuzig (1983, S. 190). Demzufolge stellen Intelligenztests u.a. keine Anforderungen an die Zielelaboration und -konkretisierung sowie Zielbalancierung und -hierarchisierung, außerdem fehlen Anforderungen an den Erwerb und die Anwendung von Kenntnissen und Anforderungen an die (Selbst-) Organisation einzelner Denkprozesse. *„Die Nichtberücksichtigung solcher Faktoren könnte verantwortlich sein für die beschränkte Möglichkeit, aufgrund der Ergebnisse von Intelligenztests etwas über die tatsächliche Leistungsfähigkeit von Personen in Problemsituationen auszusagen"* (Dörner, 1984, S. 10). Intelligenztests lassen demzufolge die höheren Organisationsformen des Denkens, die Regulationsebene oder eben die „operative Intelligenz" außer acht (Dörner, 1984, S. 18).

Putz-Osterloh (1993a, S. 299) zieht in ihrem Aufsatz zu *„Complex Problem solving as a diagnostic tool"* das folgende Resümee: *„There are individual differences in complex abilities"* (Erläuterung: gemeint sind die mit Problemlöseszenarien erfassten individuellen Differenzen) *„that are not testable by usual tests of intelligence. These differences are of interest for personnel selection if one wants to test, for example, organizing and decision-making ability."* Neben weiteren Ausführungen über die - im Vergleich zu *Assessment Center Aufgaben* (!) - angeblich bessere Standardisierung und Replizierbarkeit der mit Problemlöseszenarios erhobenen Kennwerte sowie über Generalitäts- und Validitätshinweise für Problemlöseleistungen wird auch in dieser Arbeit der diagnostische Vorteil computergestützter Problemlöseszenarien daraus abgeleitet, daß Intelligenztestleistungen und Problemlöseleistungen in einigen Untersuchungen nicht miteinander kovariierten.

Zusammenfassend kann festgehalten werden, daß mit den stimmig aufgezeigten Unterschieden zwischen den Anforderungen von Intelligenztestaufgaben und computergestützten Problemlöseszenarien sowie mit den unterschiedlichen Auswertungsmöglichkeiten einige *Voraussetzungen* für eine Erweiterung der mit Intelligenztests geleisteten Diagnosen bestehen. Unklar ist, ob damit Fähigkeiten erfaßt werden sollen, die dem Konstrukt Intelligenz zuzuordnen sind oder ob es um die Diagnostik eines neuen Konstrukts „Problemlösefähigkeit" oder „operative Intelligenz" geht (siehe dazu den Abschnitt zur Konstruktvalidierung Abschnitt 9.1.2). Dieses Kon-

strukt „Problemlösen" wäre gegebenenfalls weiter zu differenzieren. (Man denke an die Debatte, die zu hierarchischen Intelligenzstrukturmodellen führte.) Unabhängig von diesen theoretischen Voraussetzungen gilt, daß eine überzeugende Validierung computergestützter Problemlöseszenarien - und insbesondere eine Validierung im direkten Vergleich zu den Intelligenztests - bislang aussteht. Die bisherigen Validierungsbemühungen werden im Kapitel 9 referiert, diesbezüglich sind auch die in Abschnitt 4.3 referierten Experten-Novizen Vergleiche von Interesse. Selbst wenn Intelligenztests und Problemlöseszenarien unkorreliert wären (das dem nicht durchgängig so ist zeigt Abschnitt 9.1.2.2) wäre dies allein kein Argument für den diagnostischen Einsatz von computergestützten Problemlöseszenarien.

3.3 Akzeptanzvermutungen

Neben inhaltlichen Kriterien wie dem vermeintlich hohen Anforderungsbezug und der Erweiterung des diagnostischen Konzeptes wird auch die mutmaßlich hohe Akzeptanz der als eignungsdiagnostische Verfahren eingesetzten computergestützten Problemlöseszenarien angeführt. Auch hier wird der Vorteil der neuen Verfahren häufig im direkten Kontrast zu den Intelligenztests herausgearbeitet: *„Wegen des gemeinsam abgedeckten Konstruktbereichs ist bzgl. der Akzeptanz insbesondere der Vergleich zu Skalen aus Intelligenz- und Persönlichkeitstests von Bedeutung. In diesem Vergleich ergeben sich bezogen auf die Akzeptanz der Verfahren in der Praxis deutliche Vorteile beim Einsatz computersimulierter, komplexer Problemstellungen"*, heißt es etwa bei Hasselmann (1995, S.255). Die mangelnde Akzeptanz wird häufig direkt auf die (in Abschnitt 3.2) beschriebenen Defizite der Intelligenztestanforderungen zurückgeführt. So schildert etwa Putz-Osterloh (1991, S, 97 f.) die Beschränkungen von Intelligenztestanforderungen, um dann daraus die Akzeptanz der Intelligenztests abzuleiten: *„Eigenständige Lösungskonstruktionen, rückmeldungsabhängige Strategieanpassungen oder längerdauernde Verarbeitungsprozesse sind z.B. nicht erfaßbar. Diese Einschränkungen sind ein plausibler Grund dafür, daß Tests von erwachsenen Bewerbern in der Regel nicht akzeptiert werden."*

Letztlich können zwei Aussagen, respektive Behauptungen, unterschieden werden. Einerseits wird behauptet, Intelligenztests riefen heftige Akzeptanzprobleme bei dem von der Management-Diagnostik betroffenen Personenkreis hervor (z.B. Hasselmann, 1993, S. 24), und andererseits wird behauptet, computergestützte Problemlöseszenarien erfreuten sich als Forschungsinstrumente (z.B. Strohschneider & Schaub, 1995, S. 189) und als Instrumente der diagnostischen Praxis einer hohen Akzeptanz (z.B. Dörner, 1992, S.84; Funke, 1998, S. 93; Funke & Geilhardt, 1996, S. 206; Funke & Rasche, 1992, S. 118; U. Funke, 1992a, S. I-3; Guthke

1996, S. 84; Hartung & Schneider, 1995, S. 221/233; Hasselmann & Strauß, 1993a, S. 9; Kreuzig 1995b, S. 397; Kreuzig & Schlotthauer, 1991, S.106; Obermann, 1995, S. 405; Schreiber, 1995, S. 279; Strauß & Kleinmann, 1997, S. 461; 1998). Auch diese beiden Facetten der dritten Argumentationslinie - der höheren sozialen Akzeptanz von computergestützten Problemlöseszenarien - sind bislang empirisch kaum untermauert.

Für die erste Behauptung, hinsichtlich des diagnostischen Einsatzes von Intelligenztests existierten Akzeptanzprobleme, könnten die Autoren auf allgemeine Daten zur sozialen Akzeptanz von Leistungstests - allein oder im Vergleich mit anderen Verfahren wie z.B. Interviews - zurückgreifen (z.B. Köchling & Körner, 1996; Fruhner, Schuler, Funke & Moser, 1991; Fruhner & Schuler, 1988; Schuler, Frier & Kaufmann, 1991; Trost, 1993). Weder diese noch andere Daten werden aber als Beleg für die Behauptung herangezogen. Dies wäre auch nicht ohne weiteres möglich, da die Datenlage kein so schlechtes Bild der sozialen Akzeptanz von Intelligenztests in der Diagnostik zeichnet. So kommen Schuler et al. (1991, S. 176) zwar zu dem Schluß, daß z.B. Vorstellungsgespräche positiver beurteilt werden als psychologische Tests (wobei in der zugrundeliegenden Befragung problematischerweise unter dem Oberbegriff Test sowohl Persönlichkeitsfragebogen als auch Intelligenztests zusammengefaßt wurden), halten aber fest, daß auch die Tests noch „neutral" beurteilt wurden. Die Behauptung, daß der Einsatz von Intelligenztests heftige Akzeptanzprobleme hervorriefe, läßt sich also bestenfalls für die spezielle Gruppe von Führungskräften - die in Akzeptanzbefragungen seltener berücksichtigt wurden - aufstellen. Selbst wenn man die Behauptung von der geringen Akzeptanz von Intelligenztests aber auf spezifische Personenkreise einschränken würde, so ersetzt diese Einschränkung dennoch keinesfalls die empirische Prüfung. Dabei sind erneut insbesondere vergleichende Studien gefragt - denn welche Auswahlverfahren finden bei Führungskräften mutmaßlich überhaupt eine hohe Akzeptanz?

Auch hinsichtlich der zweiten Behauptung muß man sich fragen, woher der Mut zum unablässig wiederholten Mantra *„Problemlöseszenarien finden eine hohe Akzeptanz"* kommt, da sich - mit wenigen Ausnahmen - die meisten Autoren offensichtlich auf ihre subjektiven Erfahrungen als Versuchsleiter und vielleicht auf unsystematische Befragungen beschränken. Zumindest wird nichts davon berichtet, ob und mit welchem Ergebnis systematische Befragungen bei den Betroffenen durchgeführt wurden. Die Ergebnisse der wenigen empirischen Studien zur Akzeptanz von Problemlöseszenarien sind widersprüchlich oder aus methodischen Gründen in ihrer Aussagekraft eingeschränkt. Während sich in der Studie von U. Funke (1992a) mit 19 Personen in einer Reihe von Akzeptanzfragen für zwei dieser Fragen ein Vorteil der computergestützten Problemlöseszenarien gegenüber Intelligenztests abzeichnete, werteten ca. 40 Anwender und Experten aus dem Personalmanagement bzw. Exper-

ten von Beratungsinstituten die unterschiedliche Akzeptanz von computergestützten Problemlöseszenarien eher als *kritischen* Punkt (Funke & Geilhardt, 1996, S. 208). Indirekt könnte man auch eine Studie von Putz-Osterloh und Haupts (1990, S. 140ff.) im Zusammenhang mit Akzeptanzgesichtspunkten betrachten. In dieser Untersuchung wurde 30 Stabsoffizieren nach der Bearbeitung des Szenarios „Feuer" u.a. die Frage vorgelegt *„Muß man bei der Feuerbekämpfung Verhaltensweisen zeigen, die auch im Beruf wichtig sind?"* Diese Frage kann man im Zusammenhang mit Akzeptanzgesichtspunkten auswerten, da ein hoher perzipierter Anforderungsbezug häufig mit einer hohen Akzeptanz einhergeht. Die gestellte Frage konnte man offensichtlich nur mit „ja" (26) oder „nein" (4) beantworten, bevor dann einzelne Fähigkeiten genannt werden konnten, die sowohl bei der Szenarienbearbeitung als auch im Berufsalltag gefordert sind.

In der im empirischen Teil der vorliegenden Arbeit dargestellten Studie wurden die Teilnehmer gebeten, die Intelligenztests einerseits und die computergestützten Problemlöseszenarien andererseits mit Hilfe eines neu konstruierten Fragebogens unter verschiedenen Akzeptanzgesichtspunkten zu beurteilen. Da die Befragung und ihre Ergebnisse bereits bei Kersting (1998) ausführlich dokomentiert sind, bleibt dieser Teil der Untersuchung im Empirieteil der vorliegenden Arbeit ausgespart. Zusammengefaßt dargestellt ergab sich, daß jedes der beiden Verfahren spezifische Akzeptanzvor- und nachteile aufzuweisen hatte. Im Gegensatz zum anderslautenden Rumor einer pauschalen Ablehnung von Intelligenztests bei gleichzeitiger Bevorzugung der computergestützten Problemlöseszenarien zeigte die Untersuchung, daß die 103[4] Teilnehmer diese Verfahren als Instrumente der Personalauswahl differenziert beurteilten. Die Bearbeitung von Problemlöseszenarien wurde vor allem allgemein positiv erlebt. Die häufig postulierte hohe Akzeptanz von Problemlöseszenarien läßt sich im Licht dieser umfassenden empirischen Studie dahingehend dechiffrieren, daß diese Aufgaben den Teilnehmern *„mehr Spaß"* machten als Intelligenztestaufgaben. Intelligenztests übertrafen die Problemlöseszenarien hingegen bezüglich der Kontrollierbarkeit im Sinne einer höheren wahrgenommenen Qualität der Messung. Den Anforderungen der Problemlöseszenarien wurde im Vergleich mit Intelligenztests zunächst eine größere Realitätsnähe ('face validity') zugesprochen. Dieser Akzeptanzvorteil ließ sich allerdings bei einer zeitlich versetzten Wiederholung der Akzeptanzbefragung im Anschluß an eine - verfahrensspezifisch unterschiedliche - Rückmeldung über die mit den Verfahren getroffenen Diagnosen nicht mehr aufrecht erhalten. Die Ergebnisse der empirischen Akzeptanzbefragung deuten

[4] Die Abweichung zur im Empirie-Teil genannten Gruppengröße von 104 Personen erklärt sich dadurch, daß für einen Teilnehmer keine Ergebnisse zur Akzeptanzbefragung vorlagen.

in die Richtung, daß es - wie Dörner (1992, S. 57) vermutet - *„sicherlich ganz »lustig« für die Versuchspersonen* [ist], *in einer simulierten Textilfirma (...) einmal den »Boss« zu spielen"*, daß der Spaß allerdings für die Testanden aufhört, sobald sie die Qualität der Messung einschätzen und sich vorstellen, daß aufgrund dieser „Ergebnisse" ernsthafte Personalentscheidungen getroffen werden sollen.

Neben der unzureichenden empirischen Absicherung muß auch der konzeptionelle Hintergrund des Akzeptanzargumentes hinterfragt werden. Ist es theoretisch überhaupt sinnvoll, für ein bei der Personalauswahl eingesetztes Instrument eine grundsätzlich hohe oder grundsätzlich niedrige Akzeptanz zu postulieren? Um zur Beantwortung dieser Frage beizutragen, sollen zunächst allgemeine und dann differentialpsychologische Gesichtspunkte der Akzeptanz von Auswahlverfahren skizziert werden. Als die wichtigsten Parameter für die soziale Akzeptanz von Auswahlverfahren benennen Schuler und Stehle (1983, 1985) sowie Schuler (1990, 1993) die Information, Partizipation, Transparenz und Kommunikation. Auch der regelbasierte Ansatz zur sozialen Akzeptanz von Auswahlverfahren, das *„Model of Applicants' Reaction to Employment Selection Systems"* von Gilliland (1993), sieht *mehrere* Einflußfaktoren auf das Akzeptanzurteil der Verfahrensteilnehmer vor. Der Facettenreichtum der Akzeptanzfrage verdeutlicht zweierlei: zum einen ist die Akzeptanz von Auswahlverfahren nicht unidimensional, sondern mehrdimensional zu konzeptualisieren. Ein Verfahren hat entsprechend grundsätzlich nicht nur ***einen*** Akzeptanzwert - wie in den oben zitierten Positionen offensichtlich unterstellt- , sondern - wie in der Studie von Kersting (1998) bestätigt - mehrere Akzeptanzwerte. Zum anderen dürfte der Facettenreichtum des Akzeptanzurteils verdeutlichen, daß durch die Wahl der Verfahren allein nur einige Komponenten der Akzeptanz betroffen sind, bzw. daß andersherum die Gestaltung der übrigen Komponenten die wahrgenommene Akzeptanz unabhängig von der Verfahrensfrage wesentlich determinieren kann. Die Information, die Partizipation und das Feedback (Kommunikation) können zum großen Teil durch die *Art der Durchführung* sowohl bei der Verwendung von Problemlöseszenarien als auch bei der Verwendung von Intelligenztests entscheidend positiv oder negativ gestaltet werden[5]. Schon aus diesen konzeptionellen Gründen müssen allgemeine Aussagen über die Akzeptanz einzelner Verfahren in ihrer Geltung beschränkt bleiben. Darüber hinaus ist anzunehmen, daß nicht nur Situations- und Verfahrensmerkmale, sondern auch Personfaktoren, insbesondere die *Fähigkeit* und die *Motivation* einer Person, das Akzeptanzurteil beein-

[5] Das Medium „Computer" kann die Gestaltung akzeptabler Situationen in bestimmter Hinsicht begünstigen, etwa durch die Möglichkeit eines unmittelbaren Feedbacks. Dies spricht aber nicht für die Akzeptanz bestimmter Verfahren, sondern bestenfalls für die Akzeptanz bestimmter Darbietungsformen. Auch Intelligenztests können computergestützt dargeboten werden.

flussen. Bei Personalauswahlverfahren gibt es - unabhängig von der äußeren Gestaltung - in der Regel stets auch für zumindest einige Individuen negative Erlebnisse des wahrgenommenen Leistungsversagens oder explizit negative Entscheidungen. Vor dem Hintergrund der Theorie der kognitiven Dissonanz (siehe z.B. Festinger, 1957) ist zu erwarten, daß diese negative Entscheidung beim Individuum eine relevante Dissonanz zwischen den Kognitionen *„die Bewerbung ist für mich sehr wichtig"*, und *„ich habe den Eindruck (oder die Gewissheit), den Anforderungen des Auswahlverfahrens nicht gewachsen zu sein"* erzeugt, auf die das Individuum z.B. mit einer selbstwertdienlichen Abwertung des Verfahrens reagieren kann. Es erscheint von daher notwendig, bei Aussagen zur Akzeptanz von Verfahren stets auch Personvariablen zu berücksichtigen. So zeigte sich in der Studie von Kersting (1998), daß sowohl die Akzeptanzbeurteilungen innerhalb einer Verfahrensgruppe als auch die Präferenzurteile zwischen Problemlöseszenarien einerseits und Intelligenztests andererseits in Abhängigkeit von verschiedenen Personfaktoren wie Leistung, Alter, Computererfahrung und -einstellung variierten.

Die Tatsache, daß es bei Auswahlverfahren immer auch „abgelehnte" Personen geben wird, die ein „Recht" auf selbstwertschützende Mißerfolgsattributionen haben, begrenzt von vornherein die maximale Höhe der Akzeptanz eignungsdiagnostischer Verfahren. Weitere Einschränkungen der maximal erreichbaren Akzeptanzhöhe ergeben sich aus den z.T. unrealistischen Erwartungen, mit denen Bewerber in ein Auswahlverfahren gehen (siehe Rynes, 1993, S. 34f.). Schließlich bleibt es auch unklar, welcher Stellenwert der Akzeptanz eines Verfahrens überhaupt zukommen kann. Wie läßt sich das Akzeptanzkriterium mit anderen Gütekriterien, wie z.B. Objektivität, Reliabilität und Validität kombinieren und gewichten? Pawlik (1997, S. 183) betont zu Recht, daß die Augenscheinvalidität (die hier als eine wichtige Determinante der Akzeptanz angesehen wird) z.B. unter dem Gesichtspunkt der Testmotivation bedacht werden kann, daß sie aber kein Ersatz der empirischen Kriteriumsvalidität ist.

Insgesamt kann die im wesentlichen lediglich postulierte Akzeptanz der computergestützten Problemlöseszenarien in der Eignungsdiagnostik ohne weitere Klärungen nicht als Argument für deren Ernstfall-Einsatz gelten.

3.4 Zusammenfassung, Schlußfolgerungen und Ausblick

Für den diagnostischen Einsatz computergestützter Problemlöseszenarien spricht nach Ansicht vieler Autoren vor allem (1.) die vermutete Korrespondenz zwischen den beruflichen Anforderungen einerseits und den Anforderungen bei der Bearbeitung der Szenarien andererseits sowie (2.) die vermeintlich hohe Akzeptanz, die diese Verfahren - im Gegensatz zu den Intelligenztests - bei den Verfahrensteilnehmern angeblich finden. Beide Argumente stützen sich aber mit wenigen Ausnahmen lediglich auf Plausibilitätsüberlegungen und bleiben eine theoretische Fundierung und empirische Prüfung schuldig. Besonders interessant ist eine dritte Argumentationslinie, derzufolge (3.) computergestützte Problemlöseszenarien eine Erweiterung der mit Intelligenztests geleisteten Diagnosen ermöglichen, da bei diesen Verfahren Anforderungen gestellt werden, die bei Intelligenztests nicht zum Tragen kommen (z.B. Zielelaboration, Wissensanwendung und -erwerb, Berücksichtigung der „Gesamtpersönlichkeit") und da bei den computergestützten Problemlöseszenarien neue Auswertungsmöglichkeiten im Sinne einer Prozeßdiagnostik bestehen. Tatsächlich erfüllen computergestützte Problemlöseszenarien einige *Voraussetzungen* für eine Erweiterung der mit Intelligenztests geleisteten Diagnosen. Bislang gilt aber, daß das Anforderungsprofil von Problemlöseszenarien nicht hinreichend untersucht ist. Weder ist geklärt, ob mit verschiedenen Szenarien das gleiche gemessen wird noch was überhaupt erfaßt werden soll (siehe Abschnitt 9.1 zur Konstruktvalidität). Auch die Schwierigkeiten einer möglichen Prozeßdiagnostik sind bislang ungelöst (siehe Kapitel 6).

Im empirischen Teil dieser Arbeit werden Kennwerte für die Anforderungskorrespondenz der computergestützten Problemlöseszenarien erhoben und den entsprechenden Kennwerten für Intelligenztests gegenübergestellt. Darüber hinaus wird geprüft, ob sich die erweiterte Diagnostik in Form von - gegenüber Intelligenztests inkrementellen - Kriteriumsvaliditäten auszahlt. Das Argument der Anforderungskorrespondenz wird im folgenden Kapitel mit einer Diskussion des Simulationsgedankens noch einmal vertiefend aufgegriffen.

4. Zur Realitätsnähe und Ökologischen Validität: Das Simulationsargument

„Lohhausen, Tanaland, Tschernobyl...“ (Dörner, 1995, S. 58) - Der Charme, der den diagnostischen Einsatz von Problemlöseszenarien zu eigen ist, entspringt ganz wesentlich der vermeintlichen „Realitätsnähe“ und „ökologischen Validität“, die aus dem Simulationsgedanken destilliert wird. Die Behauptung, daß computergestützte Problemlöseszenarien „realitätsnah“ und „ökologisch valide“ sind, gehört zu den zentralen Legitimationsgrundlagen der Forschung zum komplexen Problemlösen. Die für computergestützte Problemlöseszenarien erhobenen Ansprüche auf einen hohen Anforderungsbezug, auf die Notwendigkeit einer Erweiterung der Intelligenzdiagnostik und auf eine hohe Akzeptanz (siehe Kapitel 3) - und somit letztendlich der gesamte diagnostische Anspruch - werden mehr oder minder direkt aus dem Anspruch auf Realitätsnähe und ökologische Validität abgeleitet. In nahezu allen Belangen in der Auseinandersetzung um die Forschung und Diagnostik mit computergestützten Problemlöseszenarien wird auf dieses Merkmal der „Realitätsnähe“ rekurriert. Einwände gegen bestimmte Charakteristika computergestützter Problemlöseszenarien werden z.B. häufig mit der Begründung zurückgewiesen, dies und jenes sei vielleicht aus meßtheoretischen Gründen ungünstig, die Szenarien könnten darauf aber keine Rücksicht nehmen, da sie schließlich der Realitätssimulation verpflichtet seien. So parieren Dörner und Reither (1978, S. 534) beispielsweise die Kritik, daß die Steuerung des Szenarios zu schwer sei (siehe auch Abschnitt 7.2) damit, daß es nicht darum ginge *„die Vpn einer 'fairen'oder 'unfairen' Situation auszusetzen, sondern einer Situation, die der Realität in Hinblick auf die (...) genannten Merkmale möglichst weitgehend entspricht.“* Mit dem gleichen Argument der Realitätsnähe werden - mit umgekehrten Vorzeichen - alle Renegaten gebannt, die solche Szenarien benutzen, welche zugunsten der experimentellen Kontrollierbarkeit auf den Anspruch auf Realitätsnähe ganz oder teilweise verzichten - und somit *„gegen die grundlegende Philosophie der Verwendung von Mikrowelten“* verstoßen (Dörner, 1992, S. 63, siehe Kapitel 5). Die Möglichkeit, Verhalten zu provozieren, welches auch unter *real-life* Bedingungen gezeigt würde, wird diesen Szenarien schlichtweg abgesprochen (z.B. Strohschneider et al, 1995, S.197). (Vergleiche aber Eyferth, Schömann und Widowski (1986), die zwischen Situationen mit augenscheinlich ökologischer Validität und Situationen mit spezifischer Validität für spezielle Leistungen unterscheiden.) Das Argument der Realitätsnähe läßt sich soweit

strapazieren, daß aus möglichen Nachteilen einzelner Szenarien Vorteile werden: So wird dem Einwand, das Programm sehe in der Realität gegebene Möglichkeiten nicht vor, etwa entgegnet, *„daß es auch in der Realität oft darauf ankommt, aus dem, was man vorfindet, das Beste zu machen"* (Kreuzig, 1995b, S. 398).

Um dem gewichtigen Argument der „Realitätsnähe" nachzugehen, muß man zunächst unterscheiden, ob diese Realitätsnähe eines computergestützten Problemlöseszenarios sich auf die Nähe zur Realität des vermeintlich simulierten *Systems* und/oder ob sich die Realitätsnähe auf die durch die Steuerung des Problemlöseszenarios gestellten *psychologischen Anforderungen* (im Sinne der ökologischen Validität) bezieht. Im ersten Fall erwartet man, daß die im Szenario abgebildeten Bereiche in derselben Weise wie das Original „funktionieren". Im anderen Fall soll das Szenario „lediglich" die gleichen Anforderungen an den Problemlöser stellen wie das Original, unabhängig davon, ob man von der Arbeitsweise und Strukturierung des Modells auf das Original schließen kann oder nicht. Die vom Szenario gestellten Anforderungen sollen bei den Diagnostikanden ein Verhalten provozieren, welches eine Verhaltensstichprobe des interessierenden Kriteriumsverhaltens darstellt. In den folgenden Abschnitten werden diese beiden Aspekte in Hinblick auf die computergestützten Problemlöseszenarien thematisiert. Anschließend werden die Ergebnisse von empirischen Experten-Novizen Vergleichen referiert, da dieser Untersuchungsansatz u.a. die Realitätsnähe der Szenarien voraussetzt.

4.1 Simulation spezifischer Realitätsbereiche

Besonders zu Beginn des Aufschwungs der komplexen Problemlöseforschung wurde das Argument der Realitätsnähe so ausgespielt, daß die computergestützten Problemlöseszenarien einzelne Realitätsbereiche simulieren würden. Beispielsweise schrieb Dörner (1987, S. 97) *„One can use computers to simulate reality by programming them to represent models of political or economic systems, for example. It is possible to define the social, psychological, economic, and ecological relations of a small city as a network of interrelations and then simulate this with a computer. The computer acts then – more or less in accordance with reality – like a small town."*

Ein solche Simulation spezifischer Realitätsbereiche setzt zunächst ein Modell des zu simulierenden Realitätsbereiches voraus. Mit Leutner (1990, S. 23f.) kann eine Simulation als eine zielgerichtete Arbeit mit einem Modell eines Systems aufgefaßt werden, die Modellbildung ist der Prozeß der Ableitung und Prüfung eines Modells. *„Die Ableitung erfordert eine eingehende Analyse des zu modellierenden Systems, d.h. die Identifikation der zu modellierenden Objekte (Systemelemente) und*

der modellrelevanten Beziehungen zwischen den Objekten." (ebd., S. 24). Das zu modellierende System – das Original – ist *„derart nachzubilden, daß zwischen Original und Modell eindeutige (homomorphe) bzw. umgekehrt eindeutige (isomorphe) Analogiebeziehungen bestehen"* (Stapf, 1995, S. 234). Der Modellbildung folgt die Modellvalidierung. Mit einem Modelltest wird nach Kastner (1995, S. 41) u.a. geprüft, ob das dynamische Verhalten (*Verhaltensgültigkeit*) und die Struktur (*Strukturgültigkeit*) des Modells dem jeweiligen Realsystem entspricht und ob die Modellergebnisse mit den Daten des Realsystems übereinstimmen (*empirische Gültigkeit*). Pawlik (1997, S. 181) bezeichnet den Grad der symmetrischen Entsprechung zwischen Ergebnissen unter Realbedingungen und unter Simulation als *Veridikalität*.

Mit der Qualität der Simulation spezifischer Realitätsbereiche ist es bei computergestützten Problemlöseszenarien meist schlecht bestellt. Voruntersuchungen zu den relevanten Attributen des simulierten Realsystems fehlen (Ausnahmen: z.B. Hasselmann et al., 1993b; Milling, 1996) ebenso wie Modellvalidierungen. Die häufiger anzutreffende Praxis der „Modellbildung" bei der Konstruktion realitätsnaher computergestützter Problemlöseszenarien beschreiben Dörner und Reither (1978, S. 530) *„...dabei hielten wir uns, soweit entsprechende Daten leicht verfügbar waren, an reale Verhältnisse, ansonsten machten wir unsere Annahmen nach Plausibilität"*. Was plausibel ist und was nicht, variiert dabei von Programmkonstrukteur zu Programmkonstrukteur, Experten für den vermeintlich simulierten Realtiätsbereich werden häufig bei der Modellbildung gar nicht erst bemüht. Inwieweit die Steuerung eines computergestützten Problemlöseszenarios erfolgreich verläuft, hängt folglich u.a. davon ab, inwieweit der Problemlöser die Plausibilitätsvermutungen der Programmautoren trifft. So erreichte man beispielsweise im „Heizölhandel" ein optimales Verhältnis zwischen der Anzahl von Tankwagen und der erreichbaren Rendite mit exakt 37 Tankwagen (Funke & Geilhardt, 1996, S. 205; Hasselmann, 1993, S. 120). Bei der Kalibrierung des Systems mußten die Programm-Verantwortlichen für das Maximum dieser Funktion irgendeinen Wert bestimmen und haben sich für diesen Wert entschieden – glücklich der Problemlöser, der sich zufällig für den gleichen Wert entscheidet. Weniger glücklich sind plausible Annahmen der Problemlöser über das Realsystem, falls diese mit den Annahmen der Programmverantwortlichen interferieren. So berichtet Putz– Osterloh (1983, S. 114) beispielsweise, daß einige Problemlöser bei dem Szenario „Schneiderwerkstatt" davon ausgingen, daß der „Hemdenverkauf" durch saisonale Einflüsse beeinflußt würde – dies war aber im Programm nicht implementiert, so daß entsprechende Thesen der Problemlöser den Steuerungserfolg negativ beeinträchtigten.

Die Zweifel an der Gegenstandstreue der frühen Version der „Schneiderwerkstatt" sind mit der Kritik von Funke (1986) bekannt geworden, so daß ein gewisses Bewußtsein für Abbildungsfehler besteht. Seltener thematisiert wird, daß für Simu-

lationen komplexer Systeme *grundsätzlich* Abstraktionen und Reduzierungen vorgenommen werden müssen, daß es sich um „Kondensationen" im Dörnerschen Sinne handelt (Dörner, 1992). Simulationen sind Projektionen der Interessen der Simulationsautoren. Dagegen ist nichts einzuwenden, sofern die Programmautoren ihre Interessen erstens explizieren und sofern sie zweitens akzeptieren, daß aus ihrer Simulations*absicht* allein keine Validität abgeleitet werden kann. Will man argumentieren, *„daß die 'Computerrealität' aufgrund ihrer Konstruktion entsprechend den Eigenschaften der 'wahren' Realität valide Ergebnisse erbringt, die man hinsichtlich der 'wahren' Realität eher generalisieren kann als die Ergebnisse aus Studien mit Denksportaufgaben"* (Dörner et al., 1983b, S. 140), so muß man seine Auffassung der realen Eigenschaften explizieren und begründen und die nachgeordnete Simulation an diesen Vorgaben validieren. Komplexe Systeme zu simulieren, bedeutet, aus der Vielzahl der Attribute des Originals (z.B. Eigenschaften, Verbindungen) einige *auszuwählen* und andere unberücksichtigt zu lassen. Zusätzlich wird eine Maßstabsreduktion in Raum und/oder Zeit eingeführt und das Modell weist stets eigene Merkmale auf. Damit ergeben sich zwei Kriterien für die Bewertung von Simulationen: (1.) Sind die für die Fragestellung *relevanten* Merkmale des Originals im Modell berücksichtigt worden? Und (2.) sind die ausgewählten Merkmale sowie ihre Wechselwirkungen und Kovariationen originalgetreu abgebildet worden?

Um diese Fragen klären zu können, bedarf es Relevanzkriterien, eingehender Modellanalysen und Modellvalidierungen, die - wie bereits ausgeführt - für die Szenarien aus der Forschungsdomäne komplexes Problemlösen überwiegend nicht vorliegen. Zumeist sind nicht einmal die Voraussetzungen für eine Validierung der Simulation, nämlich die Explikation der Simulationsabsichten (oder -ziele) sowie eine Dokumentation aller Phasen der Modellbildung, erfüllt. *„Das Ausmaß an Realitätsnähe, welches so erreicht werden kann, ist"* eben nicht nur - wie Dörner (1981, S. 165) annimmt - *„abhängig von der Fähigkeit des Programmierers und vom Programmumfang (...)"*, sondern im wesentlichen abhängig von der Beschaffenheit des zu simulierenden Originals und von theoretischen Setzungen über das Untersuchungsziel der Simulation. Die Zielsetzung der Simulation bestimmt, welche Attribute welches Realitätsbereiches mit welcher Abbildungsgenauigkeit simuliert werden sollen. Die Wahrscheinlichkeit, daß ein Modell valide ist, ist umso größer, umso weniger Attribute ein modelliertes Objekt aufweist bzw. umso weniger Attribute eines Objektes modelliert werden sollen. Das Vorhaben, bestimmte Elemente einer Flugzeugsteuerung zu simulieren, kann gegenüber der beabsichtigten Simulation des Lebens in einem Entwicklungsland als vergleichsweise erfolgversprechender beurteilt werden. Hinsichtlich der Eignungsdiagnostik gilt es zu bedenken, daß mit zunehmender Realitätsspezifität der Szenarien die Anwendungsbreite/Flexibilität dieser Szenarien als diagnostische Instrumente sinken kann (siehe oben, Abschnitt 3.1.2).

Unter anderem aufgrund der Schwierigkeiten, die sich bei ernsthaften Simulationsversuchen komplexer Realitätsbereiche ergeben würden, verzichten einige Forschungsgruppen explizit auf den Anspruch auf Realitätsnähe und verwenden „künstliche" Umgebungen für ihre komplexen Problemlöseaufgaben, wobei diese „künstlichen" Umgebungen besonders für experimentelle Manipulationen genutzt werden können (z.B. die abstrakten Systeme der „SIM"-Reihe der Hamburger Arbeitsgruppe um Kluwe (z.B. „SIM006", siehe Kluwe et al, 1990) oder das System „Sinus" auf der Basis des universellen „Dynamis"-Programms der Bonner Arbeitsgruppe um Funke (siehe z.B. Funke & Müller, 1988)). Will man mit eignungsdiagnostisch eingesetzten computergestützten Problemlöseszenarien unbedingt „reale Systeme" simulieren, so muß die Auswahl der abzubildenden Attribute des Realitätsbereiches dem spezifischen Verwendungszweck Rechnung tragen. *„Beispielsweise mögen Zufallsprozesse in Scenarios als Abbildung von Unwägbarkeiten in der realen Welt und damit verbundener Intransparenz zunächst plausibel erscheinen. Aber welche diagnostische Folgerung bezüglich* ***kognitiver*** *Leistungen soll aus einer solch oberflächlichen Analogie zur Realsituation möglich sein (will man nicht Glück und Pech als Fähigkeitsdimensionen verwenden)? Desgleichen erzeugen Scenarios, die mittels ›Überraschungen‹ Analogien zu realen Notfällen herstellen sollen, eignungsdiagnostisch fragliche, unstandardisierte Situationen."* (U. Funke, 1991, S. 114)

Im Anschluß an die Modellbildung folgt die zielspezifische Modellvalidierung. Ein Validierungsverfahren für komplexe Simulationsmodelle stellt Page (1983, S. 152 f.) vor. Zunächst folgt eine fünfgliedrige Verifikationsphase zur Überprüfung der Korrektheit von Modellverhalten und -struktur. In einer zweiten Phase wird mit Hilfe einer Sensitivitätsanalyse geprüft, wie der Modelloutput auf Veränderungen des Modellinputs oder der Modellstruktur reagiert. Dem folgt eine Phase der Kalibrierung und des Outputvergleichs. Die letzte Phase der prognostischen und dynamischen Gültigkeitsprüfung gilt nur für Modelle, die in der Praxis für Entscheidungen und Planungen eingesetzt werden.

Weder dieses noch andere Validierungsverfahren wurden auf die Simulationen, die angeblich den computergestützten Problemlöseszenarien zugrundeliegen, angewendet. Die Realitätsnähe wird zumeist lediglich behauptet, die Modellbildung und Modellvalidierung wird aber häufig nicht einmal thematisiert (Ausnahmen: siehe die Ansätze zur „realitätsnahen Modellierung" Abschnitt 3.1.2). De facto kann - wie Funke (1985, S. 446) ausführt - ein und dasselbe Simulationsprogramm mit leichten Modifikationen einmal ein Gebiet in der Sahelzone und ein anderes mal den Verlauf einer Epidemie in einer kleinen Stadt darstellen - daran spürt man, wieviel Realitätsliebe und Kenntnis bei der Auswahl der Simulationen waltet. Kluwe (1995, S. 572) kommt hinsichtlich der Güte der Simulationen zu folgender Einschätzung: *„Bei den bislang in der Kognitionspsychologie verwendeten Programmen kann man nicht*

von Simulationen in dem Sinne sprechen, daß es sich dabei um gezielt konstruierte Modelle zur validen Abbildung spezifischer Realitätsbereiche handeln würde." Er spricht den Programmen zwar eine gewisse Plausibilität und das Potential zur Aktivierung von Vorwissen zu, hält aber insgesamt fest, daß es sich um *„fiktive Umgebungen, mit losen Bezügen zur Realität*" (ebd.) handelt. Als Argument für einen diagnostischen Einsatz läßt er das Postulat der Realitätsnähe nicht gelten, für einen diagnostischen Einsatz genügt es nach Kluwe (1995, S. 575)„ *...nicht, anregende Spiele mit Augenscheinvalidität vorzugeben.*"

Vergleichbar mit der von Kluwe angesprochenen Plausibilität und der Aktivierung von Vorwissen bedeutet „Realitätsnähe" für Putz-Osterloh lediglich (1985, S. 204) *„daß Problemlöser den Systemvariablen Bedeutung beimessen und aus ihrem Wissen Hypothesen über mögliche weitere Variablen und über die Verknüpfungen zwischen ihnen generieren können."* Mittlerweile (Putz-Osterloh, 1995, S. 405) empfiehlt die Autorin anstelle der Bezeichnung „realitätsnahe Systeme" die Bezeichnung „semantisch eingekleidete Systeme" und konstatiert, daß der Umgang mit diesen Systemen deutlich von Alltagssituationen abweicht.

4.2 Realitätsnähe der Anforderungen und der Verhaltensweisen; Ökologische Validität

In anderen Arbeiten wird das Argument der Realitätsnähe so ausgelegt, daß sich der Begriff Simulation nun nicht mehr auf ein bestimmtes „Real-System", sondern auf die *Anforderungen* bezieht, die bei der Simulationssteuerung gestellt werden (z.B. Mané & Donchin, 1989, S. 17). Dörner (1981, S. 165) räumt beispielsweise ein, daß man sich darüber streiten kann, *„ob eine solche Spielsituation tatsächlich verallgemeinerbare Ergebnisse bringt"* (Unterstreichung hinzugefügt), um dann fortzufahren: *„Auf alle Fälle stellt aber eine solche Spielsituation ähnliche Anforderungen an das Denken wie entsprechend reale Systeme (...)"*. (Vergleiche auch Dörner, 1993, S. 130.) Auch Schaub und Strohschneider (1992, S. 117) betonen, daß es nicht um eine naturgetreue Simulation der Realität, sondern um die Konstruktion von Problemstellungen mit bestimmten Anforderungen (Komplexität, Intransparenz usw.) geht. Ebenso gehen Dörner et al. (1988, S. 217) davon aus, daß zwar die Problemsituationen nur auf Rechnern nachgebildet sind, *„die Verhaltensweisen, die wir dabei beobachten können, aber alles andere als künstlich*" sind - hier bezieht sich der Anspruch auf Realitätsnähe offensichtlich auf die bei der Steuerung computergestützter Problemlöseszenarien gestellten *Anforderungen* und auf das bei der Steuerung gezeigte Verhalten der Versuchspersonen. Dörner et al. (1983b, S. 321)

sprechen aufgrund dessen, was in der „Lohhausen"-Studie von den Versuchspersonen gefordert wird und was, *„sehr charakteristisch ist für reale Anforderungen im Alltag bei der Bewältigung von schwierigen Situationen"*, dieser Studie ein hohes Maß an *„ökologischer Validität"* zu, welche sie zugleich den Intelligenztests absprechen. Der Anspruch auf „ökologische Validität" entspricht somit dem aufgrund der „plausibilitätsbedingten Anforderungskorrespondenz" erhobenen diagnostischen Anspruch (siehe Abschnitt 3.1.1). Das Szenario soll als Arbeitsprobe ein Verhalten provozieren, das eine Stichprobe des relevanten Kriteriumsverhaltens darstellt.

Was bedeutet „ökologische Validität"? Die ökologische Validität kann nach Stapf (1995, S. 239) als begriffliche Variante der „externen Validität" aufgefaßt werden und wird auch unter dem Labels „ökologische Repräsentativität", „mundane realism", „phenomenon legitimacy" verwendet (siehe Stapf, ebd.). Nach Guthke (1996, S. 80) wird die ökologische Validität in der russischen Psychologie als „Tätigkeitsbezogenheit" thematisiert. Eine Erhebungs- oder Beobachtungsmethode ist mit Pawlik (1976, S. 61) in dem Maße für eine Person (Personengruppe) *„ökologisch valide"*, in dem *„die mit dieser Methode eingeführten S-Bedingungen eine unverzerrte Stichprobe der in der Grundgesamtheit aller Lebensbedingungen dieser Person (Personengruppe) repräsentierten S-Bedingungen sind (...)."* (Erläuterung: S-Bedingungen steht hier für Umwelt oder Reizbedingungen.) Damit erscheint die „ökologische Validität" verwandt mit der *Kontentvalidität*, die ebenfalls das Verhältnis zweier Mengen zueinander thematisiert: *„In general, content-related evidence demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content"* (APA, 1985, p.10). Während sich bei der Kontentvalidität die Repräsentativität auf das Itemuniversum bezieht, sind ökologisch valide Erhebungs- und Beobachtungsmethoden als Repräsentativum eines definierten Universums an *Stimulusbedingungen* einer Person oder Personengruppe zu sehen. Bezogen auf computergestützte Problemlöseszenarien müßte es sich bei den Stimulusbedingungen um die Anforderungen handeln, die das Szenario an den Problemlöser stellt, also z.B. den Umgang mit Komplexität, Intransparenz, Dynamik. Um ökologische Validität, also ein repräsentatives Verhältnis zweier Mengen zueinander, zu gewährleisten, müßten zunächst beide Mengen hinreichend definiert sein. Hinsichtlich der Beurteilung des erzielten Grades der Übereinstimmung zwischen den beiden Mengen gelten inhaltlich weitgehend die weiter oben (Abschnitt 4.1) für die Simulation spezifischer Realitätsbereiche geschilderten Kriterien. Auch hinsichtlich der Realitätsnähe der Anforderungen und der Verhaltensweisen gilt, daß die Wahrscheinlichkeit der ökologischen Validität sich tendenziell umgekehrt proportional zur Anzahl der im Realitätsbereich bestehenden Anforderungen und Verhaltensweisen verhält. Die an einen Flugkapitän bei der Steuerung des Flugzeugs gestellten Anforderungen und seine Verhaltens-

möglichkeiten sind - wie Pawlik (1997, S. 183) ausführt - im Vergleich zu den Verhaltensmöglichkeiten eines Managers oder Entwicklungshelfers auch in der Realsituation relativ regelgeleitet (normiert). Eine „Simulation" des Fluggeschehens kann nun mit einiger Erfolgswahrscheinlichkeit so gestaltet werden, daß sie die definierte Zahl der möglichen Anforderungen umfaßt und die in der Simulation möglichen Verhaltensweisen kaum eine Einschränkung gegenüber den Verhaltensmöglichkeiten in der Realität bedeuten. Demgegenüber dürfte es ungleich schwerer sein, die an einen Manager oder einen Entwicklungshelfer gestellten Anforderungen und die nicht definierten Verhaltensmöglichkeiten eines Managers oder Entwicklungshelfers „nachzubauen". In Untersuchungen zum komplexen Problemlösen wurde den Anforderungen, denen die untersuchten Personen oder Personengruppen - z.B. Manager - im Alltag ausgesetzt sind, bislang selten systematische Aufmerksamkeit zuteil, diesbezüglich verläßt man sich zumeist auf allgemeine Annahmen (siehe oben, Abschnitt 3.1). Vollständig uneinlösbar wird der Anspruch auf ökologische Validität aber dann, wenn nicht nur die Anforderungen und Verhaltensmöglichkeiten in dem interessierenden Realitätsbereich unbekannt sind, sondern darüber hinaus auch über die Anforderungen, die das eingesetzte Szenario an die einzelnen Personengruppen stellt, lediglich spekuliert werden kann (siehe Abschnitt 2.3). Computergestützte Problemlöseszenarien können daher nicht als „ökologisch valide" Erhebungs- oder Beobachtungsmethode für eine bestimmte Person(engruppe) bezeichnet werden, nicht einmal die Voraussetzungen für eine „Korrespondenzbeurteilung", nämlich die Beschreibung der subjektiven und der Kontext-Seite sind erfüllt (siehe Kaminski, 1988, S. 160). Die Frage nach der externen Validität der mit computergestützten Problemlöseszenarien gewonnenen Verhaltensdaten, die in der vorliegenden Arbeit interessierende Frage also, inwieweit diese Daten u.a. auf andere Situationen und Zeitpunkte verallgemeinert werden können, läßt sich nicht mit dem Hinweis auf den angeblichen Simulationscharakter der Szenarien abtun.

Selbst wenn man von der Repräsentativität zweier definierter Mengen - und somit von der Fachbedeutung des Begriffs „ökologische Validität" - absieht und sich auf Plausibilitätsdiskussionen einläßt, entdeckt man deutliche Unterscheide zwischen den *psychologischen* Anforderungen, die bei der Steuerung computergestützter Problemlöseszenarien und bei den vermeintlich simulierten realen Problemen gestellt werden. Folgende Unterschiede können hier beispielhaft - und unvollständig - aufgezählt werden: sinnlich-körperliche Erfahrung versus abstrakte Zugangsweise (siehe Schönpflug, 1993, S. 213 zu „symbolischen" und „anschaulichen" Simulationen), sozial vermittelte Informationen versus Informationsabfragen, Echtzeit- versus Zeitrafferprozesse sowie „Ernstfall" versus „Spiel". Das Bewußtsein, Teilnehmer einer wissenschaftlichen Untersuchung zu sein, verändert das Verhalten. Dieser „Hawthorne- Effekt" (z.B. Bortz & Döring, 1995, S. 472) gilt auch bei

Untersuchungen, bei denen Szenarien eingesetzt werden. Bei den Szenarien handelt es sich um virtuelle Situationen, und *„virtuelle Situationen erzeugen virtuelles Verhalten"* (Funke, 1995a, S. 208). Im Ernstfall werden außerdem in der Regel Personen mit solchen Problemen konfrontiert, welches ihnen wenigstens etwas vertraut ist, möglicherweise sind es sogar Experten mit problemlösungsrelevantem Fachwissen. Demgegenüber betrachtet die Forschung zum komplexen Problemlösen überwiegend das Verhalten naiver Versuchspersonen, die *„ohne Vorerfahrung in kurzer Zeit ungewöhnliche, neuartige Anforderungen bewältigen"* müssen (Kluwe, 1990a, S. 251). Die Diskrepanz zwischen der postulierten ökologischen Validität und den tatsächlich eingesetzten Programmen wird schließlich auch darin deutlich, daß die Steuerung computergestützter Problemlöseszenarien überwiegend auf eine *Überforderung* der Szenarienteilnehmer hinaus läuft. Schon bei Dörner und Reither (1978, S. 527) heißt es, daß die Probanden *„fast ausnahmslos das ursprünglich stabile Gefüge der Variablen des simulierten Landes zerstörten und dadurch häufig katastrophale Zustände schufen"*, und diese Überforderung setzt sich in zahlreichen Problemlöseszenarien fort. Kreuzig (1995a, S. 99) konstatiert, daß computergestützte Problemlöseszenarien für das Entdecken von potentiellen *Krisenmanagern* prädestiniert seien (für weitere Beispiele und zur meßtheoretischen Bedeutung dieses Überforderungsaspekts siehe Abschnitt 7.2). Selbst bei einem pessimistischen und misanthropischen Weltbild mag man doch darauf bestehen, zwischen dem Alltag und einer Katastrophe, einer Krise, zu unterscheiden. Wie will man eine Krise anders definieren als eine Abweichung von der Normalität? Vergleichbar der Themenwahl der Medien werden die in der Forschung eingesetzten Szenarien offensichtlich häufig nach den Kriterien „Krise, Katastrophe, Naturereignis und Folklore" ausgewählt, realitätsorientierte Alltags-Anforderungen und problemrelevantes Hintergrundwissen der Versuchspersonen bleiben dabei gerade unberücksichtigt. Die Überforderung der Versuchspersonen kann sinnvoll sein, um extremes Verhalten wie z.B. „Notfallreaktionen" (z.B. Dörner 1981, S. 171) untersuchen zu können. Ein solches Ziel ist aber unverträglich mit dem Anspruch auf Repräsentativität für *Alltags*anforderungen und für *typische* Berufsanforderungen.

4.3 Experten-Novizen Vergleiche

Auf die Annahme der „Realitätsnähe" bauen Experten-Novizen Vergleiche. Solche Vergleiche werden neben anderen Methoden eingesetzt, um die Effekte unterschiedlichen Wissens auf das Problemlöseverhalten zu untersuchen (siehe Putz-Osterloh, 1988, S. 252). Bei der Problembearbeitung soll sowohl bereichsspezifisches Wissen

als auch bereichsübergreifendes, heuristisches Wissen angewendet werden. Versuchspersonen können Experten für die „Realität" sein, auf die das Szenario Bezug nimmt (so sollen etwa Wirtschaftswissenschaftler Experten für Szenarien mit wirtschaftswissenschaftlicher Semantik sein) oder aber allgemeine „Problemlösungsexperten" im Sinne des heuristischen Wissens. Heuristisches Wissen wird von Dörner (1989c, S. 133) gleichgesetzt mit der „operativen Intelligenz" als *„Fähigkeit, jeweils am richtigen Ort die richtige intellektuelle Operation durchzuführen."* Experten-Novizen Vergleiche betreffen daher sowohl die Güte der Simulation spezifischer Realitätsbereiche (falls diese gelungen ist, können Experten ihr bereichsspezifisches Wissen bei der Problembearbeitung vorteilhaft nutzen) als auch die Realitätsnähe der Anforderungen solcher Aufgaben. Sofern die computergestützten Problemlöseszenarien *typische* Anforderungen des alltäglichen Problemlösens widerspiegeln, müßten „Experten für komplexe Probleme" - und als solche werden z.B. Manager angesehen, siehe oben Abschnitt 3.1.1 - sich in *allen* Problemlösungsaufgaben dank ihres *„im Laufe einer erfolgreichen Managementkarriere (...)"* entwickelten *„generell anwendbaren heuristischen Wissens"* (Strohschneider & Schaub, 1991, S. 327) als überlegen erweisen. (Diese Annahme eines allgemeinen, domänenübergreifenden Expertentums im Sinne eines „general problem solvers" (Newell & Simon, 1972) steht allerdings im Widerspruch zu Erkenntnissen der Experten-Novizen Forschung, die überwiegend die Domänen*spezifität* als definierendes Merkmal des Expertentums ansieht (siehe z.B. Frensch & Sternberg, 1989, S. 160f).) In den Experten-Novizen Vergleichen im Kontext der Problemlöseforschung sind die beiden postulierten Aspekte des Expertentums (bereichsspezifisches und heuristisches Wissen) häufig konfundiert. Aufschluß darüber, ob nun zur Problembearbeitung heuristisches oder bereichsspezifisches Wissen eingesetzt wird, verspricht man sich u.a. davon, daß die Experten mit mehreren Problemen konfrontiert werden, bei denen nur eines ihrem bereichsspezifischen Wissen entgegenkommt. Von einer anderen Prüfungsmöglichkeit, dem Experten-Novizen Vergleich bei der Steuerung „künstlicher" Systeme, wurde bislang kein Gebrauch gemacht, obwohl diese Systeme eine isolierte Betrachtung des „heuristischen Wissens" ermöglichen würden.

Putz-Osterloh (1987) verglich sieben Professoren und 30 Studenten der Wirtschaftswissenschaften hinsichtlich ihrer Leistung bei der Steuerung eines wirtschaftswissenschaftlichen Problemlöseszenarios („Schneiderwerkstatt") und bei der Steuerung des Entwicklungshilfeszenarios „Moro". Für die „Schneiderwerkstatt" - nicht aber für „Moro" - konnte die Autorin einen Leistungsvorteil der Experten gegenüber den Novizen ausmachen. Bei einem von Putz-Osterloh und Lemme durchgeführten (1987) Experten-Novizen-Vergleich mit den gleichen Szenarien erzielten Studenten der Wirtschaftswissenschaften als Experten in beiden Szenarien bessere Leistungen als Studenten anderer Fachgruppen. Ein Vergleich der Daten der Wirt-

schaftsstudenten dieser Studie mit den Daten der Professoren der zuerst genannten Studie zeigte für das betriebswirtschaftlich eingekleidete Szenario keine bedeutsamen Unterschiede in der Steuerungsleistung. In einer anderen Studie dieser Arbeitsgruppe erzielten 27 Stabsoffiziere der Führungsakademie in Hamburg als „Strategieexperten" bei der Bearbeitung des Szenarios „Feuer" weder bessere Leistungen als zwei studentische Gruppen mit 30 und 25 Personen noch setzten die „Experten" effizientere Strategien ein (Putz-Osterloh & Haupts, 1990, S. 137).

18 Studenten der Betriebswirtschaftslehre dienten in einem Experiment von Renkl, Gruber, Mandl und Hinkofer (1994) als (Semi-)Experten für das Planspiel „Jeansfabrik". Hinsichtlich der Zielvariablen „Gewinn" schnitten diese Experten schlechter ab als die Novizen (17 Pädagogik- und Psychologiestudenten).

Strohschneider und Schaub (1991, siehe auch Schaub & Strohschneider, 1992) konstatierten für 45 leitende Angestellte (Manager) geringfügige Leistungsvorteile bei der Bearbeitung des „Moro"-Systems gegenüber 45 Studenten. Eine Teilstichprobe von 23 leitenden Angestellten bearbeitete außerdem noch das Szenario „Manutex", welches nach Ansicht der Autoren Variablen und Relationen realisiert, *„die auch im beruflichen Alltag des typischen Managers von eminenter Bedeutung sind"* (Strohschneider & Schaub, 1991, S. 327 f.). Die Ergebnisse, die die „Manager" bei „Manutex" erzielten, wurden den entsprechenden Leistungen einer Gruppe von 25 Studenten gegenübergestellt. Die Vergleichbarkeit ist u.a. dadurch eingeschränkt, daß die Manager im Gegensatz zu den Studenten vorab schon Erfahrungen mit dem Problemlöseszenario „Moro" sammeln konnten. Als Ergebnis sahen die Autoren zwar auf der „strategischen Ebene" Unterschiede zwischen den Gruppen. Diese strategischen Unterschiede waren aber leistungsirrelevant. Weder den Managern noch den Studenten gelang es, die Firma auf Gewinnkurs zu bringen, hinsichtlich der Variable „Gesamtvermögen" zeigten sich abschließend keine Gruppenunterschiede. Verhaltens- oder Strategieunterschiede zeigten sich auch bei Reither (1981) bei einem Vergleich von 12 erfahrenen und 12 unerfahrenen Entwicklungshelfern bei der in Kleingruppen zu dritt vorgenommenen Steuerung eines Entwicklungshilfeszenarios. Ungeachtet dieser Verhaltensunterschiede steuerten sowohl die Experten als auch die Laien das System in katastrophale Zustände.

Reichert und Stäudel (1991) berichten, daß Manager Studenten mit betriebswirtschaftlichen Kenntnissen hinsichtlich der Problemlösegüte und -strategien bei der Steuerung des betriebswissenschaftlich eingekleideten Szenarios „Schoko-Max" überlegen waren, obwohl sich auch bei der Managergruppe *„typische strategische Fehler"* (ebd., S. 105) fanden. Die Stichprobengröße wurde ebenso wie andere Angaben, die für die Beurteilung der Studie notwendig sind- nicht berichtet. Im gleichen Buch berichten Kreuzig und Schlotthauer (1991) über einen Leistungsvergleich von Studenten wirtschaftswissenschaftlicher und technischer Fachrichtungen

bei der Steuerung des ebenfalls wirtschaftlich eingekleideten Szenarios „Manage!". Dabei zeigte sich keine Fachrichtungseffekte. Dieser Befund läßt die Autoren nicht an der „Realitätsnähe" zweifeln, die ja dazu hätte führen können, daß die wirtschaftswissenschaftlichen Studenten aufgrund ihres bereichsspezifischen Wissens besser abschneiden. Das Ergebnis zeigt für die Autoren vielmehr hypothesenkonform die „Robustheit" des Instruments. Trotz der angeblichen Simulation eines wirtschaftswissenschaftlichen Realitätsbereiches soll der Einfluß von Fachwissen auf die Steuerungsleistung nämlich - laut Kreuzig (1995b, S.388) - ausgeschlossen werden.

Bei E. Müller (1991) erzielten 48 Tankwagenfahrer als Experten in der Simulation „Tankwagen" nicht einmal 10% der Problemlöseeffizienz der Novizen (Studenten). Dieser Befund wird u.a. darauf zurückgeführt, daß Tankwagenfahrer als Sicherheitsexperten ('inventive worrier') im Sinne Schönpflugs (1989) sich mehr auf die Identifikation möglicher Gefahren orientieren, was sich zu Lasten der Produktivität des Verhaltens auswirken würde. In der Studie wurde es allerdings versäumt, die Intelligenz der bildungsheterogenen Gruppen zu kontrollieren.

Hasselmann (1993) verglich die Leistungen, die 17 Führungsnachwuchskräfte einer Bank (nähere Angaben siehe unten, Abschnitt 9.2.3.3) bei der Steuerung des betriebswissenschaftlich eingekleideten Szenarios „Textilfabrik" erzielten, mit den Leistungen von 11 Studenten der Betriebswirtschaftslehre und 41 Studenten anderer Studienrichtungen. Die Bank-Führungsnachwuchskräfte zeigten bessere Leistungen als die Studentengruppe der nicht-wirtschaftswissenschaftlichen Fachrichtungen. Im Vergleich zur Gruppe der Studenten der Betriebswirtschaftslehre erzielten die Führungskräfte zwar nominell bessere Leistungen, dieser Leistungsvorsprung ließ sich aber nur für eines der drei bestimmten Gütemaße statistisch absichern, der multivariate Test verfehlte knapp die Signifikanz.

Insgesamt gesehen sind die Befunde zum Experten-Novizen Vergleich uneinheitlich und aufgrund spezifischer (Extremgruppenvergleich, unzureichende Parallelisierung der nicht interessierenden Personmerkmale der Gruppen wie Alter, Intelligenz, Motivation usw., keine Kontrolle des Vorwissens und der indiviudellen Unterschiede zwischen den Experten) und allgemeiner Probleme (fragliche Reliabilität der Gütemaße, fragliche Definition von „Strategien") schwer zu interpretieren. Als empirischer Nachweis der Realitätsnähe oder der ökologischen Validität können die Studien nicht gewertet werden. Selbst wenn der Nachweis eines Experteneffekts gelingen würde, wäre dieser für den diagnostischen Einsatz computergestützter Problemlöseszenarien differenziert zu bewerten. Sofern dieser Effekt eindeutig auf das allgemeine heuristische Wissen oder die operative Intelligenz zurückzuführen wäre, könnte dies - im Verbund mit dem Nachweis der Kriteriumsvalidität - für den diagnostischen Einsatz der Instrumente sprechen. Ein solcher Effekt müßte sich dann aber über verschiedene Systeme empirisch generalisieren lassen. Wäre der Exper-

ten-Effekt hingegen auf das bereichsspezifische Vorwissen der Experten zurückzuführen, so würde dies die diagnostischen Einsatzmöglichkeiten einschränken: computergestützte Problemlöseszenarien könnten unter dieser Voraussetzung nur in den Fällen verwendet werden, in denen neben anderen Fähigkeitsunterschieden auch Unterschiede in der entsprechenden Wissensdomäne untersucht werden sollen. Als diagnostisches Instrument für Berufsanfänger oder Einsteiger in einen neuen Bereich eignen sich vorwissensabhängige Instrumente hingegen nicht in jedem Fall.

4.4 Zusammenfassung, Schlußfolgerungen und Ausblick

Computergestützte Problemlöseszenarien bieten keinesfalls ohne weiteres die Möglichkeit, *„interessierende Problemsituationen mittels Simulationen im Computer aus der Realität in das psychologische Labor zu transferieren"* (Strohschneider & Schaub, 1995, S. 188). Es handelt sich bisher fast ausschließlich um *„fiktive Umgebungen, mit losen Bezügen zur Realität"* (Kluwe, 1995, S. 572). Der Anspruch auf eine valide Simulation eines realen Systems ist für die bislang in der denkpsychologischen Forschung und psychologischen Diagnostik eingesetzten computergestützten Problemlöseszenarien unbegründet, entsprechende Modellanalysen, Spezifikationen von Abbildungsvorschriften und Modellvaldierungen wurden nicht vorgenommen. Ebensowenig kann ein Anspruch auf ökologische Validität im Sinne einer repräsentativen Abbildung der Anforderungen, die im Alltag an die untersuchten Personen(gruppen) gestellt werden, geltend gemacht werden. Insgesamt macht es daher wenig Sinn, bezüglich der in der Psychologie bislang eingesetzten computergestützten Problemlöseszenarien von computer*simulierten* Szenarien oder Systemen oder von Computer*simulationen* zu sprechen.

Die Relevanz der bei der Steuerung eines computergestützten Problemlöseszenarios gezeigten Verhaltensweisen für alltägliche (z.B. berufsbezogene) Verhaltensweisen, die externe Validität der mit den Szenarien erhobenen Daten, kann nicht a priori gesetzt werden, sondern muß - wie für andere psychologische Instrumente auch - empirisch geprüft werden.

Experten-Novizen Vergleiche mit computergestützten Problemlöseszenarien haben bislang uneinheitliche Ergebnisse erbracht, insgesamt können diese Untersuchungen weder als Beleg der Realitätsnähe und ökologischen Validität der Szenarien noch als Beleg für die praktische Relevanz der im Labor bei der Steuerung des computergestützten Problemlöseszenarios gezeigten Verhaltensweisen gewertet wer-

den. Die Befunde *„rechtfertigen schon gar nicht die kommerzielle Nutzung solcher Systeme als Instrumente für eignungsdiagnostische Zwecke"* (Kluwe, Schilde, Fischer & Oellerer, 1991c, S. 307).

Durch die mehr oder minder losen Bezüge der Rahmengeschichte einiger Szenarien zu Realitätsbereichen können sich bei der Bearbeitung computergestützter Problemlöseszenarien Vorwissenseffekte ergeben. Solche Vorwisseneffekte können sich bei der eignungsdiagnostischen Verwendung von computergestützten Problemlöseszenarien je nach diagnostischer Fragestellung sowohl hinderlich als auch förderlich erweisen und sind auf jeden Fall zu kontrollieren. Eine solche Kontrolle der Vorwissenseffekte ist für den empirischen Teil der Arbeit vorgesehen. Außerdem gibt die geplante Studie zur Kriteriumsvalidität Auskunft über die praktische Relevanz der bei der Problembearbeitung beobachteten Verhaltensweisen.

5. Warum einige Regeln der „Philosophie der Verwendung von Mikrowelten" nicht auf die Verwendung von computergestützten Problemlöseszenarien zur Fähigkeitsdiagnostik angewandt werden können

Mit der Verwendung von computergestützten Problemlöseszenarien war und ist auch ein fundamentaler wissenschaftstheoretischer Streit *„Über die richtige Art, Psychologie zu betreiben"* (Grawe, Hänni, Semmer & Tschan, 1990) verbunden. Dabei werden zum Teil vertraute Debatten, wie z.B. die Diskussion um den Stellenwert der Dispositions-(Struktur-) Forschung einerseits und der Prozeßforschung andererseits wiederbelebt (siehe z.B. Jäger, 1984, S. 33f.; Wittmann & Matt, 1986, S. 310ff.). Dieser wissenschaftstheoretische Streit muß hier aufgegriffen werden, da die aus der »Philosophie der Mikrowelten« abgeleiteten Regeln zur Verwendung computergestützter Problemlöseszenarien den Voraussetzungen eines Einsatzes dieser Instrumente zur Fähigkeitsdiagnostik teilweise widersprechen.

Werden computergestützte Problemlöseszenarien zu Diagnosezwecken eingesetzt, so sind sie - ebenso wie Tests - als diagnostische Instrumente zu klassifizieren und den differentiellen Methoden zuzuordnen. Differentielle Methoden dienen nach Stapf (1995, S. 233) der Beobachtung und Messung individueller Differenzen zwischen Individuen bzw. individuellen Ausprägungen von Personvariablen. Die folgenden Ausführungen zu Testverfahren von Stapf (ebd.) gelten somit auch für diagnostisch eingesetzte Problemlöseszenarien: *„Aus Gründen der Vergleichbarkeit der Testresultate muß ein Testverfahren allen zu untersuchenden Personen in gleicher, standardisierter Weise appliziert werden. D.h. jegliche Variationen der Situations-Variablen sind untersagt, da ja das Testergebnis allein von der individuellen Merkmalsstruktur der Testperson und keinesfalls von fördernden oder hemmenden Bedingungen der Untersuchungssituation abhängen soll."* Wenn nach Pauli (zitiert nach Stapf, ebd.) ein Test ein unvollständiges Experiment darstellt (siehe auch Michel & Conrad, 1982, S. 1), so gilt dies auch für diagnostisch eingesetzte computergestützte Problemlöseszenarien.

Im folgenden wird referiert, daß sich der Gedanke des Experimentierens nach Ansicht Dörners (1989a, 1992) teilweise nicht mit der *„Philosophie der Verwendung von »Mikrowelten« oder »Computerszenarios«"* verträgt. Es wird die These vertreten

und begründet, daß man explizit gegen einzelne „Regeln" verstoßen muß, die Dörner aus seiner „Philosophie" ableitet, wenn man computergestützte Problemlöseszenarien als Instrumente zur Fähigkeitsdiagnostik - als Tests im weiteren Sinne und somit als unvollständige Experimente - verwenden will.

5.1 Die „Philosophie der Verwendung von Mikrowelten"

Nach Dörner (1992, S. 57) tritt bei der Verwendung computergestützter Problemlöseszenarien anstelle des einfachen und konstanten „Reizes" der klassischen experimentalpsychologischen Situation *„eine relativ komplexe und inkonstante dynamische Mikrowelt"*, und anstelle der „Reaktion" der klassischen experimentalpsychologischen Situation tritt *„eine komplizierte Interaktion des Probanden mit der Mikrowelt, die sichtbar aus Informationsabfragen und aus Maßnahmen besteht, weiterhin - oft »unsichtbar« - aus Prozessen der Hypothesenbildung, des Planens, des Prognostizierens, usw."* (ebd.). Durch die Reaktionen der Teilnehmer verändern sich die Variablenausprägungen des jeweiligen Systems, jede Systemveränderung bestimmt mit über den nächsten Eingriff des Problemlösers. Eingriffe in das System schaffen also folgenreiche Tatsachen - wer auf dem Tiger reitet, kann nicht herab, lautet ein chinesisches Sprichwort. Zwei Szenarien stellen daher - vernachlässigt man einmal den Aspekt der Vorwissensaktivierung bei semantisch eingekleideten Systemen - lediglich zu Beginn der Steuerung „gleiche Aufgaben" dar, ansonsten bestehen (1.) inkommensurable Bedingungen und somit im Vergleich verschiedener Personen inkommensurable Verhaltensformen. Mikrowelt-Interaktionen sind (2.) nicht-wiederholbar. Aufgrund der Menge der Variablen und ihrer Ausprägungen ist (3.) keine isolierte Bedingungsvariation möglich, aufgrund der Datenmassen entsteht (4.) eine „Signifikanzabundanz" in dem Sinne, daß sich mit großer Wahrscheinlichkeit irgendwelche signifikanten Zusammenhänge per Zufall aufzeigen lassen (Aufzählung nach Dörner, ebd., S. 60). Diese Schwierigkeiten beim Umgang mit Mikrowelten können nach Ansicht Dörners nicht einfach ausgeräumt werden, indem man Maßnahmen zum Zwecke der besseren experimentellen Hantierbarkeit ergreift. Eingriffe zur Vereinfachung der Systeme (z.B. durch Verminderung der Komplexität), Kontrollen der Vorwissenseffekte (z.B. durch abstrakte Systeme) und auf Einzelaspekte konzentrierte Auswertungen (z.B. Auszählung der Maßnahmehäufigkeit) verstoßen nach Ansicht Dörners (1992, S. 62f) gegen die grundlegende Philosophie der Verwendung von »Mikrowelten«, da mit jeder dieser Maßnahmen eine Einschränkung des beobachtbaren Verhaltens verbunden ist. *„Eine Mikrowelt ist eine »Kondensation« einer Handlungssituation, und zugleich soll das Verhalten einer Vp beim Um-*

gang mit einer Mikrowelt eine »Kondensation« ihres Verhaltens in »echten« Realitätsausschnitten sein.[6] *D.h., daß das Verhalten nicht nur einzeln aus »Entscheiden«, »Planen«, »Schlußfolgern«, »Hypothesenbilden«, »Urteilen« besteht, sondern aus all diesen Prozessen zusammen. Damit soll die Konstellation des Gesamtprozesses erhalten bleiben, und man kann dessen »kompositorische« und »kontextualen« Merkmale erforschen. – Dies ist die »Philosophie«, die hinter der Verwendung der Mikrowelten steht!"* (Dörner, 1992, S. 59). Explizit wertet Dörner auch die Vorgabe fester Ziele als Verstoß gegen das Kondensationsprinzip und somit gegen die grundlegende »Philosophie der Mikrowelten«: *„...so sollte man sich darüber im klaren sein, daß man z.B. mit festen Zielvorgaben für eine Versuchsperson die Prozesse der Zielpräzisierung, der Zielauswahl, der Schwerpunktbildung, der Zielbalancierung nicht mehr ansichtig werden kann, da sie nicht mehr stattfinden werden"* (S. 62).

Die weiter oben genannten Schwierigkeiten beim Umgang mit Mikrowelten sind nach Dörners Ansicht nicht primär der Verwendung computergestützter Problemlöseszenarien geschuldet, sondern spiegeln genuine Forschungsprobleme der Psychologie wieder. In seinem 1989 erschienenen Beitrag zur Zeitschrift *Sprache & Kognition* veranschaulicht Dörner seine Position mit einer Fabel über kleine grüne außerirdische Schildkröten. Diese Fabel soll aufzeigen, daß die seiner Auffassung nach ritualhaft betriebene Forschung nach dem varianzanalytischen Prinzip zur Erforschung der Merkmale der verborgenen Maschinerie „Seele" untauglich sei. Herrmann (1990, S. 8 f.) hat diese Kritik der Experimentiermethodik als *Systemargument* bezeichnet und wie folgt zusammengefaßt: *„Der Erkenntnisgegenstand der psychologischen Forschung hat Systemcharakter. Beobachtbare bzw. meßbare Systemoutputs sind nicht nur von Umgebungsbedingungen des Systems, sondern von einer Vielzahl gegenwärtiger und früherer Systemzustände determiniert, wobei diese Systemzustände in komplizierten dynamischen Wechselwirkungen zueinander stehen. Da man im Experiment notwendigerweise nur relativ wenige Bedingungen kontrolliert, und da man per Randomisierung die Vielzahl von nicht beherrschbaren Systemzuständen und deren Wechselwirkungen sozusagen überspielt, so rächt sich das durch eine nur sehr mäßige Varianzaufklärung und durch unbefriedigende Replizierbarkeit. Also fort mit dem Experiment! Das Seelenleben ist für das Experimentieren zu komplex, zu sehr vernetzt und zu dynamisch."* In seiner Replik auf dieses Argument betont Herrmann (ebd.) u.a., daß die Komplexität eines Systems – eine theoretische Durchdringung des Systems vorausgesetzt – nicht das Vorliegen relativ einfacher, empirisch gehaltvoller (und experimentell prüfbarer) Aussagen über das Sy-

[6] Anmerkung: Hier wird der doppelte Anspruch auf die Simulation spezifischer Realitätsbereiche einerseits (siehe Kapitel 4.1) *und* auf die ökologische Validität (siehe Kapitel 4.2) andererseits, explizit zum Programm erhoben.

stem inhibiert und daß man zur Theorieprüfung im allgemeinen auch die Experimentiermethodik benötigt, sofern man nicht eine bessere alternative Datenerhebungsmethode aufweisen kann. Dörner (1989a, 1992) schlägt als alternative Datenerhebungsmethode Einzelfallbeobachtungen über lange Zeiträume vor, wobei computergestützte Problemlöseszenarien diesbezüglich den Vorteil aufweisen, umfassende Verhaltensprotokolle zu generieren. Diese Protokolle sollen nach Dörners Vorstellungen ohne „numerische Reduktion" mit allen ihren Details studiert werden (*„»Philosophie« der Einzelfallbetrachtung"*, Dörner, 1992, S. 71). U. a. mit Hilfe der *„»Idealstrategie« des Verhaltens"* (ebd., S. 73) kommt man dann von der Phänomenbetrachtung zur Erklärung, indem man sich ein möglichst gutes Verhalten für die jeweilige Situation überlegt und diese Idealstrategie mit dem tatsächlichen Verhalten vergleicht. Eine andere Methode besteht darin, das Verhalten von Versuchspersonen mit und ohne Erfolg zu vergleichen und sich Gründe für den Erfolg oder Mißerfolg zu überlegen. Die so erzielten Einzelerklärungen können in Form von „Wenn ... dann" Aussagen formuliert werden - womit den unterschiedlichen inneren und äußeren Bedingungen des Verhaltens Rechnung getragen werden soll - und dann in einer netzartigen Ablaufstruktur vereint werden. Durch die „Wenn ... dann" Aussagen bleibt nach Dörner (ebd. S. 58) der konstellative Aspekt bei der Datenreduktion erhalten. Dieser „Kondensation" stellt der Autor der üblichen „Dissektion" gegenüber. Die so gewonnenen Aussagen über unwiederholbare Prozesse können z.B. mit Hilfe der „Tonscherben-Rekonstruktionsstrategie" (oder „Amphorenstrategie" oder „laterale Validierung") geprüft werden. Dabei wird über die Gültigkeit eines Einzelergebnisses aufgrund der Existenz 'stützender' Nebenergebnisse entschieden, Einzelergebnisse gelten als bewährt, wenn sie mit anderen Ergebnissen konkordant sind (Strohschneider, 1996b, S. 46). Die Einzelfälle sollen aufgrund eines allgemeinen theoretischen Systems erklärt werden, ein „erzeugendes System" soll gefunden werden. Neben der „Tonscherben-Rekonstruktionsstrategie" sieht Dörner eine weitere Alternative zur Experimentiermethodik in der „Verhaltensraumstrategie". Diese Strategie besteht darin, das verhaltenserzeugende System selbst zu simulieren und somit künstliches Verhalten zu generieren, welches man dem natürlichen Verhalten gegenüberstellt. Beispielsweise kann man dann vergleichen *„...ob sich in dem künstlichen Verhaltensraum und in dem natürlichen Verhaltensraum bei der Bewältigung des jeweiligen Problems das gleiche Ausmaß an Erfolg bzw. Mißerfolg herausstellt."* (S. 82). Dieser Ansatz des „kognitiven Modellierens" läuft daraus hinaus, z.B. mit Hilfe der Simulation realitätsnaher Aufgaben selbst eine Simulation ganz anderer Art erstellen zu können: eine Simulation psychischer Prozesse. Ziel sind lauffähige Software-Programme, operierende Modelle psychischer Abläufe, die das gleiche Verhalten hervorbringen wie die Versuchspersonen, die also selbst neue Daten erzeugen, an denen dann Hypothesen geprüft werden können.

(Zum theoretischen Ansatz der Modellbildung mit Hilfe von Computersimulationen von Informationsverarbeitungsprozessen siehe z.B. Dörner, 1995 sowie Opwis & Plötzner, 1996; für einzelne Arbeiten siehe beispielsweise Kluwe, 1991; Kluwe, Misiak & Haider, 1989, 1991b; Ringelband, Misiak & Kluwe, 1990; Schaub, 1993, Schoppek, 1996). Herrmann (1990, S. 6 f.) hat aufgezeigt, daß die Methoden des kognitiven Modellierens und der Experimentalpsychologie keinen strukturellen Gegensatz darstellen, sondern die Proponenten des kognitiven Modellierens vielmehr auf die Ergebnisse der Experimentalpsychologie angewiesen sind, die z.B. als *constraints* bei der Modellierung berücksichtigt werden.

Im folgenden Abschnitt wird herausgearbeitet, daß sich einige Regeln der „Philosophie der Verwendung von Mikrowelten" nicht auf die Verwendung von computergestützten Problemlöseszenarien zur Fähigkeitsdiagnostik angewandt werden können. Die „Philosophie der Mikrowelten" wird somit aus einer bestimmten (diagnostischen) Perspektive heraus betrachtet und bewertet. Nimmt man die Ausführungen Dörners innerhalb ihrer eigenen Prämissen wahr, fällt vor allem ein zentraler Widerspruch auf: Einerseits wertet Dörner (ebd., S. 62 f.) die Vorgabe fester Ziele explizit als Verstoß gegen das Kondensationsprinzip und somit gegen die grundlegende »Philosophie der Mikrowelten« andererseits setzt sowohl die im Rahmen der »Philosophie der Mikrowelten« empfohlene *„»Idealstrategie« des Verhaltens"* als auch die *„Verhaltensraumstrategie"* die Vorgabe fester Ziele eigentlich voraus: Gleichviel ob man sich – wie von Dörner empfohlen – *„ein möglichst »gutes« Verhalten"* für die jeweilige Situation überlegt (S. 73), ob man Personen *„die bezüglich Erfolg und Mißerfolg bei der entsprechenden Aufgabe extrem sind"* miteinander vergleicht (ebd., S. 74) oder ob man prüft, ob das Computerprogramm *„das gleiche Ausmaß an Erfolg oder Mißerfolg produziert"* wie die Versuchspersonen (ebd., S. 82): stets soll man das Verhalten der Versuchspersonen oder des Verhalten hervorbringenden Rechners ganz offensichtlich hinsichtlich Erfolg bzw. Mißerfolg ***bewerten***. Dies setzt ein Bewertungskriterium voraus, welches sich zwingend aus dem Ziel der steuernden Person ableiten muß. Da dieses Ziel nicht vorgegeben werden darf, ist eine solche Bewertung äußerst fraglich. Dörner selbst schreibt: Wenn die Ziele offen sind *„... ist es möglich, daß die eine Person dem einen Ziel, eine zweite einem anderen, eine dritte wechselnden Zielen zustrebt; wie soll man hier die Verhaltensformen miteinander in Beziehung setzen? Man vergliche Äpfel nicht nur mit Birnen, sondern mit Kohlköpfen."* (ebd, S. 61). Auch die Bewertung im Einzelfall funktioniert ohne Zielvorgabe nicht. In den Verlaufsprotokollen ist kein „Ziel" eingetragen, diese Ziel könnte auf Einzelfallebene höchstens von den Forschern bei der Interpretation der Daten nachträglich unterstellt werden. Auch die Versuchspersonen selbst stehen im nachhinein nicht als valide Informanten über ihr individuelles Ziel zur Verfügung. Die individuelle Zielsetzung hat sich wo-

möglich im Laufe der Bearbeitung verändert oder das Ziel wird - selbstwertdienlich - nachträglich passgenau zu den Ergebnissen konstruiert. Ohne Zielvorgabe ist keine Bewertung möglich. Die Bewertung des Verhaltens nach Erfolg bzw. Mißerfolg wird also ebenso eindeutig als Strategie der »Philosophie der Mikrowelten« etabliert wie die Voraussetzung einer solchen Bewertung - nämlich die Etablierung eines Bewertungskriteriums - eindeutig als Verstoß gegen die »Philosophie der Mikrowelten« bezeichnet wird. Während die Ausführungen zur „Kondensation" darauf hinauslaufen, daß man das Verhalten - zwar vergröbert, aber mit intakter Konstellation - ohne Reduktion abbildet und sich von den Daten quasi zur Theoriebildung und kognitiven Modellierung anmuten läßt, wird de facto in zahlreichen Publikationen von Dörner eine Verhaltensbewertung vorgenommen. Auf diesen Verhaltensbewertungen baut z.B. Dörners' Systematik von menschlichen Fehlern (d.h. Mißerfolgsfaktoren) auf (z.B. 1981, 1989b, 1993).

5.2 Zu einigen zentralen Unverträglichkeiten der „Philosophie der Verwendung von Mikrowelten" mit der Zielsetzungen einer Fähigkeitsdiagnostik

Ausgangspunkt der Überlegungen zur »Philosophie der Verwendung von Mikrowelten« ist die Tatsache, daß bei der *Provokation* des Verhaltens mit Hilfe computergestützter Problemlöseszenarien die Standardisierung der Untersuchungsbedingungen nicht gewährleistet ist. Dies führt dazu, daß bei der *Auswertung* und *Interpretation* der Daten die jeweilige konstellative Bedingtheit im Einzelfall betrachtet werden muß. Die vorhandenen konfiguralen Analysemethoden (z.B. die „Konfigurationsfrequenzanalyse" von Krauth und Lienert [1973] oder „Hypag" von Wottawa [1987a]), die es erlauben, Personen mit gleichen Merkmalskombinationen in den für die jeweilige Analyse relevanten Merkmalen zusammenzufassen (Wottawa, 1987b, S. 101), stoßen nach Dörners Auffassung angesichts der Vielzahl der Variablen(interaktionen) rasch an ihre Grenzen. Dies führt seiner Ansicht nach zu der Notwendigkeit einer - nicht näher spezifizierten - Art der konstellativen Analyse der Verhaltensprotokolle, die sich - wie der Autor selbst einräumt (Dörner 1992, S. 85) - nicht so einfach automatisieren läßt. Man kann auch sagen: die sich nicht (einfach) standardisieren läßt. Erfolgt die Informationsgewinnung aber nicht mehr nach standardisierten Vorschriften, so ist eine der wesentlichen Voraussetzungen diagnostischer Instrumente verletzt: Diagnostische Instrumente sind an die Bedingung geknüpft, *„daß die Messung der in Frage stehenden psychischen Sachverhalte auf eindeutigen, meßtheoretisch fundierten Regeln basiert, nach denen den Informations-*

stichproben der Pbn Meßwerte zugeordnet werden können." Diese Voraussetzung, die Michel und Conrad (1982, S. 2) hier für psychometrische Tests ausführen, gelten auch für computergestützte Problemlöseszenarien, die als diagnostische Instrumente verwandt werden. Es ist daher nur konsequent, wenn Mitglieder der Bamberger Gruppe sich in jüngeren Veröffentlichungen von dem früher eindeutig erhobenen diagnostischen Anspruch (siehe Einführung zu Kapitel 3) distanzieren (z.B. Dörner, 1992, S. 84f.) und den Einsatz von computergestützten Problemlöseszenarien für diagnostische Zwecke nun *„ausgesprochen kritisch"* betrachten (Strohschneider & Schaub, 1995, S.201). Mit dem Verzicht auf den früher geäußerten Anspruch auf eine *„Diagnostik der operativen Intelligenz"* (Dörner, 1986), büßt das Forschungsprogramm allerdings etwas Attraktivität ein, schließlich verdankte die Problemlöseforschung - wie in Kapitel 3 gezeigt wurde - ihren anfänglichen Aufschwung sicherlich auch der theoretischen Konfrontation mit der Intelligenzforschung und deren angeblichen diagnostischen Defiziten. Die nun hilfsweise propagierte kontextspezifische individuum-zentrierte diagnostische Perspektive und der ihr inhärente methodologische Individualismus sollten von Anfang an vor dem Hintergrund der in der Diagnostik häufigen Situation der Konkurrenzentscheidung problematisiert werden. Diagnostik bedeutet auf eine Fragestellung bezogene Erhebung von Informationen; die Entscheidung, bei der der Diagnostiker hilfreich sein soll, erfordert u.a. *Vergleichswissen*. Es genügt nicht, das Verhalten einer Person zu beschreiben, sondern es geht auch darum, *„das Verhalten einer Person im interindividuellen Vergleich zu beurteilen"* (Westmeyer, 1976, S. 75). Für die diagnostische Urteilsbildung, für die Entscheidung, werden Normen im Sinne von Präskriptoren benötigt, die sich entweder auf regelbasierte oder auf empirische Theorien stützen, assoziative Verknüpfungen phänomenologisch beschriebener Verhaltenskonstellationen in der »Mikrowelt« mit möglichem Verhalten in der „realen" Welt reichen nicht aus. Die in jüngeren Arbeiten geforderte kontextspezifische individuum-zentrierte diagnostische Nutzung von computergestützten Problemlöseszenarien (siehe auch Abschnitt 6.4) ist nicht Gegenstand der vorliegenden Arbeit, diesbezüglich sollte aber kritisch geprüft werden, ob es sich nicht um den Beginn einer Echternacher Springprozession handelt.

5.3 Begründung einiger für den diagnostischen Einsatz notwendigen Abweichungen von den Regeln der „Philosophie der Verwendung von Mikrowelten“

Die Angemessenheit von Methoden wie der Experimentalmethode oder der im Kontext der Philosophie der Verwendung von Mikrowelten propagierten Methode der „lateralen Validierung“ kann nicht für sich genommen beurteilt werden, sondern muß vor dem Hintergrund des jeweiligen Forschungsproblems betrachtet werden. Die Anwendung von Methoden dient der Erreichung von Forschungs(teil-)zielen, unterschiedliche Ziele erfordern unterschiedliche Methoden. Herrmann (1995) betrachtet Forschung als Problemlösen und typisiert Forschungsprobleme u.a. nach der Klarheit der Zielvorgabe (siehe oben, Abschnitt 2.1). Bei der Entwicklung und Evaluation eines diagnostischen Instruments wie z.B. einem Test ist das Ziel, der Soll-Zustand, klar definiert. Es geht um die Effizienz- und Rationalitätssteigerung der nicht-forschenden Handlung des Diagnostizierens und somit letztendlich um ein *technologisches Problem*. Davon abgrenzbar sieht Herrmann *wissenschaftliche* oder auch *grundlagenwissenschaftliche Probleme* wie die Entwicklung und Prüfung einer kognitionswissenschaftlichen Theorie als ein Problem ohne Klarheit der Zielkriterien. (Siehe aber die Einwände Westmeyers (1993) gegen die konzeptionelle Trennung von grundwissenschaftlichen und technologischen Theorien.) Während die grundlagenwissenschaftliche Forschung das Problem u.a. rekonstruiert und das Auftreten bestimmter Ereignisse erklärt (und vorhersagt), stellen technologische Forschungsprogramme operatives Hintergrundwissen und standardisierte Techniken (im Sinne normierter Handlungsanweisungen) zur Effizienzsteigerung bereit. Technologische Regeln haben keine Wahrheitswerte, sondern Effektivitätswerte (Westmeyer, 1976, S. 79). Technologische Lösungen können effizient sein, ohne das man weiß, *wieso* das so ist (siehe dazu auch Jäger, 1970, S. 621) - man hat auch das Thermometer vor der Theorie der Thermodynamik zu nutzen gewußt und den Wirkmechanismus von Acetylalicylsäure erst 70 Jahre nach der erfolgreichen Einführung des Arzneimittels „Aspirin“ erforscht. Demgegenüber sind theoretische Untersuchungen geradezu „verpflichtet“, zu erklären, *wie* etwas zustande kommt. Während der Erfolg der Lösung technologischer Probleme oder pragmatischer Fragen von deren *Nutzen* bestimmt wird, hängt der Erfolg der Lösung (grund-)wissenschaftlicher Probleme und theoretischer Fragen von deren Fähigkeit zur Erklärung und Vorhersage ab. (Vgl. auch die Unterscheidung von pragmatisch und theoretisch orientierten Untersuchungen bei Meehl 1978, S. 823 ff.). Insbesondere für die Berufseignungsdiagnostik können die Regeln der »Philosophie der Verwendung von Mikrowelten« aufgrund der unterschiedlichen theoretischen Vorannahmen und der unterschiedlichen Zielsetzungen des Instrumenteneinsatzes nicht akzeptiert werden. Während

Dörner's Ansatz auf den Einzel- und Sonderfall von Verhalten als Produkt einer einzigartigen Konstellation motivationaler, emotionaler und kognitiver Faktoren zielt, zielt der eignungsdiagnostische Ansatz auf die Messung (zumindest partiell) konstanter, individuell unterschiedlich ausgeprägter Personmerkmale, die (zumindest partiell) konstant mit beruflichem Erfolg kovariieren. Situationsspezifische Ausprägungen allgemeiner Verhaltensformen (z.B. Entscheiden, Schlußfolgern, Urteilen), welche nur unter nicht generalisierbaren situationsspezifischen Randfaktoren (wie z.B. spezifischen Hypothesen, Planungen, extreme emotionale Zustände usw.) auftreten, sind für die Eignungsdiagnostik im Regelfall unbedeutend.

Herrmann (1995) betont, daß die angestrebten Problemlösungen in der Praxis *funktionieren* müssen, also auch verläßlich, nebenwirkungsfrei, routinisierbar und nicht zuletzt wirtschaftlich sein müssen. Hinsichtlich der diagnostischen Verwendung computergestützter Problemlöseszenarien ist es bedeutend, sich mit Herrmann (ebd., S. 32) zu vergegenwärtigen, daß Technologie nicht dasselbe ist wie angewandte Wissenschaft. *„Vielmehr werden im Kontext technologischer Problemlösungsprozesse (grundlagen-)wissenschaftliche Problemlösungsresultate genutzt, indem man sie aus ihrem wissenschaftsimmanenten Zusammenhang löst, sie für den technologischen Zweck selektiert und entsprechend aufbereitet."* Computergestützte Problemlöseszenarien können für diagnostische Zwecke genutzt werden, wenn sie diesem Zweck entsprechend aufbereitet werden, d.h. vor allem, wenn die Untersuchungsbedingungen sowie die Auswertung und Interpretation der Daten standardisiert werden. Es ist nicht nur unproblematisch, sondern geradezu sinnfällig, daß bei dem technologischen Problem der Entwicklung und Evaluation diagnostisch genutzter computergestützter Problemlöseszenarien andere Methoden zum Einsatz kommen als bei dem (grund-)wissenschaftlichen Problem, mit Hilfe der »Mikrowelten« *„etwas darüber in Erfahrung zu bringen, wie die menschliche Seele Verhalten hervorbringt"* (Dörner, 1992, S. 63).

Die vorliegende Arbeit beurteilt nicht die Frage, ob die experimentelle Methode grundsätzlich dem Fortschritt in der Problemlöseforschung abträglich ist (siehe z.B. die Argumente zugunsten der Experimentiermethode von Funke, 1995a) und ob die Strategie der „lateralen Validierung" der Lösung der (grund-)wissenschaftlichen Probleme dienlich ist. Für die vorliegende Fragestellung ist eine andere Überlegung bedeutsam: Selbst wenn man hilfsweise davon ausgeht, daß der den Regeln der »Philosophie der Mikrowelten« gerechte Einsatz von computergestützten Problemlöseszenarien zu beobachtbarem Verhalten führt, welches besser mit Theorien des menschlichen Denkens in Übereinstimmung gebracht werden kann als das Verhalten beim Bearbeiten von Intelligenztestaufgaben, so geht daraus allein noch nicht hervor, daß diese Szenarien sich besser als Grundlage praktischer Entscheidungen eignen als andere Verhaltensmessungen. *„Theoretical interpretation alone is not a*

sufficient reason for using a test. A test that is used to make social decisions must meet traditional psychometric criteria for reliability and validity." (Hunt, 1983, S. 146). Wichtig für die vorliegende Fragestellung ist die Annahme, daß unterschiedliche Ziele häufig mit unterschiedlichen Methoden erreicht werden. Stern hob bereits hervor, daß der Test als Spezialfall des Experiments keine rein wissenschaftlich-theoretische Aufgabe verfolgt, sondern eine diagnostische. Der Test (und beim diagnostischen Einsatz von Problemlöseszenarien werden auch diese zu Tests) *„will nicht unbekannte Gesetze und neue Zusammenhänge erforschen, sondern die Einordnung eines Einzelfalls in einen bereits bekannten Zusammenhang vollziehen"* (Stern, 1911, S. 87). Die Regeln der »Philosophie der Mikrowelten« betreffen den heuristischen, hypothesengenerierenden Gebrauch von Problemlöseszenarien im Rahmen einer Forschungsstrategie und können nicht auf diagnostische Fälle angewendet werden, bei denen mit Hilfe von Problemlöseszenarien Daten zur interindividuellen Unterscheidbarkeit und Vergleichbarkeit von Personen auf vorab definierten Kategorien gewonnen werden sollen. Dieses zuletzt genannte Vorhaben setzt voraus, daß die computergestützten Problemlöseszenarien für den technologischen Zweck der Diagnose entsprechend aufbereitet werden. In dieser „aufbereiteten Form" mögen die Szenarien gegen die für andere Referenzsysteme und für andere Zielsetzungen formulierte »Philosophie der Verwendung von Mikrowelten« verstoßen. Eine solche Diskrepanz zwischen Instrumenten, die mit unterschiedlichen Zielsetzungen eingesetzt werden, ist plausibler als das Ansinnen, mit Szenarien, die der für (grund-)wissenschaftliche Probleme formulierten »Philosophie der Verwendung von Mikrowelten« entsprechen, *gleichzeitig* technologische Fragestellungen beantworten zu wollen. Wer ein technologisches Problem wie das Diagnose-Problem lösen will, hat sich pragmatischen Fragen wie der Frage nach dem Funktionieren in der Praxis und der Frage nach dem Nutzen für die Praxis zu stellen und kann sich nicht hinter dem Paravent »Philosophie der Verwendung von Mikrowelten« verstecken, welcher vielleicht für andere Schlachten taugen mag. Zwar ist es für bestimmte Fragestellungen denkbar, daß ein Instrument unterschiedliche Ziele gut erfüllt, im konkreten Fall sind bestimmte Vorgaben der „Philosophie der Mikrowelten" mit dem diagnostischen Zielen aber unvereinbar. So verstößt man beispielsweise durch die Verwendung objektiver Problemlösegütemaße, durch das Ausschalten unkontrollierbaren Vorwissens, durch die Vorgabe klarer Zielvorgaben und durch die Reduktion der Komplexität zur Vereinfachung der Systeme nach Ansicht Dörners (1992, S. 62f) bereits gegen die Regeln der »Mikrowelt-Philosophie«. Zur Leistungsdiagnose eingesetzte Szenarien *müssen* hingegen z.B. mit objektivierbaren Problemlösegütemaßen und mit eindeutigen Zielen vorgegeben werden. An dem Erreichen dieses vorgegebenen Zieles wird der Erfolg bemessen, nur so ist es diagnostisch angemessen - an welchem Kriterium sollte man die Fähigkeit oder Leistung

des Individuums sonst vergleichend messen? Wie wollte man die anhand des Verhaltens bei der Steuerung des computergestützten Problemlöseszenarios getroffene diagnostische Entscheidung sonst vermitteln, d.h. transparent gestalten? Die Präzision bzw. Bestimmtheit der vorgegebenen Ziele ist mit Jäger (1991, S. 289) die „conditio sine qua non". *„Vage Zielvorgaben erlauben die Untersuchung individueller Zielfindungs- und Absichtsregulationsprozesse, aber keine Leistungsvergleiche zwischen Probanden".*

Die Vorgabe eindeutiger Zielvorgaben, die Darbietung steuerbarer und somit nicht übermäßig komplexer Probleme, der Verzicht auf „Überraschungseffekte" und Zufallseinflüsse, die Schaffung von Transparenz hinsichtlich der Handlungsmöglichkeiten und die Etablierung von objektiv und standardisiert bestimmbaren, reliablen Problemlösegütemaßen, die sich unmittelbar an den Zielvorgaben orientieren - all dies sind einige Aspekte, die berücksichtigt werden müssen, falls man computergestützte Problemlöseszenarien für diagnostische Zwecke einsetzen will. Einzelne Abschnitte werden in den folgenden Kapiteln noch vertieft, weitere Abschnitte ergänzt. Diese Vorgaben weichen teilweise von der »Philosophie der Verwendung von Mikrowelten« ab - z.B. die Vorgabe des Steuerungsziels. Durch diese, in Abweichung von der »Philosophie der Mikrowelten« vorgenommenen Änderungen der Aufgabenstellung ändert sich das beobachtbare Verhalten, insbesondere wird die Messung eindeutiger dem Bereich der Fähigkeiten zugeordnet, während die Personeigenschaften „Motivation" und „Temperament" an Einfluß auf das so gemessene Problemlöseverhalten verlieren können. Es bleibt den Protagonisten der »Philosophie der Mikrowelten« unbenommen, dem derart restriktiv erfassten Verhalten die Zuordnung zum Konstrukt „Problemlösen" abzusprechen, in dem gerade die auf der Meßebene vorgenommene Kontamination verschiedener Klasse von Personeigenschaften zum definierenden Element des Problemlöseverhaltens stilisiert wird. Als Maßstab für die Bewertung der unterschiedlichen Ansätze aus diagnostischer Perspektive gilt u.a. der Nachweis erfolgreicher Anwendungen. Der bloßen Behauptung eines prinzipiellen Vorteils des „Mikrowelt"-Ansatzes kann in diesem Zusammenhang keine Bedeutung zukommen.

Die Berücksichtigung der hier beispielhaft genannten Rahmenbedingungen der Szenarienverwendung allein ist selbstverständlich noch keine hinreichende Begründung für einen diagnostischen Einsatz dieser Instrumente, diesbezüglich kommt dem Validitätsnachweis eine zentrale Rolle zu (siehe Kapitel 9).

5.4 Zusammenfassung, Schlußfolgerungen und Ausblick

Dörner verbindet mit computergestützten Problemlöseszenarien eine »Philosophie der Verwendung von Mikrowelten«, die sich als Kritik der und Alternative zur Experimentalmethodik versteht. In der vorliegenden Arbeit wird die Frage ausgespart, ob diese „Philosophie" für die Entwicklung und Prüfung kognitionswissenschaftlicher Theorien dienlich ist. Beim Einsatz von computergestützten Problemlöseszenarien zur Fähigkeitsdiagnostik geht es um die Effizienz- und Rationalitätssteigerung bei den mit Hilfe dieser Instrumente getroffenen Entscheidungen und somit letztendlich um ein ***technologisches Problem***. Forscher, die mit computergestützten Problemlöseszenarien zur Lösung dieses Problems beitragen wollen, müssen nicht der für (grund-)wissenschaftliche Fragestellungen entworfenen Mikrowelten-Philosophie gehorchen, sondern sie müssen Problemlösungen vorlegen, die in der Praxis nachweislich *funktionieren*. Computergestützte Problemlöseszenarien sind für den diagnostischen Zweck aufzubereiten, indem alle Voraussetzungen für eine *standardisierte* Anwendung und Auswertung geschaffen werden. Dazu sind eine Reihe von Vorkehrungen nötig, die auszugsweise genannt wurden. Eine vervollständigte Liste findet sich im Diskussionsteil (Kapitel 18) am Ende des Buches. Unter anderem ist es notwendig, mit eindeutigen Zielvorgaben zu operieren. Wichtig sind außerdem reliable Problemlösegütemaße, die sich unmittelbar an den Zielvorgaben orientieren und - im Falle von mehreren Problemlösegüteindikatoren - Regeln zur konfiguralen oder integrativen Bewertung erlauben. Die hier genannten Anpassungen computergestützter Problemlöseszenarien an den diagnostischen Zweck verstoßen mit Bedacht gegen die Dörnersche »Philosophie der Mikrowelten«.

In den Kapiteln acht und neun werden Befunde zur Reliablilität und Validität von Problemlösegütemaßen und Strategieindikatoren referiert, zunächst sollen im Kapitel sechs aber verschiedene Möglichkeiten zur Bildung von solchen Indikatoren vorgestellt und diskutiert werden. Kapitel sieben widmet sich der Bedeutung der Steuerbarkeit einzelner computergestützter Problemlöseszenarien und der Bedeutung dieses Aspekts für die Diagnose.

6. Problemlösegütemaße

Wo rohe Kräfte sinnlos walten,/Da kann sich kein Gebild gestalten.
FRIEDRICH SCHILLER; Das Lied von der Glocke, V 350f.; 1799

Der Einsatz von computergestützten Problemlöseszenarien als diagnostische Instrumente zielt darauf ab, Differenzen zwischen Individuen bzw. individuellen Ausprägungen von Personvariablen zu *messen*. Messen bedeutet, daß das interessierende empirische System (z.B. die Problemlösefähigkeit) durch ein numerisches System, welches einfache Relationen enthält, homomorph abgebildet (repräsentiert) wird (siehe Gigerenzer, 1981, S. 45 und S. 47).

Einer der bislang ungelösten neuralgischen Punkte bei der Verwendung von computergestützten Problemlöseszenarien besteht darin, daß die abgeleiteten Indikatoren des Problemlöseerfolgs den Anforderungen an eine Messung (Vollständigkeit, Transitivität, Eindeutigkeit und Bedeutsamkeit) häufig nicht genügen. Allein die Vielzahl der existierenden Operationalisierungen/Indikatoren des Problemlöseerfolgs (für eine Übersicht siehe z.B. Funke, 1986) weist auf den Interpretationsfreiraum in der Zuordnungs- und Bewertungsfrage hin. Die - gelegentlich auch ex post facto vorgenommene - Ableitung der Maße des Problemlöseerfolgs erscheint nicht selten arbiträr, insbesondere „Strategien" werden häufig nicht diagnostiziert, sondern kasuistisch retrognostiziert. Die Bedeutung der Schwierigkeiten mit den Problemlösegütemaßen kann nicht hoch genug angesiedelt werden. Hussy hat bereits 1985 (S. 59) darauf hingewiesen, daß die Problemlöseforschung mit der Frage der Operationalisierung der Problemlösegüte steht und fällt (siehe auch Kluwe, 1990a, S. 244). Einige inkonsistente und unplausible empirische Ergebnisse im Kontext der Problemlöseforschung sind möglicherweise schlicht darin begründet, daß das Problem der Messung - welches jeder statistischen Analyse und inhaltlichen Interpretation vorausgeht - vernachlässigt wurde. Im Abschnitt 7.3 wird am Beispiel einer älteren Version der „Schneiderwerkstatt" aufgezeigt, welche Konsequenzen die Verwendung intern nicht-valider Problemlösegütemaße nach sich ziehen kann. Die Vielzahl unterschiedlichster Problemlösegütemaße führt außerdem dazu, daß unter dem Label „Problemlöseforschung" de facto eine Vielzahl unterschiedlichster und unvergleichbarer Ansätze subsumiert werden. Einem weit verbreiteten Mißverhältnis in der psychologischen Forschung folgend, wird auch in der Problemlöseforschung *„über die Frage, wie man zu Zahlen kommt, d.h. über das Meßproblem (...), meist schnell hinweggehuscht"* (Gigerenzer, 1988, S. 98 in Bezug auf die Illusion, daß Statistik der wichtigste Teil der wissenschaftlichen Methodik sei), um sich dann umso intensiver mit der Analyse und Interpretation dieser möglicherweise ungültigen Zahlen zu beschäftigen. Es ist alarmierend, daß aus ein und derselben Problembearbeitung - je nachdem, welcher Beurteilungsmaßstab mit jeweils plausibler Begründung heran-

gezogen wird - widersprüchliche Beurteilungen der Problemlösefähigkeit abgeleitet werden können (siehe die Beispiele zur Dissoziation zwischen Verhaltensmaßen und Maßen der Steuerungsleistung in Abschnitt 6.3.2.3). Oft werden auch parallel eine Vielzahl unterschiedlicher und teilweise widersprüchlicher Indikatoren berechnet. Auch wenn sich Laien möglicherweise durch den Hinweis auf den Umfang der verfügbaren Datenmassen beeindrucken lassen: ein Konglomerat von (widersprüchlichen) Daten erleichtert die zu treffende diagnostische Entscheidung nicht, sofern keine Regeln zur Integration der gewonnenen Daten zu einem diagnostischen Urteil angegeben werden. Häufig können lediglich die Programmierer des Problemlöseszenarios sowie wenige Experten, nicht aber die diagnostischen Anwender oder die Problemlöser selbst, nachvollziehen wie die einzelnen Indizes zustande kommen. Nicht selten fehlt jegliche Begründung der jeweils gewählten Bewertungsvariante. Hinsichtlich der - weiter unten erläuterten - Verhaltens- oder Strategiemaße sind auch die in Fachveröffentlichungen getroffenen Angaben selten so präzise, daß eine Replikation der Bewertung des Problemlöseerfolgs möglich wäre (siehe z.B. das weiter unten dargestellte Beispiel von Jansson, 1994).

U. Funke (1995a, S. 149) unterscheidet drei Ebenen der Ergebnisquantifizierung: Die Ebene der Steuerungsleistungen, die kognitive Ebene (z.B. Wissensstrukturen) und die Ebene der Verhaltensmaße. Diese Dreiteilung wird im folgenden aufgegriffen und weiter differenziert (siehe Abbildung 2). Insbesondere die Verhaltensmaße werden kritisch diskutiert, wobei die theoretischen Defizite und die Varianz in Ableitung und Bewertung von Verhaltenweisen herausgearbeitet werden. In einem weiteren Punkt wird darauf aufmerksam gemacht, daß die Beurteilung der Verhaltensweisen häufig im Widerspruch zu der Beurteilung der eigentlichen (instruktionsgemäßen) Problemlösung - wie sie sich anhand der Steuerungsleistung darstellt - steht. Abschließend werden Überlegungen zur Auswahl eines für die Fähigkeitsdiagnostik adäquaten Problemlösegütemaßes angestellt, die in einem Plädoyer für das Primat von Steuerungsleistungen enden.

Das bei allen möglichen Problemlösegütemaßen unbedingt zu klärende Problem der Reliabilität wird in Kapitel 8 gesondert thematisiert. Bei allen folgenden Ausführungen wird vorausgesetzt, daß der Systemsteuerung ein klares Ziel vorgegeben wurde. Bei „ill-defined problems" liegt die Zielspezifizierung und somit das Urteil über die Zielerreichung und die Lösungsgüte zum Teil bei dem Diagnostikanden und nicht beim Diagnostiker. Offene Problemstellungen sind daher nur in Ausnahmefällen - z.B. bei der Diagnose von Zielbildungsprozessen - für die Diagnostik interessant.

Weiterhin wird davon ausgegangen, daß die Diagnostikanden direkt mit dem System interagieren. Bei einer über einen Versuchsleiter vermittelten Interaktion ist die (Durchführungs-) Objektivität bei der Erhebung der Daten schwer zu gewährleisten.

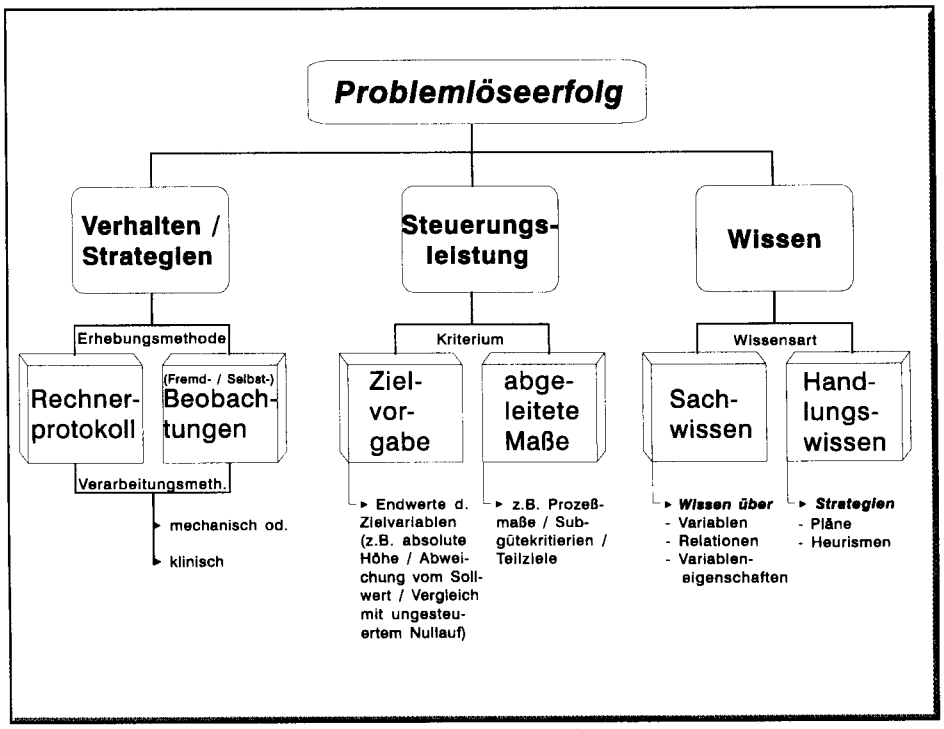


Abb. 2: Operationalisierungen des Problemlöseerfolgs

6.1 Steuerungsleistungen

Bei der Quantifizierung der Problemlösegüte über Indikatoren der Steuerungsleistung läßt sich zunächst unterscheiden, ob sich die Bewertung (a) direkt und unmittelbar an der Erreichung des vorgegebenen Ziels orientiert oder ob man (b) abgeleitete Maße benutzt, die mehr oder weniger eng an den Zielvariablen orientiert sind. Orientiert sich die Bestimmung des Problemlöseerfolgs direkt und unmittelbar an der Erreichung des vorgegebenen Ziels, so wird zur Definition des Problemlöseerfolgs häufig die Ausprägung der Zielvariable(n) (z.B. das Gesamtvermögen in einem betriebswirtschaftlich eingekleideten Szenario) herangezogen, oder es werden Abweichungsmaße (Vergleiche mit Optimallösungen [falls bestimmbar], mit vorgegebenen Sollwertgrößen oder mit ungesteuerten „Nulläufen" vorgenommen). Auswertungsvariationen entstehen auch im Zusammenhang mit dem Prozeßcharakters der Bearbeitung, betrachtet werden kann entweder die Endabweichung oder die durchschnittliche Abweichung von der Zielvariablen, gelegentlich werden auch einzelne Bearbeitungszeitpunkte der Bestimmung der Problemlösegüte zugrundegelegt.

Schwierigkeiten können sich unter anderem dadurch ergeben, daß der Wertebereich der Zielvariablen uneingeschränkt variieren kann. (So kann beispielsweise das Gesamtvermögen als Zielvariable betriebswirtschaftlich eingekleideter Problemlöseszenarien extreme postive und negative Werte annehmen.) Sofern keine weiteren Transformationen vorgenommen werden, können die dadurch häufig entstehenden Ausreißerwerte die mit parametrischen Verfahren durchgeführten Analysen verzerren, falls - wie zumeist - Aggregatshypothesen geprüft werden. Außerdem wird mit der untransformierten Verwendung einer Zielvariablen mit uneingeschränktem Wertebereich die Wahrscheinlichkeit einer Messung auf Intervallskalenniveau geringer. Bei mehreren Zielvariablen stellt sich außerdem das Problem der Kombination der resultierenden Werte.

Als Beispiel für abgeleitete Maße können „Trendmaße" aufgeführt werden, etwa die Anzahl der Takte mit einem Aufwärtstrend in der Zielvariablen. Die Trends werden entweder über die gesamte Bearbeitungszeit oder aber auch über (mehr oder minder willkürlich) ausgewählte Phasen berechnet. Ein anderes Beispiel sind Indikatoren, die als Distanz zwischen dem realisierten Eingriff und demjenigen Eingriff bestimmt werden, welcher in dem jeweiligen Systemzustand optimal gewesen wäre[7]. Gelegentlich wird auch noch der weitere Systemverlauf (ohne neue Eingriffe des Problemlösers) getestet, um die „Stabilität" des Systemzustandes zu bestimmen. Praktiziert wurde und wird auch eine „Relativierung" der Werte für die Steuerungsleistung am Null-Lauf (Ablauf des Szenarios ohne Eingriffe).

Insgesamt ist die Verwendung von Indikatoren der Steuerungsleistung, die sich nur „indirekt" an der Zielvorgabe orientieren (abgeleitete Maße) aus diagnostischer Perspektive aufgrund der mangelnden Transparenz für den Diagnostikanden und den Diagnostiker zu kritisieren (siehe z.B. Kluwe et al, 1991c, S. 298). So *interpretiert* z.B. das für die „Schneiderwerkstatt" und verwandte Systeme häufig verwendete Trendmaß der Anzahl an Monaten mit Aufwärtstrend im Flüssigkapital („Trendpo") oder im Gesamtkapital („Trendfu") z.B. die Gewinnmaximierung als einen notwendig „stetigen" Prozeß. Versuchspersonen, die während der Bearbeitung kostenintensive Investitionen tätigen, die ihnen zum Ende der Bearbeitungszeit ein instruktionsgemäß hohes Gesamtvermögen verschaffen, werden unter Umständen aufgrund der abgeleiteten Maße relativ negativ beurteilt - obwohl ihnen ein stetiger Anstieg der Zielvariable nicht abverlangt wurde. Unter Umständen kann die Beurteilung einer solchen Versuchsperson negativer ausfallen als die Beurteilung einer vom Gesamtkapitalendwert her gesehen deutlich schlechteren Versuchspersonen, die sich nach einem anfänglich ruinösen Start kontinuierlich auf niedrigem Niveau verbessert hat.

---

[7] Zum Auffinden optimaler Referenzpunkten können z.B. ableitungsfreie Suchverfahren des Operation Research angewendet werden (Kolb, Petzing & Stumpf, 1992).

Da man einen ungünstigen Anfangswert - von dem aus man sich dann ohne große Schwierigkeiten sukzessive verbessern kann - leicht willkürlich herbeiführen kann, wird durch die Trendmaße auch das Problem der Verfälschbarkeit (siehe Abschnitt 10.2) abgeleiteter Maße veranschaulicht.

In der Regel werden die hier beschriebenen Indikatoren der Steuerungsleistung „mechanisch" („automatisch") verarbeitet, d.h. sie ergeben sich aus einer formalisierten Analyse der Rechnerprotokolle. Es wurden allerdings auch „Experten-Ratings" von Variablenausprägungen, -konstellationen und -verläufen vorgenommen (z.B. Dörner et al., 1983b; Putz-Osterloh, 1981; Reichert & Dörner, 1988). Zu den Schwierigkeiten der Interpretation von Systemverläufen siehe z.B. Müller (1993).

6.2 Kognitive Ebene/Wissen

Nicht selten wird - überwiegend *zusätzlich* zu den Steuerungsleistungen - auch der Grad des Wissens über Variablen, Variablenrelationen und über Variableneigenschaften sowie das Wissen über erfolgreiche Strategien und Heurismen als ein Indikator für die Problembewältigung herangezogen (z.B. Funke & Müller, 1988; Hörmann & Thomas, 1989, S. 27). Dabei ergibt sich - ähnlich wie bei dem weiter unten in Abschnitt 6.3.2.3 ausführlicher beschriebenen Problem des fehlenden Zusammenhangs zwischen Verhaltensmaßen und dem Steuerungserfolg - die Besonderheit, daß eine Bewertung des Wissens *als Indikator für die Problemlösefähigkeit* diagnostisch nur dann vertretbar erscheint, wenn das Ausmaß an Wissen tatsächlich den Problemlöseerfolg bestimmt. Ist dies nicht der Fall, läßt sich die Bewertung des Wissens im Zusammenhang mit der Problembearbeitung unter Transparenzgesichtspunkten nur dann rechtfertigen, falls - wie z.B. bei dem System „DISKo" von U. Funke (1992) - die Erarbeitung von Wissen über das System explizit als Ziel der Systemsteuerung vorgegeben wurde. Auch unter dieser Voraussetzung kommen allerdings in der Behandlung der Wissensmaße als Indikatoren der Problemlösefähigkeit weitreichende theoretische Annahmen zum Ausdruck (siehe Abschnitt 9.1.3), die zumindest expliziert werden sollten. Neben den jeweils diagnostizierten Wissensarten, die im Abschnitt Kapitel 9.1.3 ausführlicher dargestellt werden, unterscheiden sich auch die zur Wissensdiagnose eingesetzten Instrumente (siehe Abschnitt 9.1.3.2).

6.3 Verhaltensmaße

In einer dritten Herangehensweise wird der Problemlöseerfolg und das Ausmaß der Problemlösefähigkeit aufgrund von „Verhaltensmaßen" bestimmt. Grundlage dieser Beurteilungen sind entweder die im Rechnerprotokoll verzeichneten Eingriffe sowie die Kennwerte der Variablen über die gesamte Zeit der Steuerung und/oder Fremd- oder Selbstbeobachtungen des Verhaltens während der Systemsteuerung. Zur Selbstbeobachtung wird häufig die Methode des lauten Denkens angewandt. Ausgewertet werden die protokollierten, beobachteten oder selbstberichteten Daten entweder mechanisch oder klinisch. Dabei werden relative und absolute Häufigkeiten von Verhaltenskategorien, ihre Verteilung, Veränderung über die Zeit, ihre Übergangswahrscheinlichkeiten usw. ausgezählt und bewertet. Die zu beobachtenden Verhaltenskategorien und die Bewertungsregeln sollten selbstverständlich *vorab* definiert werden. Sofern das Verhalten automatisch registriert und die so erhobenen Daten mechanisch verarbeitet werden, ist diese Bedingung der Vorab-Definition erfüllt. So müssen in einigen Programmen (z.B. in dem Programm „DISKo" von U. Funke, 1992a) bestimmte - vorab als relevante Verhaltensweisen definierte - „Aktionen" - wie z.B. das Einholen von Informationen oder das Prognostizieren von Systementwicklungen - gesondert aufgerufen werden (vgl. die Technik der „withheld information", z.B. Marshall, Duncan und Baker, 1981).

Gerade auf die Auswertung und Interpretation von Verhaltensmaßen und Strategien gründet sich der Anspruch auf Überlegenheit der mit Hilfe computergestützter Problemlöseszenarien gestellten Diagnosen (siehe Abschnitt 3.2), erlauben diese Instrumente es doch vorgeblich, Prozeßkomponenten zu diagnostizieren, die für das Zustandekommen des erfolgreichen Problemlöseverhaltens verantwortlich sind. Demgegenüber haftet der „Endproduktdiagnostik" bei herkömmlichen Leistungstests der Vorwurf der bloßen Beschreibung an. Ein Blick in die Literatur zeigt aber, daß de facto unterhalb der Ebene eines pauschalen Plädoyers keinerlei Konsens über die praktische Vorgehensweise einer Prozeßdiagnostik mit Hilfe computergestützter Problemlöseszenarien existiert. Weder die Definition von Verhaltenskategorien noch die Bewertung von Verhaltensmaßen bzw. Strategieindikatoren ist einheitlich. Welche Verhaltenskategorien beobachtet werden und wie man von den Verhaltensbeobachtungen zu „Strategiekennwerten" gelangt, variiert nicht nur von Problemlöseszenario zu Problemlöseszenario, sondern auch von Autor zu Autor und nicht selten auch von Veröffentlichung zu Veröffentlichung bei ein und demselben Autor. Eine systematische Abhängigkeit der jeweils angewendeten Verhaltensbeobachtung und -bewertung von der jeweiligen Fragestellung ist dabei nicht zu erkennen.

Schon wenige Beispiele mögen genügen, um einen Eindruck von der Vielzahl möglicher Verhaltensmaße zu gewinnen. So schlägt z.B. Dörner (1986) zur Operationalisierung der operativen Intelligenz u. a. die folgenden Verhaltensindizes vor: Häufigkeit sowie Art und Gegenstand der Fragen, Gegenstand der Entscheidungen, Dosierung der Entscheidungen, Bündelung von Fragen und Entscheidungen, Stabilität des Gegenstandes von Fragen und Entscheidungen über mehrere Takte. Uwe Funke (1991) mißt als „Strategiemerkmale" u.a. die Häufigkeit der Informationsabfragen und die Anzahl explorierender Tests. Von anderen Autoren (z.B. Strohschneider, 1986, S. 44f.) wird zusätzlich oder alternativ der Bedarf an Hintergrundinformationen, die Gesamtzahl an Maßnahmen, die Zahl unterschiedlicher Maßnahmen, die Übergangswahrscheinlichkeit Maßnahme-Maßnahme und/oder Frage-Frage erhoben und verwertet, wiederum an anderer Stelle werden u.a. die „Reaktionszeiten für Entscheidungen" als Verhaltensmerkmale erfaßt usw. (Dörner & Preußler, 1990, S. 209). Die hier genannten Verhaltenskategorien zählen, einer Unterscheidung von Schmuck und Strohschneider (1995) zufolge, überwiegend zu den quantitativ erfassten „einfachen Verhaltenskategorien". Diese werden von den Autoren gegenüber „Konstellationen von Verhaltenseinheiten" abgegrenzt. Hierzu zählt vermutlich bereits die Verteilung der Arbeitszeit auf die Takte, (z.B. ob man sich für die ersten Takte mehr Zeit läßt als für die letzten usw.) wie sie z.B. von Strohschneider und Schaub (1991, S. 330f.) geprüft wurden, aber auch die (ebd. thematisierte) „Integriertheit" des Vorgehens als zeitliche Distanz zwischen Frage- und Entscheidungsphase. Die Vielfalt möglicher „einfacher Verhaltenskategorien" führt zu einer noch größeren Vielfalt an möglichen Datenintegrationen oder „Konstellationen von Verhaltenseinheiten". Auf zumeist nicht näher beschriebene und begründete Art und Weise werden bestimmte Verhaltensweisen zur *Charakterisierung* des Denkens benutzt. Die Bearbeitung der Problemlöseszenarien ist dann Ausgangspunkt einer Klassifikation. So beschreibt z.B. Dörner (1981) das Verhalten von „Vagabundierern", „Verkapselern" usw., und Strohschneider (1996c) gelangt für das Verhalten der Probanden bei der Steuerung des „Moro"-Szenarios zu einzelnen *Problemlösetypen* wie den „Stabilisator", den „Ausbeuter", den „Zerstörer", den „Verzettler" oder den „Zögerer". Praktiziert wird auch die (oft implizite) Setzung von positiven Strategien oder Verhaltensweisen, um dann einzelnen Diagnostikanden auf diesen Dimensionen Werte zuzuweisen. Auf diese Weise kommt man zu diagnostischen Aussagen über die „Prioritätensetzung" oder „Balancestrategie" einzelner Testanden (Hartung & Schneider, 1995, S. 228). Putz-Osterloh (1989) gewinnt anhand der Protokolldaten Indikatoren für die „Strategievariablen" „Entscheidungsfähigkeit", „Auffassungsgabe", „Urteilsvermögen" und „Organisationsvermögen".

6.3.2 *Probleme der Verhaltensmaße*

Die Verwendung von Verhaltensmaßen erscheint aus mehreren Perspektiven unbefriedigend. Zunächst fehlt eine ausreichende theoretische Definition der Verhaltensmaße oder Strategien sowie eine theoretische Einordnung der so gewonnenen Maße in das bereits vorhandene Netz von Fähigkeits-/Leistungs- und Persönlichkeitsindikatoren. Des weiteren muß konstatiert werden, daß die Bildung von Beobachtungskategorien, die Zuordnung von Beobachtungseinheiten und die Bewertung von Verhaltensweisen weitgehend beliebig erscheint. Schließlich ist es in der diagnostischen Praxis nicht tolerabel, daß die Beurteilung der Verhaltensweisen häufig im Widerspruch zu der (offensichtlichen) Beurteilung der eigentlichen, d.h. instruktionsgemäßen Problemlösung - wie sie sich anhand der Steuerungsleistung darstellt - steht. Diese drei problematischen Aspekte werden in den folgenden Abschnitten erläutert.

6.3.2.1 Theoretische Probleme der Verhaltensmaße

Die Verwendung von Verhaltensmaßen/Strategien erfolgt zumeist ohne ausführliche theoretische Begründung und Einbettung. Einzelne Strategien und Verhaltensweisen werden häufig lediglich anhand der vorliegenden Szenarienbearbeitungen beschrieben, und es wird dann behauptet, solche Verhaltensweisen und Strategien würden auch „Problemlöser in der Realität" kennzeichnen. Der Nachweis, daß die positiv ausgewählten Verhaltensweisen in der „Realität" überhaupt vorkommen und in der „Realität" oder aber zumindest in dem Problemlöseszenario tatsächlich zum Problemlöseerfolg führen, wird nur selten geführt. Über die Zeit und die Problemlöseszenarien hinweg werden immer wieder neue Verhaltensdaten induktiv gewonnen, so daß die derart arbeitenden Ansätze über ein deskriptiv-phänomenologisches Stadium nicht hinauskommen. Eine Theorie des Problemlösens fehlt nach wie vor, Dörners „Logik des Mißlingens" verweist - wie Fillbrandt (1992, S. 5 u. 8) betont - vor allem auf das Fehlen der eigentlich vorgeordneten „Logik des Gelingens".

Die bei Problemlösegütemaßen grundsätzlich zu klärende Frage der Geltungsbereiche (Generalität, Universalität, zeitliche Konstanz) gewinnt bei der Verwendung von Verhaltensweisen und Strategien eine besondere theoretische Brisanz. Während Chi (1984, S. 218) beispielsweise die Bereichs-Unabhängigkeit gerade zum definierenden Merkmal von Strategien erhebt, werden in anderen Arbeiten auch problemspezifische Verhaltensweisen als Strategien klassifiziert. So grenzt z.B. Klauer (1995, S. 25) problemspezifische „starke" heuristische Strategien gegenüber den weitgehend bereichsunabhängigen „schwachen" Strategien ab. Zusätzlich zur Frage der Bereichsabhängigkeit der erhobenen und bewerteten Verhaltensweisen müßte auch ihre Situationsabhängigkeit thematisiert werden. Dörner (1989c, S. 138 f.) be-

tont z.B., daß die strategisch bedeutsamen „Großmutterregeln" zwar angeben, *was* zu tun ist, nicht aber *wann* es zu tun ist, es handelt sich um konditionalisierte Regeln, die entsprechend flexibel gehandhabt werden müssen. Nimmt man diesen Aspekt der Situationsspezifität ernst, läßt sich eine Verhaltensweise überhaupt nicht mehr generell, sondern nur noch *konstellativ* als positiv oder negativ bewerten (siehe Abschnitt 6.4). Will man aber Problemlöseszenarien verwenden, um Personen generell als Vertreter bestimmter Strategie- oder Problemlösetypen zu klassifizieren, muß der Nachweis erbracht werden, daß diese Personen die genannten typischen Verhaltensweisen/Strategien auch bei anderen Problemstellungen anwenden (zur Generalität siehe Abschnitt 9.1.1). Dieser Nachweis steht bislang aus.

Ordnungsbedarf besteht auch hinsichtlich der Ähnlichkeit oder Unterschiedlichkeit von Verhaltensstrategien einerseits und (prozeduralem) Wissen über Strategien und Heurismen andererseits. So wird der *Problemlösestil* z.B. bei Andresen und Schmid (1993, S. 7) explizit mit dem *strategischen Wissen* der Probanden gleichgesetzt; Schoppek (1996, S. 25) identifiziert Strategien als Handlungswissen.

All diese theoretischen Probleme wären bei der Verwendung von Verhaltensweisen oder Strategien als Problemlösegütemaße zumindetst zu erörtern. Anstelle von Definitionen der Strategien finden sich - wie Baron (1985, S. 372 f.) in Bezug auf Strategien im Rahmen der Intelligenzforschung bereits beklagte - oft lediglich Erläuterungen anhand von Beispielen. In seiner kurzen forschungshistorischen Einordnung des „Strategieansatzes" und der Prozeßdiagnostik stellt Baron (ebd.) den naheliegenden Bezug zwischen Strategien und kognitiven Stilen her. Kognitive Stile sollen laut Tiedemann (1995, S. 508) *„spezifische Strategien, Operationen und Neigungen in komplexen Problemlöse- und Lernprozessen organisieren und auf diesem Wege den Informationsverarbeitungsprozeß kontrollieren"*. Die Erkenntis der Ähnlichkeit zwischen dem Ansatz der kognitiven Stile und dem Ansatz der Verhaltensmaße/Typen/Strategien in der Problemlöseforschung könnte zur Ernüchterung führen. Tiedemann (ebd, S. 526) zieht das folgende Resümee: *„Die anfänglich enthusiastische Akzeptanz des Kognitiven-Stil-Konstrukts findet in den zurückliegenden Jahrzehnten intensiver empirischer Überprüfung allerdings keine zureichende Rechtfertigung"*, und konstatiert, daß sich kognitive Stile auf der Grundlage ihrer Operationalisierung weitgehend dem Intelligenz-Struktur-Modell zuordnen lassen, da im Rahmen der Forschung zu kognitiven-Stilforschung oft nur Verhaltenskorrelate unterschiedlicher Fähigkeitsdimensionen analysiert werden.

6.3.2.2 Die Beliebigkeit der Ableitung und Bewertung von Verhaltensweisen

Die in Abschnitt 6.3.2.2. konstatierten theoretischen Defizite wirken sich negativ auf die Etablierung von Beobachtungskategorien und Bewertungsmaßstäben aus. Üb-

licherweise dient das System der Beobachtungskategorien als Operationalisierung theoretischer Begriffe. Da keine empirisch fundierte Prozeßtheorie des Problemlösens existiert, können die Verhaltensweisen nicht entsprechend dieser Theorie geordnet werden, den Beobachtungen mangelt es an der notwendigen Zielgerichtetheit. Dies führt dazu, daß sowohl die im Kontext der Bearbeitung von Problemlöseszenarien zur Bestimmung der Problemlösegüte vorgenommene Definition einzelner Verhaltensweisen, Strategien und Typen als auch die Zuordnung einzelner Daten zu diesen definierten Einheiten sowie schließlich die vorgenommene Bewertung bislang nicht den Grad der Standardisierung erreicht haben, der für den diagnostischen Einsatz notwendig ist. Problematisch ist auch die Auswertungsobjektivität, die voraussetzt, daß die Auswertung der Verhaltensindizes unabhängig ist von den auswertenden Personen und/oder eingesetzten Apparaten (siehe hierzu beispielsweise die weiter unten dargestellte Untersuchung von Putz-Osterloh und Köster (1988)).

Bislang wurden kaum Bemühungen unternommen, die Bildung dieser Kategorie von Problemlösegütemaßen der Subjektivität und Spekulation der einzelnen Forscher/ bzw. Programmanbieter zu entziehen. Prüfkriterien für die Definition von Verhaltenskategorien, „Strategien" und „Typen" existieren nicht, diese Frage wird zumeist nicht einmal thematisiert. Die statt dessen anzutreffende Veranschaulichung der Sinnfälligkeit der vorgenommenen Strategie- und Typenbildung durch die Erläuterung von Einzelfällen ist unzureichend. Unbestritten haben die „Strategien" und „Typen" ihr Recht an zahlreichen Einzelfällen, ebenso werden damit aber Fälle erfaßt, an denen sie sich vergreifen. Die vorhandenen methodischen Typisierungsverfahren sind nicht anwendbar, solange es - wie bei den meisten Problemlöseszenarien - keine Ideal-Kombinationen der Verhaltensweisen gibt. Kubinger (1993, S. 136) macht darauf aufmerksam, daß unter diesen Umständen bereits bei sechs dreikategoriellen Variablen 729 Kombinationen resultieren, so daß sowohl die „Konfigurations-Frequenz"-Analyse als auch die „Latent-Class-Analysis" eine unrealistisch hohe Personenzahl zur Identifizierung von Typen benötigen. Der Einsatz der Cluster-Analyse scheitert ebenso, weil diese Methode intervallskalierte Variablen voraussetzt. Nach Ansicht von Kubinger (ebd.) verbleibt lediglich die pro Diagnostikand *„willkürliche (sicher nicht multivariat begründete) Datenreduktion"* durch den Diagnostiker, *„um vermeintlich das wesentliche aus dem komplexen Testverhalten zu abstrahieren - von Verrechnungsfairness ist da nicht zu sprechen, abgesehen davon, daß die inhaltliche Relevanz solcher intuitiven ‚Typisierungen' niemals prognostisch valide werden kann."* Eine präzise Definition von Verhaltensweisen/ Strategien wird bei operierenden Modellen geleistet, die mit einprogrammierten Strategien einzelne Szenarien steuern (siehe z.B. Ringelband et al., 1990). Dieser Ansatz birgt einerseits den Vorteil einer expliziten Beschreibung von Strategien und erlaubt es andererseits, deren Einfluß auf die Steuerungsleistung zu bestimmen.

Selbst wenn man hypothetisch einmal davon ausgeht, die „intuitive" Definition von Strategien oder Typen wäre in der diagnostischen Praxis akzeptabel, ergibt sich im nächsten Schritt das Problem der Zuordnung einzelner Daten zu den Datenkategorien. Sofern die Daten nicht automatisiert zugeordnet werden (siehe hierzu beispielsweise den Ansatz von Heeg und Kleine (1995), anhand von Idealmustern und mit Hilfe neuronaler Netze zu einer automatisierten Verhaltens-Profilbildung zu gelangen) entsteht hinsichtlich dieser Zuordnung eine unerwünschte Varianz. Eine Versuch zu einem System von Beobachtungskategorien zu kommen sowie einige Daten zur Bewährung dieses Ansatzes, stellt Jansson (1994) vor. Sein System von Beobachtungskategorien umfaßt sieben Kategorien, die bestimmte Fehlertypen (vom Autor als „pathologies" bezeichnet) beim Umgang mit Problemlöseszenarien charakterisieren sollen (z.B. „thematisches Vagabundieren"). Positiv hervorzuheben ist, daß in dieser Arbeit die Kategorien durch Beobachtungseinheiten bestimmt sind. Leider enthalten die Definitionen der Beobachtungseinheiten aber nicht näher definierte Wertungen, so daß eine Replikation der vorgenommenen Verhaltensbewertung aufgrund der in dem Artikel getroffenen Angaben auch hier nicht möglich ist. Beispielsweise wird als Indikator für die „Pathologie" „acting directly on feedback" u.a. die Beobachtungseinheit „higher decision stability" genannt (Jansson, ebd., S. 164). Eine Operationalisierung dafür, wie man dazu kommt, ein Verhalten als „higher decision stability" zu klassifizieren, fehlt aber. Letztendlich handelt es sich hierbei nicht um eine simple Beobachtungseinheit, sondern selbst schon um das Resultat einer umfassenden Verhaltensbeobachtung und *-bewertung*, die aber nicht dokumentiert ist. Die Mangel an Präzision bei der Definition der Beobachtungseinheiten ist ein fundamentales Problem des Artikels von Jansson. Der Autor setzt voraus, daß die qualitative Ausprägung der einzelnen Verhaltensweisen darüber entscheidet, welcher Typ eines fehlangepaßten Verhaltens durch die Beobachtungseinheit indiziert wird. Eine andere qualitative Ausprägung der genannten Einheit „decision stability", nämlich das Verhalten „lower decision stability" ist beispielsweise bereits als Indikator für einen anderen Fehlertyp („insufficient systematization") anzusehen. De facto werden alle von Jansson genannten Beobachtungseinheiten in ihrer Bedeutung durch die vorangestellten Worte „lower", „higher" oder „less" entscheidend modifiziert, ohne daß an irgendeiner Stelle nachvollziehbar erläutert wird, wie die entscheidende Bewertung von „high", „less" oder „low" funktioniert. Um „high", „less" oder „low" einstufen zu können, müßte man zunächst festlegen, was „normal" ist. Damit bleibt die eigentliche Aufgabe der Definition einer Beobachtungskategorie, nämlich die Sachverhalte festzulegen, die auftreten müssen, damit eine Beobachtungseinheit einer Kategorie zugeordnet wird, ausgespart. Unkommentiert bleibt auch die Frage der Vollständigkeit der Beobachtungskategorien, die sich insbesondere deshalb aufdrängt, da von Jansson keine Restkategorie vorgesehen

ist. Thematisiert wird hingegen die Notwendigkeit der Mehrfachkodierungen. So dienen mehrere Beobachtungseinheiten - wie z.B. die Einheit „lower information gathering stability" als Indikator von gleich drei verschiedenen charakteristischen Fehlertypen, andere indizieren zwei Fehlertypen. Dies führt dazu, wie Jansson auf Seite 170 ausführt, daß die Fehlertypen nicht unabhängig voneinander sind. Die entscheidenden Angaben zur Interkorrelation der „Pathologien" in der von Jansson durchgeführten Studie werden nicht berichtet. Die große Anzahl nicht unabhängig voneinander definierter Kriterien ist auch ein Punkt der umfassende Kritik der Arbeit von Jansson durch Funke (1995c). Funke macht darauf aufmerksam, daß drei Paare von Beobachtungseinheiten jeweils exakt die gleichen „Pathologien" (Fehlertypen) indizieren, ein Umstand der eine Reduzierung der Beobachtungseinheiten nahe legt. Zwei der sieben „Pathologien" (Fehlertypen) nach Jansson werden - wie Funke (ebd.) hervorhebt - durch den gleichen Indikatorsatz (Kombination von Beobachtungseinheiten) beschrieben. Obwohl es sich also um das gleiche Verhaltensmuster handelt, wird dieses einmal „insufficient systematization" und einmal „thematic vagabonding" genannt. Überraschenderweise berichtet Jansson für diese beiden - durch exakt die gleichen Beobachtungseinheiten indizierten - „Pathologien" *unterschiedliche* empirische Häufigkeiten! Man muß Funke (ebd.) schließlich auch zustimmen, daß es wenig überzeugend ist, wenn eine der genannten „Pathologien" (nämlich „selective information gathering") überhaupt nicht zu der vom Autor vorgenommenen Unterscheidung zwischen „guten" und „schlechten" Problemlösern beiträgt. Die Arbeit von Jansson (1994) zeigt deutlich die Defizite der Verhaltensmaße und einer darauf aufbauenden „Prozeßdiagnostik" auf. Für einen Ernstfalleinsatz in der (eignungs-)diagnostischen Praxis kommt ein solches Vorgehen nicht in Frage. Es bleibt zu befürchten, daß die Schwächen anderer Ansätze der Bildung von Verhaltensmaßen ebenso kraß auffallen würden wie die Schwächen des Ansatzes von Jansson, wenn die Autoren ihr jeweiliges Vorgehen vergleichbar nachvollziehbar dokumentieren würden.

Die Abhängigkeit der Beurteilung von dem jeweils gewählten Protokollmodus demonstrieren Putz-Osterloh und Kösters (1988). Bei 30 Probanden führten die Autorinnen zusätzlich zu den bis dahin verwendeten Tonbandprotokollen auch ein neues Protokollsystem für die Verhaltensdaten ein. Dieses automatisierte Protokollsystem erlaubte es den Versuchsleitern, u.a. schon während der Szenarienbearbeitungen eine Kategorisierung der Verhaltensdaten vorzunehmen. Die Ausprägung der Verhaltensmerkmale auf verschiedenen Dimensionen wurde nun einmal anhand der Tonbandprotokolle und einmal anhand des neuen Protokollsystems vorgenommen. Für drei Merkmalskategorien korrelierten die mit unterschiedlichen Protokollverfahren bestimmten Maße zwischen .40 und .80, die Korrelationen zwischen den Verhaltensmerkmalen und der Steuerungsleistung variierten in Abhängigkeit von

dem Protokolliermodus. Die Autorinnen ziehen das Resümee, daß es nicht möglich sein dürfte, mit Hilfe der Online-Protokolle zu vergleichbaren diagnostischen Aussagen zu kommen wie durch die Auswertung der vollständigen Tonbandprotokolle.

Nimmt man hilfsweise an, über die Frage der Informationsauswahl (welche Verhaltensweisen sollen beobachtet werden) und über die Frage der Definition von Verhaltenskonstellationen im Sinne von Strategien oder Typen könnte ein Konsens erzielt werden und die Verhaltensdaten könnten standardisiert beschrieben und zugeordnet werden, so bliebe noch das diagnostische Problem der Verhaltens*bewertung*: Welches Verhalten ist nun positiv und welches negativ zu bewerten? Auch diesbezüglich existiert offensichtlich kein verbindlicher Maßstab: Während sich die guten Problemlöser bei Dörner et al. (1983b) z.B. durch geringe Innovativität und Stabilität auszeichneten, berichten Kühle und Badke (1986, S. 101f.) für ein anderes Problemlöseszenario entgegengesetzte Befunde: gerade gute Problemlöser seien flexibel und innovativ. Die Diskrepanz der Befunde werten die Autoren als Hinweis auf die Systemspezifität der Strategien; eine gute Strategie sei jeweils die, die dem zu steuernden Szenario angemessen sei. Wäre dem so, ließe sich der Strategieansatz aber auch darauf verkürzen, daß eine gute Strategie diejenige ist, die im jeweiligen Szenario zum Erfolg führt. Naheliegend und nachvollziehbar wäre es, nur solche Verhaltensweisen zur positiven Kennzeichnung eines Problemlösers heranzuziehen, die auch tatsächlich *erfolgreiche* (im Sinne der Instruktion) Problemlöser kennzeichnen. Daß dies häufig nicht so ist, wird im folgenden Abschnitt dargestellt.

Vorab soll zusammenfassend festgehalten werden, daß die Beliebigkeit sowohl bei der Bestimmung relevanter Verhaltensweisen/Strategien/Typen als auch bei der Zuordnung einzelner Daten zu den definierten Einheiten sowie schließlich bei der vorgenommenen Bewertung gegen die diagnostische Verwendung von Verhaltensmaßen spricht. Ernsthafte Probleme dürften sich bei der Verwendung von Verhaltensmaßen auch hinsichtlich des Transparenzgebots der Diagnostik ergeben. Dieses Gebot schreibt vor, die Zielsetzung, Durchführung, Datenbasis und deren Interpretation offen zu kommunizieren, um so fehlerhaften Einschätzungen des Prozesses auf Seiten des Diagnostikanden vorzubeugen. Je nach Komplexität der angewendeten Verhaltensmaße führt eine solche Offenlegung entweder (bei geringer Komplexität der Verhaltensmaße, z.B. „Informationsabfragen zu Beginn stellen" als positive Verhaltensweise) zu einem hohen Ausmaß an Verfälschbarkeit (siehe unten, Abschnitt 10.2) oder aber - z.B. bei Verhaltensbewertungen, die konstellative Bedingungen berücksichtigen - zu einer Überforderung des Diagnostikanden (und wohl nicht selten auch des Testanweisers).

6.3.2.3 Das Problem der Unabhängigkeit von Verhaltensmaßen und Steuerungsleistungen

You must take the will for the deed(JONATHAN SWIFT; Polite Conversations, 1738)

Wenn bestimmte Verhaltensweisen und/oder Strategien bei der Bearbeitung von Problemlöseszenarien als „positiv" zu kennzeichnen und zu bewerten sind, so sicherlich diejenigen, welche zu einer vergleichsweise positiven Lösung und somit zu einem instruktionsgemäßen Steuerungserfolg führen. Wäre dies der Fall, so müßten die Verhaltensmaße und die Maße der Steuerungsleistung substantiell korrelieren und der Streit darüber, welches Maß angemessen ist, wäre für die praktische Diagnostik von untergeordneter Bedeutung. Erstaunlicherweise führen die als positiv gekennzeichneten und bewerteten Verhaltensweisen und Strategien aber häufig keinesfalls zum Problemlöseerfolg, während andererseits ein positiver Steuerungserfolg die Diagnostikanden offensichtlich nicht vor einer negativen Bewertung des Verhaltens schützt. Auf diesen Befund hat Neubauer (1995, S. 161) hingewiesen, der über eine Analyse der mit dem Szenario „Manage!" gewonnenen Daten berichtete. Dieser Analyse zufolge hatte das vom Rechner automatisch registrierte Verhalten wenig mit den Ergebniszuständen der Simulation zu tun.

Über den Zusammenhang von Verhaltensmaßen und Steuerungsleistungen wird nur gelegentlich explizit berichtet. Obermann (1991, S. 22ff) stellt beispielsweise den Zusammenhang von 15 Verhaltensmaßen und dem Problemlöseerfolg im Szenario „Airport" dar. Für fünf der Verhaltensmaße ließ sich erwartungswidrig kein Zusammenhang zum Problemlöseerfolg aufzeigen, so unterschieden sich bessere Problemlöser beispielsweise *nicht* in der Abfolge der Fragen zur Informationserhebung oder in der Überprüfung ihrer Entscheidungen von schlechteren Problemlösern. Selbst bei den Verhaltensmaßen, die laut Obermann (ebd.) im systematischen Zusammenhang zur Steuerungsleistung standen, handelte es sich oft um Zusammenhänge auf niedrigem Niveau. Beispielsweise korrelierte der Verhaltensindikator für die Risikobereitschaft erwartungsgemäß negativ mit der Problemlöseleistung, die Korrelation betrug aber lediglich $r. = -.22$.

Schaub (1990, S. 88) berichtet für eine Untersuchung mit 30 Versuchspersonen Korrelationen von $r = .40$ bis $r = .68$ zwischen der Steuerungsleistung (Szenario „Maschine") und einigen Verhaltensindizes.

Laut U. Funke (1992a, S. III-6) variierte die Steuerungsleistung im System „DISKo" bei 124 Personen weitgehend unabhängig von den Verhaltensmaßen „Effizienz der Informationsgewinnung" und „Analyse/Feedbacksuche". Auch der Verhaltensindex „Konzentration der Aktivitäten auf die wichtigsten Einflußgrößen" war lediglich zu $r = .09$ mit der Systemsteuerung (rang-)korreliert. Deutliche Zusammen-

hänge (Rangkorrelationen mit der Steuerungsleistung in Höhe von r=.33, r=.39 und r=.42) zeigten sich hingegen für die Indikatoren „Probehandeln" (Anzahl und Art der Testdurchläufe), „Systemwissen" und „Verhalten bei Entscheidungen".

Putz-Osterloh (1983) bewertete u.a. die Güte der Entscheidungen, die ihre 90 Versuchspersonen im ersten Drittel der Steuerung der „Schneiderwerkstatt" trafen. Dieser Verhaltensparameter korrelierte je nach Experimentalgruppe zwischen .45 und .55 mit Indikatoren der Problemlösegüte im ersten und zwischen .30 und .40 mit der Problemlösegüte in den folgenden Dritteln der Bearbeitung. Demgegenüber stand der Umfang des explorativen Verhaltens nur unter den Experimentalbedingungen in systematischer Beziehung zum Erfolg.

In der Studie von Jansson (1994) mit 40 Personen wurden sowohl die einzelnen Verhaltensweisen (pathologisch oder nicht) als auch der Erfolg im „Moro"-Szenario dichotomisiert. Die 7 Verhaltensweisen waren im Bereich von r=.22 bis r=.54 (Kontingenz-Koeffizienten) mit dem so kategorisierten Steuerungserfolg assoziert.

Die korrelativen Beziehungen zwischen Verhaltensmaßen und einem Leistungsindex für die Steuerung der „Schneiderwerkstatt" betrugen laut einem Bericht von Putz-Osterloh und Schroiff (1987, S. 213) für 32 Personen zwischen r=.27 und r=.44 (Rangkorrelationen). Später wurde diese Stichprobe um 68 weitere Personen ergänzt (Putz-Osterloh & Köster, 1988), für die Gesamtgruppe von 100 Personen ergaben sich entsprechende Rangkorrelationen von r=.14 bis r=.26. Daß der Zusammenhang zwischen Verhaltensmaßen und Steuerungsleistungen über verschiedene Untersuchungen hinweg schwankt, berichtet Putz-Osterloh (1989, S. 95) für das Beispiel des Strategiemerkmals „Organisationsvermögen". *„Einmal kovariiert es am höchsten von den vier erhobenen Strategiemerkmalen mit der Leistung (.50 bei N=48), während in einer anderen Stichprobe diese Korrelation (als einzige) das Signifikanzniveau nicht mehr erreicht (.13 bei N =100)."*

Indirekt offenbaren auch andere Untersuchungen den (fehlenden) Zusammenhang zwischen Verhaltensmaßen und Steuerungsleistungen. So zeigten sich beispielsweise Kreuzig und Schlotthauer (1991, S. 109) überrascht, daß Frauen in der Simulation „Manage!" signifikant schlechter abschnitten, obwohl die Frauen *„sich in keiner der Verhaltensvariablen in nennenswerter Weise von den Männern"* unterschieden. Umgekehrt unterschieden sich die „Experten" bei Strohschneider und Schaub (1991, siehe auch Schaub & Strohschneider, 1992) gerade auf der „strategischen Ebene" von den Laien, indem sie das Problem ausführlich explorierten und zurückhaltend eingriffen, während sich auf der Ebene der Steuerungsleistungen keine Unterschiede zeigten. Ähnliche Dissoziationen zwischen Verhaltens- oder Strategieunterschieden einerseits und Steuerungsleistungen andererseits ereigneten sich offensichtlich auch in der Studie von Reither (1981). Die erfahrenen Entwicklungshelfer unterschieden sich zwar durch ihr Entscheidungsverhalten von den Novizen, richteten aber ebenso

wie die Novizen in der vermeintlich simulierten Welt Unheil an. Auch die Experten der Studie von Putz-Osterloh (1987) unterschieden sich hinsichtlich des Erfolgs im „Moro"-System nicht systematisch von den Laien; die Manager konnten von der Autorin aber gleichwohl durch ihre besonderen Problemstrategien charakterisiert werden: Die Experten verbalisierten *„im größeren Umfang als Nichtexperten Problemlöseprozesse, die erfolgreiche Problemstrategien kennzeichnen"* (Putz-Osterloh, 1987, S. 80, siehe aber Putz-Osterloh und Lemme (1987, S. 295), die diese Strategieunterschiede nicht replizieren konnten). Fritz und Funke (1988) berichten, daß sich bei Jugendlichen mit minimalen zerebralen Dysfunktionen, die das „Ökosystem" bearbeiteten, hinsichtlich des Lösungserfolgs der Problemstellung keine Nachteile gegenüber der Kontrollgruppe zeigten - Strategieunterschiede in die erwartete Richtung konnten aber nachgewiesen werden. Bei Hesse, Spies und Lüer (1983, S. 412f.) führte äußerlich gleiches Verhalten (Informationsbeschaffung, Anfertigen von Aufzeichnungen) nur bei erhöhter persönlicher Betroffenheit zu guten Leistungen. Ebenso wirkte sich die von Dörner und Pfeifer (1991, 1992) realisierte experimentelle Variation einer Streßbedingung nicht auf den Steuerungserfolg, wohl aber auf verschiedene Formen des Steuerungsverhaltens aus.

Zahlreiche Befunde deuten also darauf hin, daß der Steuerungserfolg oft weitgehend unabhängig ist von den Verhaltensweisen der Steuerer. Neubauer (1995, S. 162) bringt diesen Sachverhalt überspitzt auf den Punkt: *„Der Simulation ist es vollkommen egal, welcher Steuerer vor ihr sitzt und sich einbildet, sie zu steuern. Sie produziert ihre Ergebnisse in der Hauptsache nach ihren eigendynamischen Gesetzen. Der Steuerer fungiert lediglich als eine Art Zufallsgenerator, der dafür sorgt, daß die Simulation weiterarbeiten kann. Mit systemischen Denken des Steuerers haben die Ergebnisse nichts zu tun."* Entsprechend dürfte der Steuerungserfolg natürlich auch nicht dem Steuerer zugeschrieben werden, der Rekurs auf Verhaltensweisen - die weitgehend unabhängig von Steuerungserfolg sind - erscheint unter dieser Perspektive als Rettung aus der Not. So betont beispielsweise Roelofsma (1995, S. 233) explizit das Auseinanderfallen zwischen Leistung und Verhalten und folgert daraus, daß *„bloße Messungen der Leistung nicht zufriedenstellend sein können"* und daß ein *„Bedarf nach prozeßorientierten Maßen"* besteht.

Die Idee, bestimmte Verhaltensweisen und Strategien, die bei der Problembearbeitung gezeigt werden, als positiv zu beurteilen, obwohl sie nachweislich nichts zur Steuerungsleistung und somit zur Problemlösung beitragen, traut man weniger den Konstrukteuren der Stadt Lohhausen als vielmehr den Bürgern von Schilda zu. Man fordert z.B. die Diagnostikanden in einem betriebswirtschaftlich eingekleideten Problemlöseszenario auf, das Gesamtvermögen ihrer Firma zu steigern und eventuell noch weitere Ziele - wie z.B. die Berücksichtigung der Interessen von Kunden und Mitarbeitern - zu verfolgen. Dann beurteilt man aber, weil es *„diagnostisch interes-*

santer" ist, die „*Wege und Wirkungen seines Handelns, zunächst ganz unabhängig davon, ob es zum wirtschaftlichen Erfolg geführt hat*" (Kreuzig, 1995b, S. 392). Anschließend klassifiziert man vielleicht einen Problemlöser, dem es nachweislich nicht gelungen ist, das Problem zu lösen und das Gesamtvermögen zu steigern, als erfolgreich, da er zum rechten Zeitpunkt die richtigen Fragen gestellt hat und dadurch eine „eigentlich" erfolgreiche Problemlösestrategie verfolgte. (Woher weiß man, daß diese Strategien „eigentlich" positiv sind?) Daß diese Strategien de facto nicht erfolgreich waren, darf dabei nicht weiter bekümmern. Ebenso ist es möglich, einen nachweislich erfolgreichen Problemlöser abzuklassifizieren, da seine Problemlösestrategien „eigentlich" wenig erfolgreiche Problemlöser charakterisieren. Oder man zwingt die inkommensurablen Verhaltensmaße und Steuerungsleistungen per Verrechnung zu einer hybriden Mesalliance in Form eines „Gesamtwerts".

Würde ein solches Vorgehen in der Eignungsdiagnostik angewendet, müßten die Diagnostiker sich unter Umständen nicht nur den Vorwurf fehlender Transparenz, sondern bereits den Vorwurf der Täuschung gefallen lassen. Es geht nicht an, eine Aufgabe mit vermeintlichem Leistungscharakter vorzugeben, aber dann nach leistungsunabhängigen Verhaltensratings auszuwerten und somit einen „Indikator" zu bestimmen, der mit der „offiziell" gestellten Aufgabe offenbar wenig zu tun hat.

Interessant ist auch eine Zusammenschau der Definition der Problemlösegüte über Verhaltensweisen die nichts mit dem Problemlöseerfolg zu tun haben einerseits und des Anspruchs auf Realitätsnähe und ökologische Validität der computergestützten Problemlöseszenarien (siehe Kapitel 4) andererseits. Greift man einmal den von Dörner (1989b) angestellten Vergleich der Problemlöseszenarienbearbeitung mit der Steuerung eines Atomkraftwerkes wie Tschernobyl auf, so könnte man die bei Problemlöseszenarienbearbeitungen vorgenommene Bewertung von Verhaltensweisen so übersetzen: die Reaktor-Havarie darf man nicht (über-)bewerten, die angewandte Strategie der Mitarbeiter im Kontrollraum war für sich betrachtet sehr positiv – oder kurz: Operation gelungen, Patient tot. Ein strategisch sinnvolles Agieren, das nicht im Zusammenhang mit der Steuerungsleistung steht, bleibt ein ritueller Regentanz: Man wartet auf Regen und verbessert den Tanz und nicht das Wetter. Ein solches Bewertungsvorgehen ist keineswegs realitätsnah.

6.4 Die Auswahl eines adäquaten Problemlösegütemaßes

Die Auswahl des in der diagnostischen Praxis verwandten Problemlösegütemaßes sollte sich zuallererst davon leiten lassen, daß der ausgewählte Indikator mit der Zielvorgabe (den Instruktionen), die den Diagnostikanden gegeben wurde, kompa-

tibel sein muß. Diese Voraussetzung läßt sich am einfachsten für Steuerungsleistungen erfüllen, die sich unmittelbar an der Zielvorgabe orientieren. Selbst bei unmittelbar an dem Steuerungsziel orientierten Problemlösegütemaßen ist die interne Validität des gewählten Leistungsindikators für die jeweils zu diagnostizierende Gruppe zu prüfen (siehe Kapitel 7). Weitere Kriterien der Auswahl sind das Ausmaß der Objektivität/Standardisierung der Zuordnung und Bewertung sowie natürlich die Reliabilität der Gütemaße (siehe Kapitel 8). Verhaltens- und Wissensmaße, die diese Kriterien nachweislich erfüllen, können *zusätzlich* zu Steuerungsleistungen als diagnostische Leistungsindikatoren verwendet werden, solange sie nachweislich im direkten Zusammenhang mit der instruktionsgemäßen Zielverfolgung stehen.

Zur Begründung der Verwendung von Verhaltensmaßen, die mehr oder minder unabhängig von der Steuerungsleistung variieren, wird in der Literatur häufig auf die unterschiedlichen Versuchsbedingungen verwiesen, die sich die Problemlöser durch ihre ungleichen Eingriffe selbst schaffen. Die Analyse von Verhaltensmaßen dürfte hier ihr fundamentum in re haben, Verhaltensmaße sind in diesem Licht betrachtet nicht nur interessante Prozeßmaße, sondern eine Art Zugeständnis an die Erkenntnis, daß die Dynamik der Systeme den Problemlösern oft über den Kopf wächst. Selbst minimale Eingriffe können in manchen Szenarien Ereignisketten von großer Tragweite auslösen, durch einen unachtsamen Befehl kann sich ein Teilnehmer u.U. in eine ausweglose Situation katapultieren usw. Würde man in dieser Situation lediglich die Steuerungsleistung beurteilen, liefe man Gefahr, die Person-Situation-Interaktion allein zu Lasten der Person zu interpretieren. *„Erfolg oder Mißerfolg in einem komplexen System kann also von einer Reihe verschiedener Faktoren abhängig sein, weshalb eine ausschließliche Berücksichtigung der Variable Erfolg/Mißerfolg zugunsten einer Prozeßbetrachtung der Handlungsorganisation beim Problemlösen aufgegeben werden sollte“*, schreiben beispielsweise Badke-Schaub und Tisdale (1995, S. 51) in einem mit „Möglichkeiten der Anwendung“ überschriebenen Kapitel zum Unterpunkt „Diagnostik“ und fordern eine kontextspezifische Diagnose der individuellen Verhaltensstrategien. Dem Anwender, der diagnostische Entscheidungen zu treffen hat, nutzt es aber wenig, wenn die mit computergestützten Problemlöseszenarien erzielten Daten, die eigentlich eine automatisierte „Diagnose“ darstellen sollen, nun selbst wieder zum Diagnostizierenden erklärt werden. Die in der jeweiligen diagnostischen Situation anstehende Frage der transparenten und nachvollziehbaren Entscheidung läßt sich nicht beantworten durch die bloße analytische Widerspiegelung eines Einzelfall-Wirrwarrs. Funke und Geilhardt (1996, S. 208) berichten, daß Anwender und Experten aus dem Bereich des Personalwesens bei der Diagnostik mit computergestützten Problemlöseszenarien *„eine »handelbare« Verlaufsdiagnostik jenseits von nicht verwertbaren Einzelprotokollen“* vermissen. Diagnostik drängt zum standardisierten Quantifizieren und zum

Vergleichen. Das interessierende Merkmal soll mit Hilfe des Meßinstrumentes numerisch abgebildet werden, über das numerische Relativ sollen Aufschlüsse über eine Person im Vergleich zu anderen gewonnen weden. Der von Badke-Schaub und Tisdale (ebd., S. 52) eingeräumte Umstand, daß automatische Auswertungsprogramme für die geforderte Prozeßbetrachtung nicht in Frage kommen, enttäuscht. Gleiches gilt für den Hinweis von Strohschneider und Schaub (1995, S. 202) auf den „Humbug" der Interpretation von automatisch ermittelten Verhaltenskennwerten. Diese zutreffenden Aussagen stellen eine Enttäuschung dar, weil mit der Einführung der Problemlöseszenarien Hoffnungen auf eine ökonomische und automatische Diagnostik, die über die bisherige Statusdiagnostik hinaus führt, geweckt wurden (siehe Kapitel 3). Die nun geforderte konstellative, strikt einzelfallorientierte (Strohschneider & Schaub, ebd, S. 201) Aus- und Bewertung „per Hand" muß bei einem tatsächlich intransparenten, komplexen, vernetzten und dynamischen System jeden diagnostischen Anwender vom Aufwand und vom Können her (Voraussetzung wäre eine vollständige Programmkenntnis) überfordern. Daß den für die diagnostische Praxis bestimmten Programmen überhaupt keine entsprechenden Auswertungsrichtlinien beigefügt sind, zeigt, daß dieses Vorgehen nicht ernsthaft für die diagnostische Praxis erwogen wird.

Verhaltens- und Wissensmaße können auch für die in der vorliegenden Arbeit interessierende Fähigkeitsdiagnostik interessant sein. Voraussetzung ist aber, daß diese Maße theoretisch und empirisch begründet definiert sowie standardisiert erhoben und bewertet werden. Die positiv bewerteten Verhaltensweisen/Strategien müssen außerdem nachweislich mit den Zielen der Steuerung in Einklang stehen, die den Diagnostikanden genannt wurden. Solche Verhaltens- und Wissensmaße stellen keine Alternative, sondern bestenfalls eine Ergänzung der für die Fähigkeitsdiagnostik *vorrangigen* Indikatoren der Steuerungsleistung dar. Der Hinweis, daß richtige Lösungen - wie Spada und Reimann (1988, S. 184) schreiben - auf sehr verschiedenen Wegen zustande kommen können und somit der Rückschluß von der manifesten auf die latente Ebene erschwert wird, zählt z.B. in der praktischen Eignungsdiagnostik wenig, solange der Nachweis der Kriteriumsvalidät gelingt. Solange die positive Steuerungsleistung mit einer akzeptablen Wahrscheinlichkeit mit zukünftigen positiven Werten auf dem Eignungskriterium einhergeht, ist es aus eignungsdiagnostischer Perspektive weniger interessant, *wie* die gute Steuerungsleistung zustande gekommen ist. Die vorliegende Arbeit beschränkt sich auf die in jedem Fall vorgeordnete Auswertung von Steuerungsleistungen - mögen diese Auswertungsstrategien einigen auch reduktionistisch vorkommen.

Die strikt einzelfallorientierte und konstellative Auswertung von Verhaltensweisen „per Hand" ist als Forschungsstrategie und hypothesengenerierendes Verfahren dem Einsatz von Problemlöseszenarien *im Forschungskontext* zuzuordnen. Aus ähnlichen

Gründen, aus denen die Regeln zur »Philosophie der Verwendung von Mikrowelten« nicht auf die diagnostische Praxis angewandt werden können (siehe Abschnitte 5.2 und 5.3), können auch die Empfehlungen einer einzelfallorientierten und konstellativen Verhaltensbewertung nicht auf die diagnostische Praxis bezogen werden. Dörner und seine aktuellen Mitarbeiter haben diese Techniken weder für die praktische Diagnostik entworfen noch dort eingesetzt. Anbieter von Lösungen diagnostischer und somit technologischer Probleme können sich daher nicht auf diese für Forschungsfragen entwickelte konstellative Verhaltensbewertung berufen und sich diese „neuen Methoden" rhetorisch zu Nutze machen, um die Schuldenkonten des „unreflektierten Methodismus" oder der „Endproduktdiagnostik" zu saldieren und *gleichzeitig* auf dem diagnostischen Markt verdienen zu wollen.

Für die diagnostische Praxis ist die Intention, mit (konstellativen oder allgemeinen) Verhaltensmaßen die durch die Eingriffe der Problemlöser diversifizierten Untersuchungsbedingungen zu kompensieren, zwar nachvollziehbar, führt aber zu den oben genannten Problemen, die letztendlich in intransparenten Beurteilungen und in einer Überforderung des Diagnostikers resultieren. Die Unvergleichbarkeit der Versuchsbedingungen und die möglicherweise daraus folgende mangelnde Korrespondenz zwischen Strategiemaßen und Steuerungsleistungen reflektieren Probleme der Meßinstrumente, die nicht den Diagnostikanden und Diagnostikern aufgebürdet werden sollten. Man kann nicht erst ein Programm verkaufen, in dem positive Verhaltensweisen nicht zum Erfolg führen und dann die so entstehenden Probleme mit einer der Erfolgsvariablen unangepassten Verhaltensbewertung „lösen". Anstelle des Versuchs, die Probleme der computergestützten Problemlöseszenarien durch die speziellen und problematischen Auswertungstechniken irgendwelcher Verhaltensmaße zu kompensieren, sollte der Versuch treten, die Instrumente und die Erfolgskriterien so zu konstruieren, daß eine individuelle Kontrolle über Art und Inhalt der Simulationsgeschehnisse jederzeit für alle Diagnostikanden zumindest möglich ist und somit vermeintlich positive Verhaltensweisen auch zu positiven Resultaten führen können. Dazu ist es notwendig, *steuerbare* Systeme mit angemessenem Schwierigkeitsgrad zu konstruieren (siehe Kapitel 7).

6.5 Zusammenfassung, Schlußfolgerungen und Ausblick

Zur Beurteilung des Problemlöseerfolgs wurden zahlreiche Gütemaße entwickelt, die in ihrer Unterschiedlichkeit (und vermutlich auch in ihrer ungeprüften unterschiedlichen psychometrischen Qualität) zu einem Großteil der konzeptionellen Konfusion in der Problemlöseforschung beigetragen haben. Grundsätzlich können die

Ergebnisse der Bearbeitung von computergestützter Problemlöseszenarien auf der Ebene der Steuerungsleistungen, auf der kognitiven Ebene (z.B. Wissensstrukturen) und auf der Ebene der Verhaltensmaße quantifiziert werden (siehe Abbildung 2). Die Ebene der Steuerungsleistungen wird in der vorliegenden Arbeit für den Einsatz von computergestützte Problemlöseszenarien zur Fähigkeitsdiagnostik bevorzugt, da derart gebildete Problemlösegütemaße im Sinne der Transparenz mit der Instruktion an die Diagnostikanden übereinstimmen. Außerdem können Steuerungsleistungen mit der notwendigen Objektivität erfaßt werden. Die Übereinstimmung von Instruktion und Gütekriterium und die Objektivität sind notwendige Voraussetzung für die Reliabilität und Validität der Gütekriterien. Weiterhin sollte gewährleistet sein, daß es zumindest einer substantiellen Teilstichprobe der Zielpopulation gelingt, das System zielgerichtet erfolgreich zu steuern (siehe Kapitel 7).

Verhaltens- oder Wissensmaße sollten in der diagnostischen Praxis nur dann ergänzend hinzugezogen werden, wenn die entsprechenden Verhaltensweisen oder Wissensbestände (1.) einer theoretisch und empirisch begründeten Definition folgen, (2) objektiv und standardisiert erhoben und bewertet werden können und (3.) nachweislich zur instruktionsgemäßen Zielerreichung dienen. Diese Voraussetzungen sind für viele der in der diagnostischen Praxis heute angewendeten computergestützten Problemlöseszenarien nicht erfüllt. Die Forderung nach einer konstellativen, einzelfallbezogenen „Prozeßbetrachtung", bei der versucht wird, die Eingriffe in das System unter Berücksichtigung des jeweiligen Systemzustandes zu bewerten, wird aufgrund der mangelnden Standardisierung und der Überforderung des diagnostischen Anwenders, für die diagnostische Praxis zurückgewiesen. Mit Hilfe einer solchen konstellativen Auswertungsmethode oder mit einer Bewertung von Verhaltensweisen, die in keinem Zusammenhang zum Steuerungserfolg stehen, soll den unvergleichbaren Versuchsbedingungen Rechnung getragen werden, die sich die Diagnostikanden durch ihre ungleichen Eingriffe selbst schaffen. Der Versuch, durch diese Arten der Auswertung die Gefahr zu bannen, die Person-Situation-Interaktion allein zu Lasten der Person zu interpretieren, muß als untauglich für die diagnostische Praxis gewertet werden. Als Alternative zu solchen kompensatorischen Auswertungstechniken wird vorgeschlagen, steuerbare Systeme zu konstruieren, bei denen positive Verhaltensweisen auch zu einem Steuerungserfolg führen. Dieser Aspekt wird im folgenden Kapitel aufgegriffen. Zusätzlich zu den in diesem Kapitel behandelten Aspekten ist bei allen Problemlösegütemaßen die Frage der Reliabilität zu berücksichtigen, die in Kapitel 8 aufgegriffen wird.

7. Zur Steuerbarkeit der Systeme

Weiter oben im Text (Abschnitt 6.3.2.3) wurde bereits darauf hingewiesen, daß unter Umständen positive Verhaltensweisen der Systemsteuerer keinen sonderlich positiven Effekt auf die Steuerungsleistung nach sich ziehen. Die Systemdaten beschreiben in diesen Fällen nicht die Personmerkmale. Einige Szenarien sind so zugeschnitten, daß unter Umständen die Aufgabenmerkmale die Ergebnisse determinieren. Diagnostisch sind solche Systemdaten unbrauchbar. Dieses Problem ist auf die Dynamik computergestützten Problemlöseszenarien zurückzuführen und somit in gewissen Grenzen grundsätzlich gegeben. Gleichwohl gibt es deutliche Unterschiede in der Steuerbarkeit und im Schwierigkeitsgrad der einzelnen computergestützten Problemlöseszenarien. Das vorliegende Kapitel thematisiert die Frage der Steuerbarkeit und der Schwierigkeit von Szenarien und plädiert dafür, im diagnostischen Kontext nur „steuerbare" Szenarien mit einer mittleren Schwierigkeit zu verwenden. Anschließend wird darauf hingewiesen, daß die Steuerung der meisten Szenarien zu schwer ist. Welche Folgen diese Überforderung für die interne Validität der Problemlösegütemaße und somit für die Interpretation der Daten haben kann, wird am Beispiel einer älteren Version der „Schneiderwerkstatt" veranschaulicht.

7.1 Erwünschte und unerwünschte Schwierigkeit

Natürlich muß einem Problem eine gewisse Schwierigkeit zu eigen sein, sonst wäre es ja kein Problem. Probleme belasten das Arbeitsgedächtnis (Anderson, 1983), etwa durch die notwendige Repräsentation des Problems oder durch die Notwendigkeit, simultan mehrere Zwischenziele und -resultate zu berücksichtigen (Kotowsky et al., 1985). Diese kognitiven Anforderungen und die daraus resultierende Schwierigkeit sollen als diagnostisch grundsätzlich *„erwünschte* Schwierigkeit" von Problemlöseszenarien bezeichnet werden, obgleich es auch bei dieser diagnostisch *„erwünschten* Schwierigkeit" natürlich auf das richtige Ausmaß ankommt. Zu wenig oder zu viel (siehe z.B. Kotowsky & Simon, 1990, für Beispiele sehr schwerer Problemlöseaufgaben) schadet natürlich auch hier der intendierten Messung.

Demgegenüber sollen als für eine Leistungs- und Fähigkeitsmessung diagnostisch *„unerwünschte* Schwierigkeiten" alle Gegebenheiten aufgefaßt werden, die den Problemlöser aus nichtkognitiven Gründen daran hindern, zu einer Problemlösung zu

kommen. Hierzu zählen Bedienungsschwierigkeiten (vgl. Abschnitt 10.1.2), frühe irreversible Systemzustände, vom Problemlöser nicht zu verantwortende und nicht vorhersehbare extreme Sytemereignisse (Überraschungs-Katastrophen/Notfälle) usw. Über die grundsätzlichen Voraussetzungen der *„Kontrollierbarkeit"*, *„Beobachtbarkeit"*, *„Steuerbarkeit"* und *„Erreichbarkeit"* (vgl. die entsprechenden Ausführungen von Hübner, 1989b, für diskret lineare zeitinvariante Systeme im Rahmen der Zustandsraumdarstellung) hinaus, sind für Problemlöseszenarien, die für eine diagnostische Verwendung bestimmt sind, durchaus auch zielgruppenspezifische mittlere Schwierigkeiten im statistischen Sinne wünschenswert (Lienert, 1967, S.39 f.).

7.2 Überforderung

Im nachhinein läßt sich natürlich nicht immer entscheiden, ob sich die computergestützten Problemlöseszenarien aufgrund erwünschter oder unerwünschter Schwierigkeiten der Steuerung durch die jeweils untersuchten Problemlöser entzogen. Dies spielt aber auch keine Rolle, da Überforderung generell zu Einschränkungen der diagnostischen Verwendbarkeit führt - ebenso wie die Unterforderung. Neben den bekannten Problemen von unzweckmäßigen Schwierigkeitsgraduierungen, die in asymetrischen Verteilungen und „Decken"- oder „ground"-Effekten münden, ist bei einer *Über*forderung der Steuerer computergestützter Problemlöseszenarien auch zu berücksichtigen, daß die erzielten Ergebnisse teilweise auf die Aufgabenmerkmale und nicht auf die Personmerkmale zurückzuführen sind. Ein „ungebändigtes" dynamisches System ist nicht mit einem unbeschriebenen Antwortbogen zu vergleichen; es zeichnet nur bedingt - wie ein mit falschen Lösungen beschriebenener Antwortbogen - die Fehler der Probanden nach. Ein „ungebändigtes" System kann vielmehr seine eigene Dynamik ausbreiten. Des weiteren führt Überforderung zur Demotivierung und zu Frustrationserlebnissen bei den Diagnostikanden - mit entsprechenden Extrem-Reaktionen. Dies - so werden Protagonisten desaströser Szenarien einwenden - sei doch durchaus wünschenswert. Denn so könnte man eben Denken unter belastenden Bedingungen untersuchen. Im günstigsten Falle würde das Verhalten in desaströsen Szenarien dieser Argumentation zufolge ein Konglomerat aus Emotion und Kognition widerspiegeln. Auch wenn man akzeptiert, daß Emotion und Kognition zusammengehören, ist es differentialdiagnostisch sinnvoll, die Komponenten getrennt zu untersuchen (vgl. Eysenck, 1988). Bei diagnostischen Anwendungen im Rahmen der Personalauswahl ist außerdem zu bedenken, daß es um die Diagnose *eignungsrelevanter* Eigenschaften und Fähigkeiten geht. Nur in wenigen Berufsbildern muß man sein Denkvermögen überwiegend unter katastrophalen Um-

ständen unter Beweis stellen. Ganz grundsätzlich darf mit Dörner (1981) angezweifelt werden, daß das Verhalten in katastrophalen Situationen überhaupt noch etwas mit „Denken" zu tun hat. Dörner (1981, S. 171) führt nämlich aus, daß das Denken in extremen Fällen bei schlechten Versuchspersonen zugunsten des Tuns ganz ausgeschaltet wird („Notfallreaktionen"). Solch eine extreme Situation, die sich den *„Manipulationsbemühungen eines Individuums gegenüber unzugänglich zeigt"* (ebd.) wurde bei einigen Probanden durch das „Lohhausen"-Experiment hervorgerufen. Damit kommen extrem schwere Szenarien möglicherweise für die Diagnostik des Verhaltens in Extremsituationen, nicht aber - wie allgemein postuliert - für eine Diagnostik des Alltagsverhaltens, und schon gar nicht für eine Diagnostik des Denkens in Frage. Es muß im nachhinein verwundern, daß die Autoren des „Lohhausen"-Experiments überhaupt annahmen, daß sich das Verhalten in einer Situation, in der das Denken ihrer Ansicht nach unter Umständen „ausgeschaltet" wird, durch Intelligenztests prognostizieren ließe.

Es bleibt festzuhalten: Jegliche Persondiagnostik (ob unter belastenden Bedingungen oder nicht) aufgrund der Kennwerte computergestützter Problemlöseszenarien wird obsolet, wenn diese Kennwerte der Systemdynamik und nicht dem Systemsteuerer zugeschrieben werden müssen.

Die Überforderung der Steuerer (aus welchen Gründen auch immer) ist bei der Konstruktion von komplexen Problemlöseszenarien offensichtlich zur Regel geworden. Schon bei Dörner und Reither (1978, p. 527) heißt es, daß die Probanden *„fast ausnahmslos das ursprünglich stabile Gefüge der Variablen des simulierten Landes zerstörten und dadurch häufig katastrophale Zustände schufen"*. Auch für zahlreiche andere Probanden handelte es sich bei der Systemsteuerung offenbar um eine Danaidenarbeit, wie die nachfolgenden Beispiele veranschaulichen können. Laut Reither (1981) steuerten Laien und erfahrenen Entwicklungshelfer ein Entwicklungshilfeszenario („Dagu") gleichermaßen in katastrophale Zustände. Für das System „Moro" berichten Schaub und Strohschneider (1992, S. 121), daß die Steuerung auch den „Experten" recht schwer fiel. Etliche rutschten in katastrophale Zustände ab. Entsprechendes berichtet Putz-Osterloh (1987, S. 76): Nur acht von 37 Experten und Novizen gelang es bei „Moro" die Zielvariablen in eine positive Richtung zu beeinflussen, neun von 37 versetzten das System in einen „katastrophenträchtigen" Zustand. Laut Strohschneider (1991a, S. 363) haben die Versuchspersonen bei „Moro" durchschnittlich 1055 Stück Vieh zu wenig. Auch für die Steuerung eines von Dörner und Preußler (1990, S. 206) als „einfaches Räuber-Beute-System" charakterisierten Szenarios ergab sich: *„Die Vpn kamen also von Anfang an mit der Steuerungsaufgabe nicht gut zurecht und lernten im Durchschnitt auch nichts dazu"* (ebd., S. 210). E. Müller (1991, S. 190) berichtet für das System „Tankwagen", daß die Probanden durch die vorliegende Einstellung der Modell-

parameter überfordert waren, und K.J. Klauer (1996, S. 101) stellt fest, daß in einem von zwei Experimenten fast alle seine Versuchspersonen den zu steuernden Betrieb schon nach der Hälfte der zur Verfügung stehenden Zeit „ruiniert" hatten. Jansson (1994, S. 164) attestiert 72,5% seiner Versuchspersonengruppe, daß „Moro" Szenario schlecht gesteuert zu haben. Putz-Osterloh (1995, S. 408) weist allgemein darauf hin, daß Versuchspersonen mit nichtlinearen Entwicklungen selten erfolgreich umgehen können, so daß, falls der Effekt von Entscheidungen auch noch zeitlich verzögert wird, oft keine Steuerung mehr gelingt. Für Szenarien, die für die eignungsdiagnostische Praxis konzipiert wurden, räumt Obermann (1995, S. 405) ein, daß die ersten Versionen des Szenarios „Airport" zu schwer waren. Für das eignungsdiagnostisch eingesetzte Szenario „Textilfabrik" stellte Hasselmann (1993, S. 152) fest, daß die Teilnehmer seiner Untersuchung mit der Steuerung nicht überfordert waren. Der Autor leitet seine Aussage dabei aus einem Vergleich der Steuerungsleistungen seiner insgesamt 52 Probanden mit dem Ergebnis eines ungesteuerten „Null-Laufs" des Systems ab. Der „Null-Lauf" als Vergleichsmaßstab bildet jedoch nur einen Aspekt der Aufgabenschwierigkeit ab. Einen anderen Maßstab bildet der Vergleich mit dem Startwert. Den auf Seite 151 der Arbeit von Hasselmann abgebildeten Verlaufsgraphiken für den Kapitalwert ist zu entnehmen, daß über 80% der Teilnehmer ein Ergebnis erzielten, welches schlechter war als der Startwert. Geht man davon aus, daß die Testanden ihre Leistungen nicht am „Null-Lauf" (dieser Wert ist den Teilnehmern nicht bekannt), sondern am Startwert messen, dürften diese sich bei der Bearbeitung zumindest überfordert gefühlt haben. Hasselmann (ebd., S. 153) berichtet, daß die Probanden nur in 25% der Simulationsmonate der „Textilfabrik" einen Anstieg des Gesamtkapitalwertes erzielen konnten. Hinsichtlich des Gütemaßes „Simulationsmonate mit Gewinn" berichtet Locher (1997, S. 81f.), daß die 20 Versuchspersonen bei der Erstbearbeitung der „Textilfabrik" im Durchschnitt nur 34,5 % der Takte mit Gewinn steuerten, beim „Heizölhandel" ergab sich für 40 Personen sogar nur ein Mittelwert von 19,4%.

Dörner und Reither (1978, S. 534) parieren das kritische Argument der Überforderung der Probanden mit dem Hinweis, daß es darum geht, die Versuchspersonen einer Situation auszusetzen, *„die der Realität in Hinblick auf die ... genannten Merkmale entspricht"*. Diese Replik ist nicht nur deshalb wenig überzeugend, weil sich die Autoren ja nicht die Mühe von Abbildungsvorschriften im Sinne einer ökologischen Validität gemacht haben (siehe Kapitel 4), denen sie sich nun verpflichtet fühlen müßten und weil die vermeintlich abgebildeten Alltagsanforderungen eben per definitionem keine katastrophale Überforderung darstellen können. Die Replik ändert vor allem nichts an den diagnostisch negativen Konsequenzen der Überforderung. Wie als Konsequenz der Überforderung die Problemlösegütemaße ihre interne Validität einbüßen können, beschreibt der nächste Abschnitt.

7.3 Potentielle Effekte der Überforderung auf die interne Validität der Problemlösegütemaße am Beispiel der Berliner (Erst-)Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen

Süß, Oberauer und Kersting (1993b) berichten über die Berliner Erstuntersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen. 214 Schüler steuerten dabei dreimal hintereinander die „Schneiderwerkstatt" (Fassung 2.3 von J. Funke) mit jeweils veränderten Startwerten. Als Problemlösegütekriterien wurden - der Steuerungsvorgabe entsprechend - zunächst das Gesamtvermögen am Ende der Bearbeitungszeit und, als Trendmaß, die Zahl der Monate mit Gewinn bestimmt. Im ersten Durchgang gelang es *keinem* Probanden, das angestrebte Ziel zu erreichen, nämlich das Gesamtvermögen des Unternehmens zu steigern. Mehr als die Hälfte aller Probanden blieb unter dem Ergebnis des „Null-Laufs" des Systems. Das heißt, diese Personen erzielten ein schlechteres Ergebnis, als wenn sie über alle Takte („Betriebsmonate") überhaupt keinen Eingriff vorgenommen hätten (vgl. Abschnitt 10.2 zur Verfälschbarkeit der Maße). Nicht viel günstiger waren die Ergebnisse für das Trendmaß: Fast 70% hatten in *keinem* Monat das Gesamtvermögen verbessert. Gerade zwei von 214 Probanden gelang dies in mehr als der Hälfte der Monate. Nur wenig besser waren die Ergebnisse im zweiten und dritten Durchgang. Die Versuchspersonen waren mit dem System überfordert. Zusammenhangsanalysen zwischen Problemlöseleistungen einerseits und Intelligenz und Wissen andererseits, die auf der Basis dieser Problemlösgütemaße (das Gesamtvermögen am Ende der Bearbeitung in normalisierter Form und das Trendmaß in einer leicht modifizierten Variante) berechnet wurden, ergaben einen eindeutigen Befund: es gab keine Zusammenhänge. Dieser Befund hätte als weiterer Hinweis auf den fehlenden Zusammenhang zwischen Problemlöseleistungen und Intelligenztestleistungen gut in die damalige Literaturlage gepaßt.

7.3.1 Aufgabenanalyse des Problemlösegütemaßes unter den Bedingungen der Überforderung der Probanden

Die eindeutige Überforderung der Probanden weckte bei den Autoren Zweifel an der internen Validität der Problemlösegütemaße. Süß et al. (1993b, S. 92 f.) prüften daher mit einer Aufgabenanalyse, ob das Gütemaß unter den gegebenen Bedingungen überhaupt das Ausmaß zweckrationalen problemlösenden Handelns erfasste. Vernünftiges Handeln im Sinne der Instruktion ist es, in der Problemsituation „Schneiderwerkstatt" zu versuchen, durch den Verkauf von „Hemden" Gewinne zu

erwirtschaften. Andere Einnahmequellen als der Verkauf von „Hemden" stehen, mit Ausnahme der vernachlässigbaren Größe der Zinseinnahmen, im Programm nicht zur Verfügung. Die Gewinne sind das *Produkt* aus der Anzahl verkaufter „Hemden" und der „Gewinnspanne pro verkauftem Hemd". Vernünftig ist es also, sowohl die „Gewinnspanne pro Hemd" zu steigern als auch die Anzahl verkaufter „Hemden". Das Ziel, Gewinn zu machen, läßt sich deshalb in diese zwei Teilziele zerlegen. Da die „Gewinnspanne pro verkauftem Hemd" im Programm „Schneiderwerkstatt" nicht berechnet wird, wurde diese Variable im Rahmen der Aufgabenanalyse wie folgt berechnet: Veränderung des Gesamtvermögens im Vergleich zum Vormonat, dividiert durch die Anzahl verkaufter „Hemden" in diesem Monat. Denselben Wert erhält man, wenn für jeden Monat von den Einnahmen die Summe aller Kosten subtrahiert und durch die Anzahl verkaufter „Hemden" teilt.

Wenn nun die Gewinnspanne pro verkauftem „Hemd" infolge der Überforderung bei *allen* Probanden negativ ist, erwirtschaften die Versuchspersonen mit jedem verkauften „Hemd" nicht Gewinn, sondern Verlust. Das bedeutet: je mehr sie verkaufen, desto mehr Verluste haben sie[8]. Genau das war bei den Probanden der Berliner Erstuntersuchung der Fall. *Alle* Teilnehmer erzielten, gemittelt über alle Simulationsmonate, negative Gewinnspannen pro verkauftem Hemd, nur wenigen ist es gelungen, das System überhaupt einmal in eine positive Gewinnspanne zu steuern.

Der Grund dafür, daß das Problem für die Probanden so schwierig war, lag nicht darin, daß das System prinzipiell nicht zu steuern war. Im Zustandsraum des Systems „Schneiderwerkstatt" (also im n-dimensionalen Raum seiner möglichen Zustände, mit n=Anzahl der Variablen) gab es aber nur einen schmalen Sektor, in dem die Gewinnspanne pro verkauftem „Hemd" positiv war. In dieses schmale Zielfenster konnten die Versuchsteilnehmer das System unter den vorgegebenen Be-

---

[8] Diese Analyse unterstellt zunächst vereinfachend, daß die Gewinnspanne pro verkauftem „Hemd" unabhängig ist von der Anzahl verkaufter „Hemden". Diese Annahme trifft für das System „Schneiderwerkstatt" größtenteils zu. Im System „Schneiderwerkstatt" kann der Betriebsleiter nicht, wie in realen Betrieben, einfach beschließen, produzierte Waren nicht zu verkaufen. Der Verkauf von „Hemden" in einem simulierten Monat erfolgt „automatisch" in Abhängigkeit von Angebot (= Zahl „Hemden im Lager") und Nachfrage. Angebot und Nachfrage hängen systemimmanent ab von den Ausgaben des Probanden für Produktion, Werbung, etc. Die Variable „Anzahl verkaufter Hemden" im System „Schneiderwerkstatt" ist also eine (direkte oder indirekte) Funktion aller anderen Variablen (außer der Variable Gesamtvermögen). Sie ist - bei *gegebenen* Relationen zwischen den Variablen - ein Indikator für die absolute Größe des Betriebs, die der Proband als Betriebsleiter selbst bestimmt. Die „Gewinnspanne pro verkauftem Hemd" ist andererseits ein Indikator für die Güte der Abstimmung der Variablen untereinander, also ein Indikator für den Grad der Optimierung der Variablenverhältnisse. Jedes suboptimale Verhältnis zwischen Variablen produziert „tote Kosten", Ausgaben, die keinen Effekt haben, weil andere Ausgaben nicht auf sie abgestimmt sind.

dingungen offensichtlich nicht gezielt bringen und halten. Vielleicht fehlte ihnen für eine erfolgreiche Steuerung das spezifische Wissen. Wesentliche Information, die ein „echter" Manager selbstverständlich hätte (beispielsweise über die Kosten einer neuen Maschine oder über die zum Bedienen einer Maschine erforderliche Anzahl der „Arbeiter") wurden den Probanden in der üblichen Präsentationsbedingung vorenthalten. In der zur Verfügung stehenden Bearbeitungszeit konnten die Probanden dieses Wissen offensichtlich nicht erwerben.

Solange sich nun das System im Bereich negativer Gewinnspannen bewegt, maximieren Probanden, die den Verkauf maximieren, dadurch nicht ihre Gewinne, sondern ihre Verluste. Eine Expansion des Betriebes muß sich unter solchen Umständen schädlich auf das Ziel der Gesamtvermögensmaximierung auswirken: Tatsächlich wiesen Indikatoren für die Größe des Betriebs (z.B. Anzahl der „gekauften Maschinen", Anzahl der produzierten „Hemden") hohe negative Korrelationen mit den untersuchten Gütemaßen auf. Die Vergrößerung des Betriebs schadete also. Expansion erhöht zwar möglicherweise die Verkaufszahlen und damit auch die Einnahmen, sie erhöht aber notwendigerweise in ungefähr gleichem Maße die Kosten. Mehr „Hemden" zu verkaufen ist nicht möglich, ohne zugleich die Kosten für „Produktion", „Werbung" oder „Verkaufsstellen" zu erhöhen. Daher führt eine Verbesserung der Verkaufszahlen in der Regel nicht zu einer Verbesserung der „Gewinnspanne pro verkauftem Hemd". Wer beispielsweise alle Ausgaben gleichzeitig verdoppelt, verdoppelt damit auch den Verkauf - die Abstimmung zwischen den Variablen wird jedoch nicht besser. Die „Gewinnspanne pro verkauftem Hemd" bleibt nahezu konstant. Wenn ein Proband aber den Verkauf durch allgemeine Expansion seines Betriebes erhöht, ohne dabei die „Gewinnspanne" wesentlich zu ändern, vermehrt er damit seine Gesamtverluste.

Nachdem die für die Berliner Probanden der Erstuntersuchung geltenden Bedingungen der Systemsteuerung beschrieben sind, kann man sich einen idealtypischen *guten* Problemlöser vorstellen, einen Probanden also, der auf der Grundlage der gegebenen Voraussetzungen der Problemsituation vernünftig handelt. Diese Voraussetzungen sind: das vorgegebene Ziel, das in die beiden genannten Subziele zerlegbar ist („Gewinnspanne pro Hemd" und Anzahl verkaufter „Hemden" steigern) und die stark begrenzten Systeminformationen, die die Probanden bekommen. Ein guter Problemlöser wird beide Teilziele gut erreichen: Er wird viele „Hemden" verkaufen, und er wird eine relativ hohe „Gewinnspanne pro Hemd" erzielen. Eine relativ hohe „Gewinnspanne" war aber aufgrund der spezifischen Systembedingungen bei den überforderten Probanden immer noch ein *negativer* Wert und bedeutete also: relativ *geringe Verluste* pro „Hemd". Wenn nun ein guter Problemlöser den Verkauf *und* die „Gewinnspanne" steigert, ist sein Gesamtverlust das Produkt aus einer *hohen* Zahl verkaufter „Hemden" und einem relativ *geringen* Verlust pro verkauf-

tem „Hemd". Beispielsweise könnte es einer vernünftig agierenden Person unter diesen Umständen gelingen, den negativen Wert der „Gewinnspanne" gering zu halten, z.B. „nur" 100.-DM Minus mit jedem verkauften „Hemd" zu erzielen. Gleichzeitig wird es dieser vernünftigen Person gelingen, viele „Hemden" (z.B. 1.000) zu verkaufen. Der Erlös aus dem Verkauf der „Hemden" wird für diese vernünftige Person den Wert -100.-DM * 1.000 „Hemd" = -100.000 DM betragen.

Beispiel für einen unter den Umständen der Berliner Untersuchung (zu schweres Szenario, *kein* Proband erzielte eine positive Gewinnspanne) relativ *guten* Problemlöser (Werte für einen Bearbeitungstakt):

- ❒ der Wert der *negativen* „Gewinnspanne" (kein Proband erzielte einen positiven Wert) ist gering ausgeprägt, z.B. wird pro „Hemd" lediglich ein *Verlust* von -100.- DM erzielt
- ❒ guter Verkaufserfolg: 1.000 „Hemden"

Ergebnis für diesen *guten* Problemlöser:

-100 DM * 1.000 „Hemden" = ***-100.000 DM***

Beispiel für einen unter den Umständen der Berliner Untersuchung relativ *schlechten* Problemlöser (Werte für einen Bearbeitungstakt):

- ❑ der Wert der *negativen* „Gewinnspanne" ist hoch ausgeprägt, z.B. wird pro „Hemd" ein *Verlust* von -1.000.- DM erzielt
- ❑ schlechter Verkaufserfolg: 100 „Hemden"

Ergebnis für diesen *schlechten* Problemlöser:

-1.000 DM * 100 „Hemden" = ***-100.000 DM***

⇨ unter diesen Umständen ist das Problemlösegütemaß *invalide*, das Maß unterscheidet nicht zwischen guten und schlechten Problemlösern

Abb. 3: Probleme mit der internen Validität des Problemlösegütemaßes bei einem zu schweren Problemlöseszenario

Ein *schlechter* Problemlöser dagegen, der *wenig* verkauft und *hohe* Verluste pro verkauftem „Hemd" hat, wird insgesamt etwa gleich viel Verlust machen. Sein Gesamtverlust ist das *Produkt* aus einer geringen Zahl verkaufter „Hemden" und einem großen Verlust pro verkauftem „Hemd". Beispielsweise kann ein schlechter

Problemlöser, der erstens 1.000 DM Minus pro „Hemd" erwirtschaftet und zweitens auch nicht in der Lage ist, „Hemden" zu verkaufen, also z.B. nur 100 „Hemden" pro Monat „absetzt", exakt den gleichen Erlös erzielen. Wieder erhält man -100.000 DM (-1.000 * 100) als Wert für den erzielten Erlös und somit (über alle Bearbeitungsmonate) als Wert für das Problemlösegütemaß. Das übliche Gütemaß (und auch darauf aufbauende Trendindizes) konnten in der Berliner Erstuntersuchung nicht zwischen guten und schlechten Problemlösern unterscheiden, denn vernünftiges Handeln führte zu genauso großen Verlusten wie unvernünftiges Handeln (siehe Abbildung 3). Solange die Gewinnspanne pro verkauftem „Hemd" negativ ist, also mit jedem verkauften „Hemd" Verlust verbunden ist, ist das zweite Teilziel, den Verkauf zu steigern, kontraproduktiv. Diese Tatsache, *„daß die Erhöhung der Verkaufszahlen bei negativer Gewinnspanne zu größeren Verlusten führt und damit dem Ziel der Kapitalmaximierung widerspricht"* (Schoppek, 1996, S. 30) wurde somit von Süß, Kersting und Oberauer keinesfalls - wie Schoppek irrtümlicherweise (ebd.) schreibt - übergangen, sondern im Gegenteil durch die Aufgabenanalyse erst erkannt. Gute Problemlöser hoben unter solchen Bedingungen mit ihrem Erfolg beim zweiten Teilziel den Erfolg beim ersten Teilziel wieder auf. Problemlöser, die bei beiden Teilzielen erfolgreich waren, hatten im Endergebnis genauso schlecht abgeschnitten wie Problemlöser, die beide Teilziele verfehlt hatten.

Die Annahmen der Aufgabenanalyse ließen sich empirisch bestätigen. Aus den Annahmen folgt, daß die Variable „Gewinnspanne pro verkauftem Hemd" hoch positiv, die Variable „Verkauf" negativ mit dem Gesamtvermögen korrelieren müßte. Genau dies war der Fall.

7.3.2 Definition eines neuen, intern validen Problemlösegütemaßes

Aus den oben genannten Gründen haben Süß et al. (1993b) für die Berliner Erstuntersuchung ein neues Problemlösegütemaß definiert (siehe Abbildung 4). Das Globalziel, die Steigerung des Gesamtvermögens, wurde in zwei Subziele zerlegt: (1) Zahl verkaufter „Hemden" steigern und (2) „Gewinnspanne pro verkauftem Hemd" steigern. Diese beiden Subziele ergeben sich unmittelbar und zwingend aus dem vorgegebenen Ziel, so daß die von Kluwe et al. (1991c, S. 298) sowie Schoppek (1993, S. 30) geäußerten Bedenken hinsichtlich der Übereinstimmung des neuen Gütemaßes mit den Zielvorgaben unbegründet sind. Von einer Dissoziation zwischen der Zielvorgabe und dem neuen Problemlösegütemaß könnte man nur sprechen, wenn die Probanden die unter den spezifischen Bedingungen gegebene Imkompatibilität der Subziele hätten einsehen können. Es ist unwahrscheinlich, daß die Probanden diesen sehr speziellen Umstand erschließen konnten. Mit „gesundem

Menschenverstand“ ist die Mehrheit der Probanden vermutlich davon ausgegangen, daß es sinnvoll ist, bei minimalen Kosten möglichst viele „Hemden“ zu verkaufen.

altes PLG	=	Verkauf	×	Gewinnspanne
neues PLG	=	Verkauf	+	Gewinnspanne

Abb. 4: Altes und neues Problemlösegütemaß (PLG)

Während sich das Gesamtvermögen in der „Schneiderwerkstatt“ im wesentlichen aus dem *Produkt* der Werte dieser beiden Variablen (Zahl verkaufter „Hemden“ und „Gewinnspanne pro verkauftem Hemd“) ergibt, wurden die Werte dieser beiden Teilgütemaße zur Bildung des neuen Problemlösegütemaßes einfach aggregiert. Die Subgütekriterien wurden dabei zunächst *z*-transformiert, wegen ihrer extrem schiefen Verteilung normalisiert und dann *additiv* zu einem neuen Problemlösegütemaß verknüpft. Mit den (normalisierten) Gütekriterien für die beiden Teilziele und dem neuen Gütemaß wurden die Korrelationen zu den Intelligenz-Skalen noch einmal berechnet, diesmal ergaben sich - ebenso wie für die Zusammenhangsanalysen zwischen Problemlöseleistungen und Wissensskalen - substantielle Zusammenhänge (siehe Abschnitt 9.1.2.2).

7.3.3 Empirische Belege für die These, daß das ursprüngliche Problemlösegütemaß unter den gegebenen Bedingungen nicht valide war

Die Aufgabenanalyse und die Zusammenhangsbefunde legten die These nahe, daß es dem ursprünglichen Problemlösegütemaß unter der Bedingung der Überforderung der Probanden an interner Validität mangelte. Als Beleg für die Richtigkeit dieser Annahme kann gelten, daß sowohl die neuen Gütewerte für die einzelnen Subziele als auch die *Summe* der beiden Teilgütewerte mit den Prädiktoren Intelligenz und Wissen korrelierten, das *Produkt* aus einem *positiven* Wert für die Anzahl verkaufter „Hemden“ und einem *negativen* Wert für die Gewinnspanne jedoch mit keinem externen Prädiktor aus dem Bereich kognitiver Leistungsvoraussetzungen korrelierte. Ein stärkerer Beleg für die Richtigkeit der These stellte jedoch die ein Jahr später vorgenommene konzeptionelle Replikation der Untersuchung dar, an der 64% der Probanden der Erstuntersuchung erneut teilnahmen (siehe Süß et al., 1991). Die oben beschriebenen Probleme des Problemlösegütemaßes ergeben sich nicht, wenn es einer nennenswerten Teilstichprobe gelingt, das System in den Bereich positiver Gewinnspanne zu steuern. Es gab verschiedene Möglichkeiten, die Voraussetzungen für diesen Fall zu schaffen: eine Verlängerung der Bearbeitungszeit (Anzahl der „Monate“ (Takte), Anzahl der Bearbeitungsdurchgänge) oder eine Vereinfachung des Problems oder eine Vermittlung zusätzlichen Wissens an die Pro-

banden. Die letzten beiden Wege hat die Berliner Arbeitsgruppe zur Überprüfung der vorgenommenen Aufgabenanalyse und der These der unter diesen Bedingungen mangelnden internen Validität des ursprünglichen Problemlösegütemaßes in der Wiederholungsstudie beschritten (Süß et al., 1991). Zunächst wurde die Programmbedienung erleichtert und gegen Eingriffsfehler abgesichert. Des weiteren wurden die Kosten für „Investitionen" im Verhältnis zu ihrem Nutzen verringert und die nachfragefördernde Wirkung der „Werbung" wurde verstärkt. Außerdem wurde das „Kreditvolumen" in Abhängigkeit vom jeweils aktuellen Wert der Variablen „Gesamtvermögen" begrenzt. Vor allem aber erhielten die Probanden steuerungsrelevante Informationen über die „Schneiderwerkstatt" (z.B. über die Kosten von „Investitionen" und über die „Betriebsorganisation", außerdem wurden alle am Bildschirm sichtbaren Variablen wurden kurz erläutert usw.). Schließlich erhielten die Probanden noch die Möglichkeit, sich während der Steuerung für die Zustandsvariablen die Werte des Vormonats anzeigen zu lassen. Diese Maßnahmen sollten die Steuerung des Systems erleichtern, gleichwohl blieb auch das modifizierte System komplex, intransparent, dynamisch und vernetzt.

In der Wiederholungsuntersuchung konnte mehr als die Hälfte der Probanden das vorgegebene Ziel erreichen und das Gesamtvermögen maximieren. Fast alle Versuchspersonen erzielten diesmal ein besseres Ergebnis als der „Null-Lauf", d.h. als wenn sie über alle Bearbeitungstakte überhaupt keinen Eingriff vorgenommen hätten. Vor allem gelang es den Probanden, eine positive Gewinnspanne zu erzielen. Unter diesen Voraussetzungen korrelierten die Teilgütekriterien „Verkauf" und „Gewinnspanne" positiv mit den herkömmlichen Problemlösegütemaßen. Auch das neu definierte Problemlösegütemaß korrelierte substantiell mit dem herkömmlichen Problemlösegütemaß. Somit können die herkömmlichen und die neu definierten Problemlösegütemaße als äquivalent bezeichnet werden, auch die Kontingenz der Korrelationsmuster der alten und neuen Gütekriterien mit den Außenkriterien unterstützten diese Schlußfolgerung. Die in der Erstuntersuchung vorgenommene Bildung eines neuen Problemlösegütemaßes war aufgrund der Überforderung der Problemlöser also notwendig und gerechtfertigt.

7.4 Zusammenfassung, Schlußfolgerungen und Ausblick

Bei der Diagnostik mit computergestützten Problemlöseszenarien kommt dem - zielgruppenspezifsch festzulegenden - angemessenen Schwierigkeitsgrad und der Steuerbarkeit des Systems eine große Bedeutung zu. Gerade bei den in der Forschung häufig eingesetzten desaströsen Szenarien kann nicht ausgeschlossen werden, daß die Systemkennwerte mehr die Dynamik des Systems als das Verhalten der Steuerer widerspiegeln. Für ein System, bei dem hingegen ein substantieller Teil der Diagnostikanden das vorgegebene Steuerungsziel erreicht, ist zumindestens gezeigt, daß die Kennwerte sich durch das Verhalten der Steuerer beeinflussen lassen, zumindest sofern der „Null-Lauf" nicht zu ähnlich positiven Ergebnissen führt. Die bei Überforderung zu erwartenden Demotivations- und Frustrationsreaktionen der Problemlöser können außerdem den Anforderungscharakter der Aufgaben verändern. Am Beispiel der Berliner Erstuntersuchung mit einer älteren Version der „Schneiderwerkstatt" wurde gezeigt, daß die Validität eines Problemlösegütemaßes von der Schwierigkeit der Steuerungsaufgabe abhängen kann. Dieser Aspekt wird im Kapitel 13 des empirischen Teils der Arbeit wieder aufgegriffen.

Als Fazit der Ausführungen zur Steuerbarkeit und Schwierigkeit der Szenarien ergibt sich ein dringendes Plädoyer dafür, die Szenarien so zu gestalten, daß sie eine individuelle Kontrolle über Art und Inhalt der verlangten bzw. gelieferten Informationen ermöglichen und die Schwierigkeit der Szenarien so abzustimmen, daß ein substantieller Teil der anvisierten Untersuchungsgruppe das System zielgerecht steuern kann.

8. Zur Reliabilität der Messung von Steuerungsleistungen

Zahlreiche Autoren sehen die Problemlöseleistung als Ausdruck eines relevanten Personmerkmals, bezüglich dessen sich Individuen unterscheiden (Eigenschaftsmodell). Der zu überprüfende Ansatz der Diagnose dieses Personmerkmals mit Hilfe computergestützter Problemlöseszenarien geht davon aus, daß die Problemlöseleistung bei der Szenariensteuerung durch zumindest partiell zeit- und situationsstabile, (eignungs-)diagnostisch erfaßbare, Dispositionen erklärt werden kann. Gerade im (eignungs-)diagnostischen Kontext kommt der Reliabilität der Merkmalsindikatoren eine entscheidende Bedeutung bei. Die Bestimmung der Reliabilität von Problemlösegütemaßen ist eine Voraussetzung für die Überprüfung von Zusammenhangsannahmen - z.B. zwischen Steuerungsleistungen an einem komplexen Problem als Prädiktor und dem Berufserfolg als Kriterium. Die Reliabilität der Prädiktoren und Kriterien begrenzt die Höhe der empirisch feststellbaren Zusammenhänge. Die Interpretation der Ergebnisse von Kriteriumsuntersuchungen setzt daher Annahmen über die Reliabilität von Problemlöseleistungen voraus. Das folgende Kapitel gibt einen Überblick über ausgewählte Reliabilitäts-Studien, wobei der Schwerpunkt des Kapitels auf den Bericht von Reliabilitäten von Steuerungsleistungen liegt.

Theoretisch liegt es nahe, daß die Reliabilität von Steuerungsleistungen geringer ist als die anderer Leistungsmessungen, z.B. als die von Intelligenztestleistungen. Die Problemlöseszenarien lassen den Probanden für die Lösung im Vergleich zu Intelligenztestaufgaben mehr Freiheitsgrade, die Steuerungsleistungen sind vermutlich durch zahlreiche heterogene Faktoren bedingt. Aufgrund der Abhängigkeit jedes Systemzustandes von den vorangegangenen können sich zunächst geringe nicht-intendierte Varianzquellen wie Motivationsschwankungen, Ermüdung, Ablenkung durch die Struktur des Systems „aufschaukeln" und über die meist sehr langen Bearbeitungszeiten kumulieren. Die im Vergleich zu Intelligenzaufgaben lange Zeitdauer einer Problembearbeitung bringt daher keine Reliabilitätssteigerung durch Testverlängerung. Jeder Simulationsdurchgang liefert in der Regel lediglich „single act"-Kriterien (Fishbein & Ajzen, 1974) mit geringer Reliabilität. (Siehe aber den weiter unten beschriebenen Ansatz von Müller zur gleichzeitigen parallelen Bearbeitung von unabhängigen und mathematisch strukturäquivalenten Teilsystemen.)

Grundlage empirischer Reliabilitätsbestimmungen sind in der Regel Einheiten, die einer Korrelationsrechnung zugrunde gelegt werden. Dabei werden die Korrela-

tionsrechnungen entweder hinsichtlich unterschiedlicher Meßzeitpunkte (Retest-Reliabilität) oder hinsichtlich unabhängiger Schätzungen des interessierenden Gegenstandes (Paralleltest-Reliabilität bei mehrmaliger Testung; Split-half Reliabilität bei einmaliger Testvorgabe) durchgeführt. Im folgenden werden die Vor- und Nachteile skizziert, die sich bei der Anwendung dieser Methoden zur Reliabilitätsbestimmung von Problemlösegütemaßen ergeben. Außerdem werden zu den einzelnen Ansätzen der Reliabilitätsbestimmung beispielhaft die Ergebnisse einzelner Studien referiert.

8.1 Mehrmalige Vorgabe der Szenarien

Bei der Bestimmung der Wiederholungs- bzw. Retest-Reliabilität wird in der Regel das gleiche Szenario zu zwei verschiedenen Zeitpunkten zur Bearbeitung vorgegeben, bei der Methode zur Bestimmung der Paralleltest-Reliabilität werden zwei äquivalente Formen des Szenarios nacheinander durchgeführt. Bei beiden Arten der Reliabilitätsbestimmung können grundsätzlich Wiederholungseffekte zu Fehleinschätzungen der Reliabilität führen, diese Gefahr droht aber besonders bei der Methode der Wiederholungs- bzw. Retest-Reliabilität. Mögliche Wiederholungseffekte können z.B. durch die Übung, durch die Vertrautheit mit der Testsituation (etwa die Vertrautheit mit der Computer-Bedienung), durch Sättigung und Problemeinsicht hervorgerufen werden. U. Funke (1995, S. 178) und H. Müller (1993, S.26 und 28) haben darauf aufmerksam gemacht, daß auf (aktualitätsbedingte) Verhaltensweisen, die nur bei der Erststeuerung notwendig sind (z.B. Informationsabfragen) bei den Wiederholungsdurchgängen verzichtet wird, so daß die mit der Retest-Methode erfasste Reliabilität unter Umständen unterschätzt wird. Andererseits weist U. Funke (ebd.) darauf hin, daß bei wiederholten Steuerungen eventuell auch ein Lernplateau erreicht wird, infolge dessen die Reliabilität mit der Wiederholungsmethode überschätzt würde. Bei mehrmaliger Vorgabe der Szenarien besteht auch die Gefahr, daß Versuchspersonen ihr Verhalten des ersten Steuerungsdurchgangs - sofern sie sich daran erinnern können - schlicht wiederholen. Vor diesem Hintergrund ist vor allem die Praxis, Systeme mit identischen Startwerten mehrfach vorzugeben, zu überdenken. Alternativ dazu können Variationen der Initialbedingungen vorgenommen werden (siehe z.B. Kluwe et al. 1990; Süß et al., 1991).[9]

[9] Sofern Variationen der Initialbedingungen vorgenommen werden, bestehen Interpretationsfreiräume hinsichtlich der Frage, ob es sich um ein Retest- oder um ein Paralleltestdesign handelt. Diese Frage läßt sich nicht immer - wie bei der Berliner Studie (Süß, Kersting & Oberauer, 1991), bei der die Erst- und Wiederholungsuntersuchung im Abstand von einem Jahr durchgeführt wurden -...(Fortsetzung Seite 103)

Wiederholungseffekte drohen insbesondere bei ausgiebigem Repetieren und bei sehr kurzen Zeitintervallen zwischen den Problembearbeitungen. So steuern die Versuchspersonen bei Schoppek (1991) und Schmuck (1992) beispielsweise gleich eine Serie von fünf Durchgängen des Szenarios „Feuer" (siehe unten).

Grundsätzlich besteht bei mehrmaliger Vorgabe der Szenarien das Problem, daß Meßwertdifferenzen zwischen den Meßzeitpunkten nicht nur auf Fehlereffekte, sondern auch auf interindividuell unterschiedliche „echte" Veränderungen, z.B. Lerneffekte zurückgeführt werden können.

8.1.1 Studien zur Bestimmung der Retest-Reliabilität

Funke (1983) konnte 14 ausgewählte Probanden - aus einer ursprünglichen Stichprobe von 53 Personen - gewinnen, nach einem Interval von zwei Wochen, die „Schneiderwerkstatt" erneut zu bearbeiten. Dabei ergaben sich für das - inhaltlich wenig sinnvolle - Trendmaß der Anzahl an Monaten mit Aufwärtstrend im Flüssigkapital („Trendpo") eine Rangkorrelation zwischen dem ersten und dem zweiten Trendwert von $r=.20$, während sich für das angemessenere Trendmaß der Anzahl an Monaten mit Aufwärtstrend im Gesamtkapital („Trendfu") ein entsprechender Wert von $r=.80$ ergab.

Strohschneider (1986) ließ 25 Studenten das „Moro"-System im Abstand von 11-55 Tagen wiederholt steuern. Aufgrund der über einen Versuchsleiter vermittelten Systemsteuerung ist die (Durchführungs-)Objektivität dieser Studie ungeklärt. Während Strohschneider für Verhaltensmaße recht hohe Retest-Korrelationen (von $r=.33$ bis $r=.85$) berechnete, ergaben sich für ein kombiniertes und am Null-Lauf relativiertes Maß der Steuerungsleistung (Systemzustand auf der Grundlage von sieben ausgewählten Variablen) geringere - überwiegend nicht-signifikante - Werte zwi-

---

[9] *Fortsetzung Fußnote 9 von Seite 102*: ... aufgrund des Zeitabstands zwischen den Messungen entscheiden. Der Interpretationsspielraum gilt umso mehr, da sich in einschlägigen Methodenbüchern zur Frage des Zeitabstands zwischen der ersten und der wiederholten Vorgabe der Tests beim Ansatz der Paralleltest-Reliabilitätsmethode inkonsistente Angaben finden. Während Heidenreich (1995, S. 355) für Studien zur Bestimmung der Paralleltest-Reliabilitäten z.B. eine Intervallzeit von mehreren Tagen empfiehlt, gehen Bortz und Döring (1995, S. 182) davon aus, daß die Paralleltests kurz hintereinander bearbeitet werden. Unmittelbar aufeinanderfolgende Wiederholungen der Szenariensteuerung können schließlich auch als Testverlängerung interpretiert werden, diese Interpretation findet in der Aggregation der Ergebnisse einzelner Steuerungsdurchgänge ihren praktischen Ausdruck. Die Zeitabstände zwischen den im Kontext der Problemlöseforschung durchgeführten Reliabilitätsstudien sind oft so kurz, daß auch die dem Retest-Design zugeschlagenen Studien allein vom Zeitintervall her noch dem Paralleltest-Design entsprechen.

schen r=.26 und r=.44 für fünf diskrete Zeitpunkte. Diese unterschiedlichenReliabilitätswerte für Verhaltensmaße einerseits und Steuerungsleistungen andererseits sind ein weiterer Indikator für die oben (Abschnitt 6.3.2.3) beschriebene Dissoziation zwischen diesen beiden Arten von Problemlösegütemaßen. Die niedrigen Reliabilitätswerte für die Steuerungsleistung sind vor dem Hintergrund zu betrachten, daß die Versuchspersonen in ihren Zielen nicht reglementiert wurden. Es ist denkbar, daß die Probanden das Szenario jedesmal mit einer anderen Zielsetzung steuerten; befriedigende Reliabilitäten sind eher bei Szenarien-Steuerungen mit eindeutiger Zielvorgabe zu erwarten.

In einer Retest-Studie von Hasselmann und Strauß (1988; berichtet nach Hasselmann, 1993) mit dem System „Textilfabrik" wurden 52 Studenten untersucht, die das System im Abstand von 14 Tagen zweimal steuerten - wobei das System mit jeweils identischen Systemstartwerten vorgegeben wurde. Für das oben bereits beschriebene Problemlösegütemaß „Trendfu" ergab sich für die 11 Studenten der Betriebswirtschaftslehre eine Retest-Reliabilität von r=.71 und für die übrigen Studenten ein Wert von r=.73. Für ein Problemlösegütemaß, welches die Entwicklung von vier wichtigen Systemvariablen berücksichtigt, ergaben sich entsprechende Werte in Höhe von r=.74 (.61) (in Klammern die Werte für die Studenten der Betriebswirtschaftslehre), für ein Maß der relativen Häufigkeit des Anstiegs des Kapitalwerts über unterschiedliche gewichtete Zeitabschnitte ergaben sich Werte in Höhe von r=.70 (.63), und für den Kapitalendwert ergaben sich Werte von r=.47 (.43).

Für das Maß „Trendfu" wurden auf Grundlage der Gesamtstichprobe von 52 Personen zusätzlich Reliabilitätsberechnungen für vier Phasen (á fünf Takten) berechnet. Während die Trendmaße für das erste Viertel des ersten und des zweiten Steuerungsdurchgangs lediglich zu r=.09 korrelierten, zeigten sich für die weiteren Viertel signifikante Korrelationen in Höhe von r=.63, r=.56 und r=.52. Die Autoren schließen daraus, daß eine stabile Messung erst nach einer längeren Problembearbietung erwartet werden kann. Der gewählte Ansatz ist allerdings problematisch, da es sich bei den einzelnen „Takten" um abhängige Messungen handelt. Es wurden noch weitere Problemlösegütemaße gebildet und hinsichtlich ihrer Retest-Reliabilität geprüft, mit den neuen Maßen wurden aber keine Reliabilitäten erzielt, die bedeutsam über der Reliabilität für das Problemlösegütemaß „Trendfu" lagen.

Für eine Retest-Studie mit dem Szenario „Heizölhandel" berichtet Hasselmann (1993) Test-Retest-Korrelationen zwischen r=.67 und r=.74 für die unterschiedlichen Maße der Steuerungsleistung. Versuchspersonen waren diesmal 14 Studenten, welche die Szenarien im Abstand von zwei Wochen wiederholt steuerten. Locher (1997) ließ 20 Teilnehmer ebenfalls das Szenario „Heizölhandel" nach einem Tag erneut steuern. Für das Gütemaß „Anzahl der Monate mit Gewinn" berichtet der Autor eine Test-Retest-Korrelation in Höhe von r=.58, für das Maß „Kapitalend-

wert" betrug der entsprechende Wert hingegen lediglich r=.12 (ebd., Anhang S. 4).

Bei Schoppek (1991) bearbeiteten 22 Studenten der Wirtschaftswissenschaften in fünf aufeinanderfolgenden (mit Ausnahme des ersten und letzten Durchgangs unterschiedlich schwierigen) Durchgängen das Szenario „Feuer". Die Endergebnisse jedes Durchgangs korrelierten zwischen r=.44 und r=.82 mit der jeweils nachfolgenden Szenarienbearbeitung. Das Endergebnis der ersten Steuerung korrelierte zu r=.42 mit dem Endergebnis der letzten Steuerung. Statistisch bedeutsame Korrelationen in unterschiedlicher Höhe zwischen den in den verschiedenen Durchgängen gewonnenen Maße zeigten sich auch für die Mehrheit der Verhaltensindikatoren.

Ebenfalls fünfmal nacheinander wurde das System „Feuer" bei Schmuck (1992) von 59 Studenten bearbeitet. Schmuck (ebd.) berichtet u.a. über die Stabilität bestimmter Verhaltensweisen, nicht aber über die Stabilität der Steuerungsleistung. Für die Hälfte der analysierten Verhaltensindikatoren ergab sich in den letzten Durchgängen eine statistisch bedeutsam höhere Stabilität (Korrelationen zwischen dem vierten und fünften Durchgang von r=.55 bis r=.97) als in den ersten Durchgängen (Korrelationen zwischen dem ersten und zweiten Durchgang von r=.33 bis r=.84). Diese „Stabilitäts-Zunahme" wurde dann noch hinsichtlich der „Kontrolleffizienz" der Versuchsteilnehmer differenziert: Aufgrund eines Vortests als „flexibel" eingestufte Versuchspersonen erzielten bei drei von acht Verhaltensmaßen in den ersten Durchgängen einer Serie statistisch bedeutsam geringere Test-Retest-Korrelationen als weniger flexible Personen. Geringe Reliabilitäten von Problemlösegütemaßen sind Schmuck zufolge möglicherweise auch auf das flexible (und somit instabile) Verhalten einer Teilstichprobe zurückzuführen. Schmuck hat seine Annahmen zu differentiellen Aspekten der Verhaltensstabilität beim Problemlösen aber explizit für Verhaltensindikatoren formuliert. Für Steuerungsindikatoren müßten die Befunde anders aussehen, da auch flexible Personen bei eindeutiger Zielvorgabe an einem kontinuierlichen, stabilen Erfolg interessiert sein müssen - auch wenn sie diesen Erfolg möglicherweise durch jeweils andere Verhaltensweisen herbeiführen.

Die Untersuchung von Schmuck (1992) wurde später mit 61 Auszubildenden konzeptionell repliziert (Schmuck & Strohschneider, 1995), wobei die Versuchsbedingungen allerdings deutlich geändert wurden. So kam z.B. mit „Moro" ein anderes Szenario zum Einsatz, welches ohne eindeutige Zielvorgabe gesteuert wurde. Die Stabilitätskennwerte wurden hier aus den vier Phasen (je fünf Takte) *einer Steuerung* berechnet. Da diese Bearbeitungszeiträume unmittelbar aufeinander aufbauen, können diese Werte - wie die Autoren ausdrücklich schreiben (S. 159) - nicht zur Reliabilitätsschätzung verwendet werden. Die Studie findet hier Erwähnung, da auch diese Arbeit Anhaltspunkte dafür liefert, daß Verhaltensstabilität beim Bearbeiten komplexer Probleme erst nach Aufbau einer Problemrepräsentation erwartet werden kann. Demzufolge kann sich erst bei einem längeren Problembear-

beitungsverlauf oder bei mehreren Problembearbeitungen Merkmalskonstanz zeigen, geringe Reliabilitäten sind möglicherweise (auch) auf zu kurze Bearbeitungszeiten (und/oder mangelnde Übungsgelegenheiten) zurückzuführen.

Schoppek (1996, S.132) ließ 31 Personen das Szenario „Jogi'91" zweimal - mit unterschiedlichen Startwerten - steuern. Für zwei Maße der Steuerungsleistung betrug die Korrelation zwischen den beiden Durchgängen r=.38 bzw. r=.53.

Putz-Osterloh (1991a) berichtet über einen wiederholten Einsatz des Systems „Feuer". Bei 50 Studenten wurde das System in unterschiedlichen Versionen in zwei Sitzungen vorgegeben. Pro Sitzung mußte das Szenario (in identischer Version) dreimal gesteuert werden. In einer anderen Untersuchung mit 80 studentischen Versuchspersonen wurde eine Version dieses Szenarios viermal hintereinander gesteuert. Bei der Untersuchung mit 50 Versuchspersonen korrelierten die Leistungsmaße des zweiten mit dem dritten Durchgangs zwischen r=.60 und .89, während die Verhaltensmaße zwischen r=.74 und r=.88 miteinander korrelierten. (Ein Bericht über die Stabilitätsschätzungen leicht abgewandelter Verhaltensmaße dieser Studie findet sich bei Putz-Osterloh, 1989, S. 96f.) In der anderen Studie mit 80 Teilnehmern ergab sich für das Leistungsmaß eine Korrelation von r=.84 und für die Verhaltensmaße Korrelationen von r=.76 und r=.79 zwischen dem dritten und dem vierten Durchgang. Vergleichsweise niedrigere Werte berichten Putz-Osterloh und Haupts (1990) aus einer weiteren Studie mit 30 Probanden, in der ebenfalls das Szenario „Feuer" eingesetzt wurde. Hier ergaben sich Retest-Korrelationen von r=.48 für die Steuerungsleistung und r=.45 und r=.52 für die Verhaltensmaße.

In der Berliner Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen (Süß et al.1991, 1993b; Süß, Kersting und Oberauer 1993a; Süß, 1996) steuerten die 214 Schüler der Erstuntersuchung das System „Schneiderwerkstatt" einmal für 12 Takte und anschließend zweimal für sechs Takte - jeweils mit veränderten Startwerten. Während sich für die herkömmlichen Problemlösegütemaße, die unter den spezifischen Bedingungen der Untersuchung keine interne Validität aufwiesen (siehe oben, Abschnitt 7.3.3), Interkorrelationen über die drei Bearbeitungsdurchgänge von r=.37 bis r=.53 ergaben, betrug die Korrelation zwischen der ersten Steuerung und dem Aggregat aus den beiden folgenden Durchgängen für ein neu definiertes - intern valides - Problemlösegütemaß (siehe oben, Abschnitt 7.3.2) r=.63. Von den 214 Teilnehmern der Erstuntersuchung nahmen 137 Schüler ein Jahr später an der Wiederholungsuntersuchung teil. Diese Teilnehmer übertrafen die Nicht-Teilnehmer in der Intelligenzdimension „Verarbeitungskapazität". Diesmal steuerten die Probanden die geringfügig modifizierte (siehe oben, Abschnitt 7.3.3) „Schneiderwerkstatt" zweimal, jeweils über 12 Takte, erneut mit jeweils anderen Initialbedingungen. Die in diesen beiden Durchgängen der Wiederholungsuntersuchung mit Hilfe des neuen Problemlösegütemaßes bestimmten Steuerungsleistungen

korrelierten zu r=.74 miteinander, die mit Hilfe des Endwerts des Gesamtvermögens bestimmten Steuerungsleistungen zu r=.70. Die Stabilität der Steuerungsleistung, berechnet als die Korrelation zwischen den aggregierten Steuerungsleistungen aus der Erst- und der Wiederholungsuntersuchung, betrug r=.46. Diese Stabilität bricht allerdings zusammen (r=.13), wenn man wie Süß (1996, S. 194f.) den Anteil der systematischen Varianz auspartialisiert, der auf Intelligenz- und Wissensleistungen zurückzuführen ist.

Die Retest-Reliabilität des Szenarios „DISKo" beträgt laut U. Funke (1993) für das Maß „Gesamtsumme" r=.54. Grundlage dieser Berechnung waren die Daten von 10 Personen, die das Szenario in einer Sitzung mehrmals steuerten. Für die Steuerungsleistung als Aggregat des Gesamtkapitalanstiegs und dem Gesamtkapital ergab sich bei 13 Personen eine Test-Retest-Korrelation von r=.53, für das Verhaltensmaß (Aggregat aus fünf Indizes) ein entsprechender Wert von r=.63.

8.1.2 Studien zur Bestimmung der Parallel-Test-Reliabilität

Kluwe et al. (1991c, S. 302) bestreiten grundsätzlich die Möglichkeit, parallele Formen für ein komplexes dynamisches System so zu entwickeln, daß zwei äquivalente Formen resultieren, da schon geringfügige Änderungen unterschiedliche Problemrepräsentationen bewirken würden. Tatsächlich eröffnet der Begriff Äquivalenz Interpretationsfreiräume, die Anforderungen an die „Ähnlichkeit" müssen bei parallelen Messungen aber anders formuliert werden als z.B. bei tau-äquivalenten oder kongenerischen Messungen (siehe z.B. Fischer, 1974, S. 32 ff.).

Die oben aufgelisteten Studien zur wiederholten Steuerung von Szenarien mit veränderten Initialwerten können je nach Interpretation entweder als Studien zur Bestimmung der Paralleltest-Reliabilität oder als Studien zur Bestimmung der Meßwiederholungs-Reliabilität eingeordnet werden. Ebenso können Studien mit ähnlichen Szenarien entweder als Parallelitäts-Studien oder aber als Generalitätsstudien (siehe unten, Abschnitt 9.1.1) bezeichnet werden. So berichten Köller, Dauenheimer und Strauß (1993) sowie Hasselmann (1993) über eine Untersuchung, in der 22 Studenten sowohl das Szenario „Textilfabrik" als auch das in enger Anlehnung an die „Textilfabrik" konstruierte Szenario „Heizölhandel" steuerten. Zwischen den beiden Bearbeitungen lagen im Durchschnitt sieben Wochen. Dabei ergaben sich Interkorrelationen zwischen den drei Maßen der Steuerungsleistungen für die beiden Szenarien in Höhe von r=.42 bis r=.51. Berichtet wird (Hasselmann, 1993) auch über den Zusammenhang zwischen den Indikatoren der Steuerungsleistung im Szenario „Textilfabrik" und im Szenario „Reifenhandel". Das Szenario „Reifenhandel" weist zwar einen deutlich reduzierten Problemumfang auf als das Szenario „Textilfabrik",

gleichwohl bestehen auch Gemeinsamkeiten zwischen den beiden Systemen. Das Szenario „Reifenhandel" wurde in drei (unterschiedlich schwierigen) Varianten eingesetzt, jeweils 10 Studenten bearbeiteten zunächst die „Textilfabrik" und dann den „Reifenhandel". Nur bei einer (relativ leichten) Version des „Reifenhandels" korrelierten alle drei Maße der Steuerungsleistung signifikant mit den entsprechenden Maßen der Steuerungsleistung der „Textilfabrik" (Korrelationen zwischen r=.43 und r=.51). Drei Szenarien dieser Konstruktionsfamilie, nämlich die „Textilfabrik", der „Heizölhandel" und der „Brennstoffvertrieb" wurden von Köller, Strauß und Sievers (1995) zu drei Meßzeitpunkten im zeitlichen Abstand von zirka einer Woche bei 60 studentischen Probanden eingesetzt. Hinsichtlich des Gütemaßes „Anzahl der Simulationstakte mit Kapitalanstieg" zeigte sich für die Szenarien „Heizölhandel" und „Brennstoffvertrieb" mit r=.69 ein engerer Zusammenhang als für das Szenario „Textilfabrik" mit den übrigen beiden Szenarien (r=.41 und r=.44).

U. Funke (1995a) konstruierte eine Parallelversion („Videofabrik") zum Szenario „DISKo". Beide Versionen wurden 14 Teilnehmern als die deutsche und französische Niederlassung eines Konzerns simultan vorgegeben, der Wechsel zwischen den Szenarien erfolgte durch Tastendruck. Für die Steuerungsleistung übertreffen die Paralleltest-Reliabilitäten mit r=.83 die Retest-Reliabilitäten von r=.53. Entsprechendes galt für die Verhaltensmaße (r=.88 zu r=.63).

8.2 Einmalige Vorgabe der Szenarien: Studien zur Bestimmung der Halbierungs-Reliabilität

Als Sonderform der Paralleltest-Methode bei einmaliger Vorgabe kann die Testhalbierungs-Methode aufgefaßt werden (Bortz & Döring, 1995, S. 183). Die oben benannten Einwände gegen die Verwendbarkeit der Paralleltest-Methode bei computergestützten Problemlöseszenarien gelten daher prinzipiell auch hier. Hinzu kommt, daß die für die Halbierungs-Methode vorausgesetzte Zerlegung des Systems in zwei gleichwertige Hälften bei computergestützten Problemlöseszenarien - in denen das Verhalten der Steuerer in einem Bearbeitungstakt stets abhängig ist von seinem Verhalten im vorherigen Takt - besonders schwierig erscheint.

Müller (1993, S. 65 f.) löste dieses Problem, indem er ein dynamisches System konstruierte, welches sich aus zwei unabhängigen und mathematisch strukturäquivalenten Teilsystemen zusammensetzte, die aber bei gleichzeitiger paralleler Bearbeitung als Gesamtsystem kombiniert und präsentiert wurden. Für insgesamt 105 Studenten und Studenten lagen verwertbare Daten vor. Die in den beiden Teilsystemen erzielten Steuerungsleistungen korrelierten zu r=.86 (Test) und r=.83 (Retest) mit-

einander. Demgegenüber war die Test-Retest Reliabilität, die Müller aufgrund einer Meßwiederholung nach fünf Monaten bestimmte, mit r=.53 deutlich niedriger.

8.3 Zusammenfassung, Schlußfolgerungen und Ausblick

Die bei der Steuerung computergestützter Problemlöseszenarien erzielten Ergebnisse lassen sich nur dann verantwortbar und sinnvoll diagnostisch verwenden, wenn die Zuverlässigkeit der Messung geklärt ist. Auch Untersuchungen zur Kriterien- und Konstruktvalidität setzen die Kenntnis der Reliabilität voraus, da die Reliabilität dem empirischen Zusammenhang und der Vorhersagbarkeit grundsätzlich Grenzen setzt. Angesichts dieser zentralen Bedeutung der Reliabilitätsfrage muß die Anzahl (und teilweise auch Qualität, z.B. Stichprobengröße, zu geringer Zeitabstand bei Stabilitätsprüfungen usw.) der vorhandenen empirischen Studien als unzureichend gewertet werden. Die referierten Studien zeichnen sich durch eine breite Variation hinsichtlich der verwendeten Szenarien und Darbietungsbedingungen (mit und ohne Variation der Initialwerte) aus. Dies führt - ebenso wie die unterschiedlichen Untersuchungsgruppen (von zum Teil beklagenswert geringem Umfang) sowie die unterschiedlichen Problemlösegütemaße - zu einem inkonsistenten Bild der Reliabilität von Problemlöseleistungen. Hinzu kommen noch die Variationen, die auf die unterschiedlichen Reliabilitäts-Bestimmungsmethoden zurückzuführen sind. Diese inkonsistente Datenlage vereitelt eine Verallgemeinerung der Ergebnisse. Die Reliabilitäts-Studien stammen häufig aus dem grundlagenwissenschaftlichen Bereich. Nur für zwei („DISKo" sowie die Szenarien aus der Verfahrensfamilie „Textilfabrik") explizit anwendungsorientierte Szenarien liegen überhaupt Reliabilitätsprüfungen vor. Die übrigen anwendungsorientierten Szenarien erfüllen damit noch nicht einmal eine der zentralen Voraussetzungen für einen diagnostischen Einsatz. (Der Hinweis auf nicht-veröffentlichte Reliabilitätsprüfungen leistet hier ebensowenig Abhilfe wie Veröffentlichungen, in denen alle wesentlichen Informationen über die angestellten Untersuchung fehlen.)

Festgehalten werden kann, daß befriedigende Reliabilitäten nur für intern valide Problemlösegütemaße zu erwarten sind. So zeigte sich etwa für das Trendmaß der Anzahl an Monaten mit Aufwärtstrend im Flüssigkapital („Trendpo") - welches die tatsächliche Entwicklung des Systems „Schneiderwerkstatt" nur unzureichend abbildet - eine deutlich geringere Reliabilität als für das intern validere Trendmaß der Anzahl an Monaten mit Aufwärtstrend im Gesamtkapital („Trendfu"). Auch in der Berliner Erstuntersuchung variierte die Höhe der Reliabilität in Abhängigkeit von der internen Validität des Problemlösegütemaßes. Reliable Messungen setzen außer-

dem eine klare Zielvorgabe voraus. Ein stabile Messung ist schließlich mit großer Wahrscheinlichkeit erst nach einer längeren Problembearbeitung und dem damit einhergehenden Aufbau einer Problemrepräsentation zu erwarten. Aus der diagnostischen Perspektive sind daher ausreichend lange Bearbeitungszeiten, ausführliche Übungsgelegenheiten sowie ggf. die Aggregation mehrerer Bearbeitungsdurchgänge mit wechselnden Initialwerten (siehe hierzu die multivariate Reliabilitätstheorie von Wittmann (1988)) zu empfehlen.

Den referierten Untersuchungen zufolge erfüllen zumindest einige spezifische Problemlösegütemaße bestimmter Szenarien trotz der ungünstigen theoretischen Voraussetzungen - wie z.B. der multiplen Leistungsbedingtheit, den Lerneffekten und dem niedrigen Aggregationsniveau - die für einen diagnostischen Einsatz zu stellenden Reliabilitätsanforderungen. Laut U. Funke (1995a, S. 189 f) liegt die Zuverlässigkeit von Problemlöseleistungen auf der gleichen Ebene wie die Zuverlässigkeit sogenannter „simulationsorientierter" Verfahren (z.B. Gruppendiskussionen). Der Vergleich mit Intelligenztestleistungen fällt hinsichtlich der Reliabilität hingegen zuungunsten der Problemlöseleistungen aus.

9. Validität der Messung von Steuerungsleistungen

Entscheidend für den diagnostischen Einsatz computergestützter Problemlöseszenarien ist die Frage der Validität, wobei es nicht um die Validität der Verfahren, sondern um die Validität der *Interpretation* der Daten geht, die mit dem Verfahren gewonnen werden (Cronbach, 1971, S. 447). Aus dieser Validitätsauffassung folgt, daß sich die Validitätsfrage auch im Kontext der Eignungsdiagnostik nicht auf den Nachweis der Kriteriumsvalidierung beschränken läßt, sofern Diagnostik nicht zur *„blinden Technologie"* degradiert werden soll (Jäger, 1986, S. 284). Zentrale diagnostische Fragen bleiben unberührt, wenn man lediglich (signifikante) Zusammenhangsmuster zwischen Prädiktoren und Kriterien betrachtet. Um die Validität eines Verfahrens erschließen zu können (denn die Validität wird nicht gemessen, sondern erschlossen, siehe Jäger, ebd., S. 272), bedarf es Annahmen darüber, was der Prädiktor mißt. Messick (1988, S. 37) hat dies mit eingänglichen Beispielen verdeutlicht. So ergibt sich z.B. eine andere Bewertung des nachweislichen Kriterien-Prädiktor Zusammenhangs, je nachdem, ob man ein und denselben Testwert als Ausprägung auf der Dimension „Flexibilität- Rigidität" oder als Ausprägung der Dimension „Konfusion- Kontrolle" interpretiert. Auch die Frage, ob man die Kriterien als solche akzeptiert, kann erst aufgrund der Interpretation der gemessenen Leistungsdimension entschieden werden. Sofern differentielle Kriteriumsvaliditäten für einzelne Gruppen auftreten, kann schließlich ebenfalls nur vor dem Hintergrund der Konstruktannahmen entschieden werden, ob es sich um eine wünschenswerte Differenzierung oder um einen unerwünschten „bias" handelt. Daten zur Kriteriumsvalidierung wären von zweifelhaften Wert, wenn ihr Geltungsanspruch sich ausschließlich auf die den Daten zugrundeliegende Situation beziehen würde. Der gängigen Interpretation von Kriteriumsvaliditäten liegen zumeist Annahmen zur zeitlichen Stabilität des prognostizierten und prognostizierenden Merkmals sowie Annahmen zur Situationsinvarianz - und somit Konstruktannahmen - zugrunde. Ohne Konstruktannahmen wären Generalisierungen von Kriteriumsvaliditäten und somit Absicherungen gegen bestimmte Arten von Stichproben- und Meßfehler (siehe z.B. F.L. Schmidt, 1988, 1992) zumindest deutlich erschwert.

Es ist daher auch im eignungsdiagnostischen Kontext nicht sinnvoll, die Validitätsfrage auf eine bloße Betrachtung der Prädiktor-Kriteriums-Zusammenhänge zu amputieren. Validität umfasst stets alle Validitätsaspekte. In dem vorliegenden

Kapitel sollen daher in jeweils eigenständigen Abschnitten sowohl Aspekte der Konstrukt- als auch Aspekte der Kriteriumsvalidität betrachtet werden. Der Aspekt der Kontentvalidität entfällt, da computergestützte Problemlöseszenarien - wie in den Kapitel 4 zur Simulationsfrage und zur Frage der ökologischen Validität gezeigt wurde - keinen begründeten Anspruch auf Kontentvalidität stellen können.

9.1 Konstruktvalidität

Der eignungsdiagnostischer Einsatz von computergestützten Problemlöseszenarien geht implizit oder explizit davon aus, daß sich in dem beobachtbaren Verhalten der Personen beim Umgang mit Problemlöseaufgaben eine latente - nicht unmittelbar beobachtbare - Fähigkeit manifestiert, so daß die aus der Aufgabenbearbeitung abgeleiteten Problemlösegütemaße Indikatoren für diese Fähigkeit darstellen. Wenn dem so ist, sollten die mit Hilfe verschiedener Problemlöseszenarien gewonnenen empirischen Indikatoren desselben Konstrukts miteinander korrelieren. Diesem Aspekt der Konstruktvalidierung wird im ersten Abschnitt (Abschnitt 9.1.1) zur Generalisierbarkeit der Problemlösefähigkeit nachgegangen.

In den anschließenden Abschnitten (Abschnitt 9.1.2 bis 9.1.5) wird dann versucht zu klären, um was für eine Fähigkeit es sich beim „Problemlösen" überhaupt handeln soll. Bislang existieren vor allem phantasievolle Benennungen der postulierten Fähigkeit, so listet U. Funke (1992, S. I-15 f.) beispielsweise die folgenden neun Bezeichnungen auf: »heuristische Kompetenz«, »komplexe Problemlösefähigkeit«, »Fähigkeit zum Umgang mit komplexen, vernetzten Systemen«, »Problemlösekompetenz«, »operative Intelligenz«, »Fähigkeit zu vernetzten Denken«, »Fähigkeit zu systemischen Denken«, »Fähigkeit zu ganzheitlichen Denken«, »Fähigkeit zu systemadäquatem Handeln«. Die Namensschöpfungen können aber - ebensowenig wie der Bericht von „miscellaneous correlations" (Cronbach, 1989, S. 155) - kein Ersatz für eine trennscharfe Einordnung der Fähigkeit in das bestehende nomologische Netz sein. Was hat die „Problemlösefähigkeit" mit etablierten Konstrukten gemeinsam, was unterscheidet sie von diesen? Weiter oben - siehe Abschnitt 2.3 und Abbildung 1 - wurde bereits dargestellt, daß die Fähigkeit zum Problemlösen theoretisch bislang mit drei Klassen von etablierten Personmerkmalen in Verbindung gebracht wurde, nämlich mit *kognitiven Merkmalen* (z.B. Wissen und Intelligenz), mit *emotionalen und motivationalen Merkmalen* sowie mit *Persönlichkeitsmerkmalen im engeren Sinne* (z.B. Selbstsicherheit, Ängstlichkeit usw.). Die vorliegende Arbeit fokussiert den Aspekt der kognitiven Merkmale. Da die Problemlöseszenarien sich in der praktischen Diagnostik de facto als eine Alternative zu herkömmlichen Intelli-

genztests verstehen (siehe Kapitel 3), werden insbesondere die Gemeinsamkeiten und Unterschiede von intellektuellen Leistungen in herkömmlichen Intelligenztests und Problemlöseleistungen bei der Steuerung computergestützter Problemlöseszenarien betrachtet. Diese Perspektive wird dann - wenngleich schon weniger ausführlich - um das Thema „Wissen und Problemlösen" erweitert. Während Intelligenz und Wissen zunächst separat in ihrem Verhältnis zum Problemlösen betrachtet werden, folgt anschließend ein kurzer Abschnitt zur Assoziation des Problemlösens einerseits mit einer Einheit aus Intelligenz und Wissen andererseits. Abschließend wird der Zusammenhang der Problemlösefähigkeiten mit emotional/motivationalen Personmerkmalen sowie mit Persönlichkeitsmerkmalen im engeren Sinne skizziert.

9.1.1 Zur Generalität der Problemlösefähigkeit

Zur Konstruktvalidierung gehört die Klärung der Generalitätsfrage. Die *Generalität* der postulierten Problemlösefähigkeit läßt sich u.a. dadurch prüfen, daß die Messung der interessierenden Fähigkeit mit unterschiedlichen Aufgaben, hier also mit unterschiedlichen Problemlöseszenarien, vorgenommen wird. Sofern die Meßinstrumente Anforderungen an ein und die gleiche Fähigkeit stellen, müssen die verschiedenen Messungen der Problemlösefähigkeit substantiell miteinander korrelieren. Die Höhe dieser Korrelationen könnte dabei in Abhängigkeit von der Ähnlichkeit der inhaltlichen und formalen Aufgabenmerkmale (siehe oben, Abschnitte 2.3.2.2 und 2.3.2.3) der in den Generalitätsstudien eingesetzten Szenarien variieren. Insgesamt sind Korrelationen zu erwarten, die etwas unterhalb der Reliabilität der einzelnen Messungen liegen. Für das Intelligenzkonstrukt berichtet z.B. Jensen (1984, S. 570) eine durchschnittliche Interkorrelation von $r=.77$ (bzw. $r=.86$ nach Attenuationskorrektur) zwischen 30 verschiedenen Intelligenztestverfahren. Aufgrund der ungünstigeren Reliabilität (siehe oben, Kapitel 8) ist zwar davon auszugehen, daß die Leistungen in unterschiedlichen Problemlöseszenarien vergleichsweise geringer miteinander korreliert sind als Intelligenztests, die im folgenden referierten empirischen Befunde stellen aber eine Enttäuschung selbst niedrigster Erwartungen dar.

9.1.1.1 Empirische Befunde zur Generalität der Problemlösefähigkeit

Strohschneider (1990, S. 217 ff; 1991a, S. 365 f.) berichtet über 20 Versuchspersonen, die sowohl das abstrakte „Vektor"-System als auch das semantisch eingekleidete „Moro"- System bearbeiteten. Weder die für beide Szenarien bestimmte Güte der Sollwertabweichung noch die Güte der Systemstabilisierung standen in einem statistisch überzufälligen Zusammenhang miteinander. Hinsichtlich der Sollwertab-

weichung waren die Varianzanteile, die die jeweiligen Werte pro System mit der Intelligenztestleistung gemeinsam hatten, sogar größer als die gemeinsamen Varianzanteile zwischen den Systemen. Auch die analysierten Handlungsstile und -strategien ließen sich nicht über die beiden Systeme hinweg generalisieren.

Die von Putz-Osterloh und Haupts (1989 und 1990) durchgeführten Versuche mit den Szenarien „Feuer" und „Schneiderwerkstatt" sprechen ebenfalls gegen die Generalisierbarkeit der Problemlöseleistungen. In diesen Experimenten gab es keinen positiven Transfer zwischen den Systemen „Schneiderwerkstatt" und „Feuer", solange die „Schneiderwerkstatt" zuerst gesteuert wurde. Die Autorinnen geben daher zu bedenken, daß die beiden Systeme eher unterschiedliche als übereinstimmende Anforderungen stellen (1990, S. 129).

Auch zwischen den Gütewerten der Szenarien „Moro" und „Schneiderwerkstatt" ist der Anteil gemeinsamer Varianz „- *wenn überhaupt - gering*", wie Putz-Osterloh (1987) aus der bereits mehrmals zitierten Experten-Novizen Studie (siehe u.a. Abschnitt 4.3) berichtet. In einer zweiten Studie mit diesen beiden Systemen (Putz-Osterloh & Lemme, 1987, S. 299) zeigten sich in der Gruppe der studentischen Experten keinerlei Transfereffekte, während die Gruppe der Nicht-Experten von ihren Erfahrungen der ersten Systemsteuerung profitierte.

Schaub (1990) ließ 30 Versuchspersonen die Systeme „Maschine" und „Simutanien" steuern. Sowohl zwischen Verhaltens- als auch zwischen System- und Gütedaten der beiden Situationen zeigten sich praktisch keine Übereinstimmungen.

Selbst bei Szenarien, für die eine relativ hohe Retest-Stabilität (hier mindestens $r=.60$) nachgewiesen werden konnte, nämlich für die Szenarien „Jogi'91" und „Feuer", zeigten sich bei 24 studentischen Versuchspersonen keinerlei bedeutsamen Übereinstimmungen zwischen den jeweiligen Endergebnissen (Schoppek, 1996, S.156), lediglich für die Tendenz, eher viele oder eher wenige Operatoren zu nutzen, konnte eine gewisse intersituative Konsistenz aufgewiesen werden.

Im Rahmen der Diskussion des Reliabilitäts-Aspektes (siehe oben, Kapitel 8) wurde bereits darauf hingewiesen, daß Studien mit sehr ähnlichen Szenarien entweder als Studien zur Paralleltest-Reliabilität oder aber als Generalitätsstudien interpretiert werden können. Aufgrund der konstruktionsbedingten besonders hohen Ähnlichkeit zwischen den eingesetzten Szenarien werden die Untersuchungen von Hasselmann (1993) sowie Köller et al. (1993, 1995) zum Zusammenhang der Steuerungsleistungen in den Szenarien „Textilfabrik", „Heizölhandel", „Brennstoffvertrieb" und „Reifenhandel" in der vorliegenden Arbeit als Reliabilitäts-Studien betrachtet und entsprechend im Kapitel über die Reliabilität (Abschnitt 8.1.2) referiert. Eine Ausnahme von dieser Klassifikation wird für die Studie von Locher (1997) gemacht. Zwar setzte der Autor ebenfalls die sehr ähnlichen Szenarien „Heizölhandel" und „Textilfabrik" ein, mit dem experimentellen Versuchsplan seiner Untersuchung

und mit seiner zentrale Frage nach Transfereffekten stellt der Autor aber den Generalitätsaspekt explizit in den Vordergrund. Der unter der Annahme der Generalisierbarkeit über verschiedene Szenarien hinweg zu erwartende positive Transfereffekt ergab sich nur bei den 20 Personen, die zunächst die „Textilfabrik" und dann den „Heizölhandel" bearbeiteten, bei der Invertierung der Reihenfolge blieb der positive Transfereffekt hingegen aus.

In den bisher referierten Untersuchungen konnte kein nennenswerter Zusammenhang zwischen dem Problemlöseverhalten bei unterschiedlichen Systemen nachgewiesen werden. Selbst wenn die Leistungen mehrerer Szenarienbearbeitungen systematisch miteinander kovariieren würden, wäre dies allein noch kein Nachweis der Generalität der Problemlösefähigkeit. Da die Problemlöseleistung nämlich - wie weiter unten gezeigt werden wird (Abschnitt 9.1.2.2) - überzufällig mit Intelligenz und Wissen zusammenhängt, läßt sich ein möglicher Zusammenhang zwischen verschiedenen Systemsteuerungen theoretisch auch auf diese Konstrukte zurückführen, so daß sich u.U. selbst im Falle eines Nachweises der Interkorrelation verschiedener Systemsteuerungen die Annahme eines eigenständigen Konstrukts „Problemlösen" erübrigen würde. In einem zweiten Schritt muß daher über die bloßen Interkorrelationen von Problemlöseleistungen bei verschiedenen Szenarien hinaus gezeigt werden, daß die mit Hilfe verschiedener Szenarien erfassten Steuerungsleistungen auch dann noch miteinander korreliert sind, wenn die Effekte der Intelligenz und des Wissens kontrolliert werden. Dieser zweite Schritt wurde bislang nur von den Arbeitsgruppen um Süß (Berlin und Mannheim) geleistet. Wittmann, Süß und Oberauer (1996) konnten zeigen, daß die Steuerungsleistungen der „Schneiderwerkstatt" zwar systematisch mit zwei anderen (untereinander nicht signifikant korrelierenden) Szenarien („PowerPlant" und „Learn") kovariierten, daß dieser Zusammenhang nach Auspartialisierung der Indikatoren etablierter Personmerkmale wie vor allem der „Verarbeitungskapazität" und des „Wissens" allerdings seine statistische Bedeutsamkeit verlor. In der Berliner Untersuchung (Süß, 1996, S. 106 und 202) zeigte sich ein ähnlicher Befund für den Zusammenhang zwischen der Steuerungsleistung in der „Schneiderwerkstatt" und dem kleinen System „Tomaten".

9.1.1.2 Zusammenfassung der Befunde zur Generalität der Problemlösefähigkeit

Bislang konnte in keiner Untersuchung ein überzeugender Nachweis dafür erbracht werden, daß die Annahme einer aufgabenunabhängigen, generalisierbaren und eigenständigen Fähigkeit zum Problemlösen sinnvoll ist. Den unterschiedlichen Szenarien liegt offensichtlich kein einheitliches, stabiles Anforderungsprofil zugrunde, die Szenarien stellen keine transsituationalen Anforderungen an eine „Problemlösefähigkeit". Es scheint daher sinnvoll, mit Pawlik (1988, S. 152) anzunehmen, daß

mit den Szenarien nicht „*»allgemeines komplexes Problemlöseverhalten«, sondern im Gegenteil höchst spezielle Leistungsfunktionen*" geprüft werden. Sofern sich überhaupt Hinweise auf aufgabenübergreifende Anforderungsstrukturen zeigen, können diese auf bereits etablierte Konstrukte, nämlich Intelligenz und Wissen, zurückgeführt werden. Die Fähigkeiten, die über Wissen und Intelligenz hinaus zur Szenariensteuerung benötigt werden, sind offensichtlich so situations- und aufgabenspezifisch, daß daraus keine diagnostisch relevanten Schlußfolgerungen gezogen werden können, da eine derartige Anforderung in einem anderen Kontext nur mit geringer Wahrscheinlichkeit wieder auftritt.

Der Appell von Putz-Osterloh (1983, S. 115), ungeachtet der fehlenden systematischen Beziehungen zwischen Leistungen bei unterschiedlichen Systemen an der Annahme einer übergeordneten Problemlösefähigkeit festzuhalten, kann nicht überzeugen. Die Autorin vermutet, daß sich die vermisste Gemeinsamkeit einstellt, sobald man die Problemlöse*prozesse* betrachtet. Diese Annahme findet in den Befunden von Schaub (1990) und Strohschneider (1990, 1991a), die jeweils vergeblich auch nach systemübergreifenden *Verhaltensweisen* Ausschau gehalten haben, keine Bestätigung. Außerdem liegen bislang überhaupt keine konsensfähigen, psychometrisch befriedigenden Operationalisierungen von Prozeßmaßen des Problemlösens vor (siehe oben, Abschnitt 6.3.1).

Bis auf weiteres ist davon auszugehen, daß mit computergestützten Problemlöseszenarien zu einem großen Teil *systemspezifische*, nicht generalisierbare Leistungen erfaßt werden. Der (relativ kleine) Anteil der über verschiedene Problemlöseaufgaben hinweg systematischen gemeinsamen Varianz läßt sich als Intelligenz- und Wissensleistung kennzeichnen. Für die Diagnostik mit computergestützten Problemlöseszenarien hat dieser Befund gravierende Auswirkungen, die weiter unten im Rahmen der Generalzusammenfassung zur Validität diskutiert werden.

Einschränkend ist allerdings einzuwenden, daß einige Untersuchungen nicht die methodischen Voraussetzungen erfüllten, einen möglicherweise existierenden Zusammenhang aufzudecken. Diesbezüglich gelten inhaltlich die weiter unten (Abschnitt 9.1.2.3) dargestellten Ausführungen zu den methodischen Defiziten der Untersuchungen zum Zusammenhang von Intelligenz und Problemlösen.

9.1.2 Intelligenz und Problemlösen

9.1.2.1 Theoretische Überlegungen zum Zusammenhang von Intelligenz und Problemlösen

Sowohl Laien als auch Experten sehen die Fähigkeit zum Problemlösen als ein Attribut intelligenter Personen an (z.B. Sternberg, 1989). Auch wenn keine einheitliche Beschreibung des Intelligenz-Konstrukts existiert, gilt doch, daß gerade die Problemlösefähigkeit als gemeinsames Element zahlreicher Definitionsversuche der Intelligenz genannt wird. *„Whatever intelligence may be, reasoning and problem-solving have traditionally been viewed as important subsets of it. Almost without regard to how intelligence has been defined, reasoning and problem solving have been part of the definition"* (Sternberg, 1982, S. 225). Bei einem 1921 und 1986 veranstalteten Symposium wurden die versammelten Experten gebeten, Intelligenz zu definieren. Jedesmal waren es vor allen anderen genannten Attributen stets Komponenten höherer Ordnung - wie z.B. das Problemlösen- , die von mindestens der Hälfte der Experten als Attribute zur Kennzeichnung von Intelligenz angeführt wurden (Sternberg & Frensch, 1990, S. 60). Nach Guthke, Böttcher und Sprung (1991, S. 202) *„verstehen die meisten zeitgenössischen Forscher unterschiedlicher Provenienz Intelligenz als Problemlösefähigkeit."* Dörners (1986) Beschreibung der „operativen Intelligenz" als zweckgerichtete Fähigkeit zur Koordination von Einzelfähigkeiten, als Regulationsebene oder höhere Organisationsform des Denkens, deckt sich mit etlichen Intelligenzdefinitionen, etwa mit Sternbergs (1990) Beschreibung der Intelligenz als *„mental self-government"*. In Sternbergs (1984) triarchischer Intelligenztheorie sind die von Dörner angesprochenen Organisationsleistungen der „operativen Intelligenz" als Metakomponenten im Rahmen der Komponenten-Subtheorie zu klassifizieren. Mit „Metakomponente" bezeichnet Sternberg kognitive Prozesse höherer Ordnung, die es einem Individuum erlauben zu planen, die eigenen Aktionen zu überwachen und die Ergebnisse der Aktionen zu evaluieren. In diesen Ausführungen zu einem Aspekt der *Intelligenz* findet sich ein Großteil der Verhaltensweisen wieder, die häufig der Problemlösefähigkeit zugeschrieben werden (siehe z.B. Putz-Osterloh, 1981).

Bei derart hohen Übereinstimmungen auf der Ebene der Theoriesprache ist zu erwarten, daß die jeweiligen Indikatoren der Konstrukte (Ebene der Beobachtungssprache) miteinander korrelieren, und daß z.B. die Bearbeitung von Problemlöseszenarien Anforderungen an die Intelligenz der Versuchspersonen stellt (Dörner, 1986, S. 297). Dies sahen auch Versuchspersonen so, die im Anschluß an die Bearbeitung eines computergestützten Problemlöseszenarios danach gefragt wurden, welche Fähigkeiten bei der Systemsteuerung gefordert sind (siehe z.B. Putz-Osterloh &

Haupts, 1990, S. 142; Dörner & Pfeifer, 1992). Nach Ansicht von Experten stellen Problemlöseaufgaben insbesondere Anforderungen an das schlußfolgernde Denken und an das Arbeitsgedächtnis. Süß et al. (1991, S. 337) gehen davon aus, daß mit Hilfe induktiver Denkprozesse aus der Fülle der Variablenwerte und deren Veränderungen Regelmäßigkeiten extrahiert werden und daß das induktive Denken die Korrektur und die Erweiterung des Vorwissens erlaubt (siehe auch K.J. Klauer, 1996). Demgegenüber sei das deduktive Denken notwendig, um aus dem Wissen über Wenn-Dann Zusammenhänge konkrete Handlungsabsichten abzuleiten. Aufgaben zum induktiven und deduktiven Denken sind im Berliner Intelligenzstrukturmodell (Jäger, 1982) der Operationsklasse „Verarbeitungskapazität" zugeordnet, entsprechend erwarten die Autoren, daß diese Intelligenzdimension der wichtigste Einzelprädiktor von Problemlöseleistungen ist. Diese besondere Bedeutung der „Verarbeitungskapazität" läßt sich auch aus den theoretischen Annahmen zur Belastung des Arbeitsgedächtnisses beim Problemlösen ableiten (z.B. Anderson, 1983; Newell & Simon, 1972; Klauer, 1993), da die „Verarbeitungskapazität" von allen Intelligenzkomponenten die engste Verbindung mit dem Arbeitsgedächtnis aufweist (siehe z.B. Süß, Oberauer, Wittmann, Wilhelm & Schulze 1996, Oberauer, 1993a).

Im Anschluß an eine Literaturübersicht führen Gardner und Sternberg (1994) verschiedene Gründe dafür an, daß die Leistung in neuen und unbekannten Situationen Zusammenhänge zur Intelligenz aufweist. Die Ausführungen der Autoren zu dieser allgemeineren Fragestellung sind hier bedeutsam, da Problemlöseaufgaben u.a. durch das Moment des Neuen definiert sind (siehe Kapitel 2). Neben der bereits angesprochenen Bedeutung des Arbeitsgedächtnisses und des schlußfolgernden Denkens sowie der weiter unten thematisierten Bedeutung des Aufbaus einer adäquaten Wissensbasis heben Gardner und Sternberg (ebd.) allerdings auch die Bedeutung motivationaler Faktoren hervor. Neue Situationen erfordern Frustrationstoleranz, eine hohe Aufmerksamkeitsspanne und bieten Lerngelegenheiten. Die beiden zuletzt genannten Aspekte sollten eine Verbindung zur Intelligenz begründen.

Aus diesen theoretischen Ausführungen ergibt sich die Erwartung einer systematischen empirschen Beziehung zwischen Intelligenztest- und Problemlöseleistungen. Befunde, die diese Erwartung enttäuschten, wurden unterschiedlich interpretiert. Entweder wurde die meßtheoretische Güte der Problemlöseleistungen (z.B. Tent, 1984) oder aber die Validität der Intelligenztests in Frage gestellt. An dieser Stelle sei bemerkt, daß es vor dem Hintergrund der oben (siehe Abschnitt 9.1.1) referierten mangelnden Generalisierbarkeit der Problemlöseleistungen verwunderlich ist, daß die Dissoziation von Intelligenz- und Problemlöseleistungen als Argument *gegen die Validität der Intelligenztests* Anklang finden konnte. Wie auch immer: beide Interpretationsweisen betrafen die *Meßebene*, die auf theoretischer Ebene formulierte Erwartung eines Zusammenhang zwischen Intelligenz und Problemlösen

wurde nur selten aufgegeben. Umstritten war für die meisten Forscher nicht die Bedeutung von intellektuellen Fähigkeiten für die Problemlöseleistung, sondern umstritten war und ist vielmehr die Bedeutung der entsprechenden *Meßverfahren*, der Intelligenz*tests*. Entsprechend führt Putz-Osterloh (1985, S. 104) aus, daß sich das Verhalten erfolgreicher Problemlöser unschwer als „intelligent" klassifizieren läßt. Nach Ihrer Ansicht *„verfügen wir bisher über kein Testverfahren, das diese Leistungen zu erfassen vermag"*.

Sofern dem Konstrukt „Problemlösen" lediglich Verhaltensweisen zugeordnet werden können, die bereits typisch für andere Konstrukte sind, erübrigt sich die Annahme einer gesonderten Fähigkeit. Insgesamt betrachtet lassen sich Problemlöseleistungen theoretisch unter zahlreiche bestehende Intelligenzbegriffe subsummieren. Neu an dem Ansatz der „komplexen Problemlöseforschung" sind weniger die kognitionstheoretischen Überlegungen als vielmehr die *Instrumente* zur Erhebung und die *Konzepte zur Auswertung* von Daten, die in der Tat deutliche Unterschiede zu den bisherigen Instrumenten der Intelligenzforschung aufweisen. Die konzeptionelle Vereinbarkeit der Fähigkeitsbegriffe „Intelligenz" und „Problemlösen" gelingt auf der theoretischen Ebene nicht zuletzt deshalb, weil die Definitionen der Intelligenz häufig wenig mit den tatsächlich vorgenommenen Leistungsmessungen korrespondieren (Carroll, 1993, S. 35). Zahlreiche Anforderungen, deren Bewältigung dem Konstrukt „Intelligenz" zugeschrieben wird, finden sich bekanntlich nicht in den herkömmlichen - auf die „akademische" Intelligenz begrenzten - Tests wieder (siehe Neisser, 1976 sowie ergänzend Sternberg und Wagner, 1993, S. 2; Wagner und Sternberg, 1985, S. 437). Diese berechtigte Kritik an Intelligenztests war seit langem gängige Münze, die mit der Problemlöseforschung in Deutschland noch einmal in schwunghaften Umlauf gebracht wurde. Auf der Ebene der Operationalisierungen, auf der Ebene der *Aufgaben* also, lassen sich deutliche Unterschiede zwischen Intelligenz und Problemlösen ausmachen, die zum Teil (siehe Abschnitt 3.4) bereits beschrieben wurden. Intelligenztestaufgaben sind nicht komplex (im engeren Sinne), nicht dynamisch, nicht vernetzt und nicht intransparent. Infolgedessen erfordern Intelligenztestaufgaben - ungeachtet der theoretischen Intelligenz-Auffassung - de facto in einem ungleich schwächeren Ausmaß als Problemlöseaufgaben die selbstständige Organisation vieler Lösungsschritte zu einer Lösungsprozedur. Bei der Bearbeitung von Intelligenztests fehlt nach Ansicht Dörners (Dörner 1976, S. 133) allgemein all das, was bei der Durchführung längerfristiger Denkakte von entscheidender Bedeutung ist. Außerdem kommt dem Erwerb und dem Umgang mit Wissen bei Intelligenztests eine andere Bedeutung zu als bei computergestützten Problemlöseszenarien: bei Intelligenztests wird das Wissen nach Ansicht von Dörner et al. (1983b S. 321) eher „abgefragt", als daß Ableitungen aus einem selbst strukturierten Erfahrungssatz zu ziehen sind (zum Aspekt Problemlösen und Wissen siehe

unten, Abschnitt 9.1.3), und schließlich spielt bei Intelligenztests die Eigeninitiative nur eine untergeordnete Rolle. Die Kritik läßt sich dahingehend zusammenfassen, daß mit den durch Intelligenztests erfaßten Fähigkeiten das Konstrukt Intelligenz unterrepräsentiert ist. Trotz all dieser kontrastierenden Charakterisierungen der beiden diagnostischen Ansätze finden sich aber auch Gemeinsamkeiten zwischen beiden Aufgabentypen, beispielsweise werden hier wie dort Anforderungen an grundsätzliche intellektuelle Fähigkeiten, wie z.B. das schlußfolgernde Denken, gestellt (siehe z.B. Putz-Osterloh, 1981, S. 83).

Aufgaben zum komplexen Problemlösen können theoretisch als eine Erweiterung der herkömmlichen Intelligenzdiagnostik betrachtet werden (siehe etwa Putz-Osterloh, 1985, S. 215). Intelligenz und Problemlösen müßten dieser Auffassung zufolge empirisch in etwa der Höhe miteinander korreliert sein, wie Aufgaben zu verschiedenen Dimensionen der Intelligenz (z.B. „Verarbeitungskapazität" und „Merkfähigkeit"). Jäger et al. (1997) berichten, daß im Test zum Berliner Intelligenzstrukturmodell („BIS-4"-Test) die Skalen der Operationsklasse im Streubereich von $r=.25$ bis $r=.47$ miteinander korrelieren. In der Höhe ähnliche Koeffizienten ergeben sich auch dann, wenn man verschiedene Tests der Intelligenz zugrundelegt. So korrelieren die mit Hilfe des „BIS-Tests" gemessenen Operationsklassen „Merkfähigkeit", „Bearbeitungsgeschwindigkeit" und „Einfallsreichtum" z.B. bei Bucik und Neubauer (1996, S. 993) zwischen $r=.37$ bis $r=.49$ mit den Raven-Matrizen als Indikator für „Verarbeitungskapazität".

Theoretisch sind Korrelationen in der genannten Höhe zwischen Indikatoren der Intelligenz- und der Problemlöseleistungen zu erwarten. Dabei sollte das Ausmaß des Zusammenhangs in Abhängigkeit von der in jedem einzelnen Untersuchungsfall realisierten Ähnlichkeit der Anforderungen variieren. Berücksichtigt man die bislang entwickelten Thesen und empirischen Ergebnisse zum Zusammenhang von Intelligenz und Problemlösen so sollten die Gemeinsamkeiten zwischen den mit unterschiedlichen Aufgabenklassen gemessenen Leistungen höher ausfallen, wenn die Problemlöseleistungen (1) reliabel gemessen werden (z.B. Funke, 1993) und wenn die Problemlöseaufgaben (2) mit einem konkreten Ziel (z.B. Strohschneider, 1991a) (3) unter Transparenzbedingungen (z.B. Putz-Osterloh & Lüer, 1981) (4) mit einer mittleren Schwierigkeit (z.B. Raaheim, 1974; Hussy, 1985) sowie (5) mit einem eher abstrakten semantischen Kontext vorgegeben werden (z.B. Hesse, 1982; Spies & Hesse, 1987). Auf Seiten der Intelligenztests sind vor allem dann Zusammenhänge zu erwarten, wenn (6) nicht vereinzelte Intelligenztestaufgaben, sondern psychometrisch hochwertige und umfassende Verfahren eingesetzt werden sowie wenn (7) Testverfahren eingesetzt werden, die eine Differenzierung auf Seiten der Intelligenz erlauben (z.B. Süß et al., 1991).

Studien zum Zusammenhang von Intelligenz- und Problemlöseleistungen wurden bereits an anderer Stelle (siehe z.B. Funke, 1986; Kluwe et al., 1991a ; Kluwe et al., 1991c) in Übersichtsform dargestellt und kritisch gewürdigt, so daß sich die folgende Darstellung auf eine thematisch fokussierte Auswahl beschränken kann.

Divergente Korrelationsmuster - zumindest unter Intransparenzbedingungen - zwischen Intelligenz- und Problemlöseleistungen wurden vor allem aus der Anfangsphase der Forschung zum komplexen Problemlösen berichtet. Neben der folgenreichen „Lohhausen"-Studie (Dörner et al, 1983) sind hier beispielhaft die Arbeiten von Kühle und Badke (1986), Putz-Osterloh (1981), Putz-Osterloh und Lüer (1981) sowie Stäudel (1987) zu nennen.

Die im Laufe der Jahre angewachsene Zahl von Untersuchungen erlaubt es, den moderierenden Einfluß der oben aufgeführten Bedingungen auf den Zusammenhang zwischen Intelligenz- und Problemlöseleistungen zu betrachten.

Daß die **Reliabilität** der Maße den möglichen Zusammenhang der beiden Leistungen von vornherein einschränkt, zeigte bereits die Studie von Funke (1983). Ein Haupteffekt für die mit Hilfe der Raven-Matrizen gemessene Intelligenz auf die Steuerungsleistung in der „Schneiderwerkstatt" zeigte sich nur dann, wenn ein reliables - und nicht das in den Studien von Putz-Osterloh (1981) sowie Putz-Osterloh und Lüer (1981) verwendete unreliable - Problemlösegütemaß verwendet wurde. Eine systematische Untersuchung über den Einfluß der Reliabilität hat Süß (1996) vorgestellt, wobei er allerdings den Zusammenhang zwischen einer spezifischen Intelligenzdimension („Verarbeitungskapazität") und der Leistung bei einer *klassischen* Problemstellung, nämlich dem Tangram-Puzzle, studierte. Ausgangspunkt der Studie von Süß war die Aussage von Dörner und Kreuzig (1983), derzufolge Intelligenztest- und Tangramleistungen - entgegen früheren Untersuchungen, z.B. Klix und Lander (1967) - nichts miteinander zu tun hätten. Süß (ebd.) zeigte in seinen beiden Studien mit 30 und 25 Studenten, daß die Tangram-Leistungen sehr wohl mit der Intelligenzkomponente „Verarbeitungskapazität" zusammenhängen. Die Höhe der Korrelationen stand aber in deutlicher Abhängigkeit von der Reliabilität der Messung der Problemlöseleistung. Erst wenn mehere Tangram-Leistungen aggregiert wurden, zeigten sich starke Effekte, die unter den Bedingungen der Studie von Dörner und Reither überhaupt nicht beobachtet werden *konnten*.

Hinweise auf die Effekte der **Zielspezifikation** leiten sich u.a. aus einer Studie von Strohschneider (1991a) ab. Der Autor ließ das System „Moro" einmal von 25 Studenten ohne Zielspezifikation und einmal von 20 Schülern mit klaren Zielvorgaben steuern. Lediglich unter der Bedingung der eindeutigen Zielvorgabe korrelierten einige Indikatoren der Systemsteuerung mit den Leistungen in den Intelli-

genzskalen. Nach einer Überlegung von Süß (1996, S. 58) stellen diese Effekte der Zielspezifikation letztendlich aber lediglich Effekte der Reliabilitätssteigerung dar.

Zum Einfluß der **Transparenz-Bedingung** auf den Zusammenhang von Intelligenztest- und Problemlöseleistungen wurden widersprüchliche Ergebnisse berichtet. Während bei Putz-Osterloh und Lüer (1981) sowie bei Hörmann und Thomas (1989) nur unter Transparenzbedingungen ein Zusammenhang zwischen Problemlöse- und Intelligenztestleistungen bestand, konnte dieser Zusammenhang bei Funke (1983), Hussy (1989) sowie bei Süß et al. (1991) auch unter Intransparenzbedingungen aufgezeigt werden. Genau entgegengesetzt zur Transparenz-Hypothese ergab sich in der Berliner Wiederholungsuntersuchung sogar der Befund, daß sich nur unter der Intransparenzbedingung ein signifikanter Zusammenhang zwischen Intelligenz- und Steuerungsleistungen aufzeigen ließ (siehe Süß, 1996, S. 163f.).

Die auf Raaheim (1974) zurückgehende These, derzufolge der Zusammenhang zwischen Intelligenz und Problemlöseleistungen durch die **Aufgabenschwierigkeit und -neuheit** moderiert wird, wurde u. a. von Hussy (1985) geprüft. Hussy variierte beim Szenario „Zielannäherung" einige Komponenten, die seiner Taxonomie nach der Komplexität entsprechen, nämlich die Variablenzahl, die Transparenz und die Variablenvernetzung. Die These von Raaheim bestätigend, konnte Hussy für die 120 studentischen Teilnehmer nachweisen, daß der Zusammenhang zwischen Problemlöseleistung und den mit Hilfe des CFT3 bestimmten Intelligenzleistungen mit zunehmender Problemschwierigkeit sinkt.

In der Berliner Erstuntersuchung wurde die auf Raaheim zurückgehende These durch die zusätzliche Applikation von zwei reduzierten Versionen der „Schneiderwerkstatt" geprüft. Der formulierten These zuwiderlaufend korrelierten die Steuerungsleistungen bei den reduzierten Versionen entweder gar nicht oder sogar niedriger mit den Intelligenztestleistungen als die bei den komplexeren, vollständigen Versionen der „Schneiderwerkstatt" erzielten Leistungen (Süß, 1996, S. 161f.).

Hesse (1982) sowie Spies und Hesse (1987) nehmen an, daß der Zusammenhang von Intelligenz und Problemlösen dann gering ausfällt, wenn die verwendeten Problemlöseszenarios eine semantische, Vorwissen aktivierende Einkleidung besitzen. In diesen Fällen - so die Überlegung - wird die auf das Wissen zurückzuführende Varianz, die vom Intelligenz-Konstrukt nicht erfaßt wird, den Problemlöseerfolg mitdeterminieren. In einem Experiment mit 120 Probanden variierten Spies und Hesse (1987) die semantische Einkleidung des „Dori"-Szenarios und korrelierten die unter den verschiedenen experimentellen Bedingungen erzielten Steuerungsleistungen mit den Leistungen im „Raven-Test". Während diese Korrelationen beim Vorhandensein eines semantischen Kontextes um Null ($r=-.06$) lagen, zeigte sich unter **Wegfall der semantischen Einkleidung** ein deutlicher Zusammenhang ($r=.42$). Ähnliche Befunde berichten Beckmann (1994) sowie Beckmann und Guthke (1995)

für ein Experiment, bei dem ebenfalls ein formal identisches Problem mit unterschiedlichen semantischen Einkleidungen vorgegeben wurde. Lediglich unter der abstrakten semantischen Bedingung (Einkleidung als „Maschine") konnten die Steuerungsleistungen der 40 Schüler durch Lerntestaufgaben (r=.57) sowie durch zwei Aufgaben aus Intelligenz-Statustests (r=.36) vorhergesagt werden.

Hervorzuheben sind solche Studien, die auf Seiten der Intelligenz **psychometrisch hochwertige und struktur-differenzierende Verfahren** eingesetzt haben. Ein solches Verfahren liegt mit dem Test zum Berliner Intelligenzstrukturmodell (BIS, Jäger et al., 1997) vor. In Untersuchungen mit verschiedenen Versionen dieses Tests konnten Zusammenhänge zwischen den so gemessenen Intelligenzleistungen und den Problemlöseleistungen aufgezeigt werden, allerdings blieben bei Reichert und Dörner (1988), Funke (1985b) sowie Putz-Osterloh und Lemme (1987) auch beim Einsatz dieses Meßverfahrens die erwarteten Korrelationen aus.

In der Berliner Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen (Süß et al., 1991, 1993a, 1993b), die ebenfalls mit dem Test zum Berliner Intelligenzstrukturmodell durchgeführt wurde, konnte die weiter oben erläuterte besondere Bedeutung der Intelligenzkomponente „Verarbeitungskapazität" für die Problemlöseleistung in der „Schneiderwerkstatt" unter Beweis gestellt werden. Für ein über die Erst- und Wiederholungsuntersuchung aggregiertes Problemlösegütemaß (zu den notwendigen Modifikationen des Problemlösegütemaßes der Erstuntersuchung siehe oben, Abschnitt 7.3.2) ergab sich ein Zusammenhang zwischen diesem Intelligenzaspekt und der Steuerungsleistung von r=.47 (N=137). Für eine Teilstichprobe von 40 Personen, bei der in der Wiederholungsuntersuchung auf eine experimentelle Induktion von Wissen verzichtet wurde, betrug der entsprechende Zusammenhang r=.65.

Der Ansatz der Berliner Untersuchungen zum Zusammenhang von Intelligenz, Wissen und Problemlösen wurde in Mannheim fortgeführt (Wittmann et al., 1996). 92 Studenten bearbeiteten eine umfangreiche Batterie von Testaufgaben (u.a. BIS-4 Test, Working-memory Aufgaben, Wissenstests, einen Test zur Erfassung der Computererfahrung sowie einzelne Skalen aus Persönlichkeitsfragebogen) und zwei Problemlöseszenarien („Schneiderwerkstatt" und „PowerPlant"). Eine Teilgruppe von 57 Personen steuerte zusätzlich noch das betriebswirtschaftlich eingekleidete System „Learn". Ein über alle drei Szenarien gebildetes Globalmaß der Problemlöseleistung war mit allen Intelligenzfaktoren substantiell korreliert, mit einer Korrelation in Höhe von r=.62 erwies sich die „Verarbeitungskapazität" als bester Prädiktor. Mit multiplen Regressionen konnte die Mannheimer Arbeitsgruppe zeigen, daß - mit Ausnahme eines bedeutsamen Beitrags des „Einfallsreichtums" zur Vorhersage der Steuerungsleistung bei „PowerPlant" - die „Verarbeitungskapazität" der einzige relevante Prädiktor unter den operativen Intelligenzkomponenten war.

Eine so eindeutige prädiktive Dominanz der „Verarbeitungskapazität" wie in der Berliner und Mannheimer Untersuchung zeigte sich allerdings nicht in allen Studien, die eine Form des BIS-Tests verwendeten. So erwies sich zwar auch bei Hussy (1989) die „Verarbeitungskapazität" als bester Einzelprädiktor der Leistung im System „Zielannäherung", aber auch alle anderen Skalen leisteten einen signifikanten Beitrag zur Vorhersage der Steuerungsleistung. Insgesamt wies der Summenwert für die allgemeine Intelligenz mit r=.38 nominell die höchste Korrelation auf. Auch bei Strohschneider (1991a) waren *alle* Intelligenzskalenleistungen mit Indikatoren der Sollwertabweichung des - mit Zielvorgabe gesteuerten - „Moro"-Systems und - mit Ausnahme des Einfallsreichtums - des „Vektor"-Systems korreliert - die „Verarbeitungskapazität" war aber auch hier zumindest nominell der beste Prädiktor. Putz-Osterloh (1987) konnte bei 30 Studenten zwar vereinzelt bedeutsame Korrelationen zwischen der „Verarbeitungskapazität" und Indikatoren der Steuerungsleistung in der „Schneiderwerkstatt" aufweisen, entsprechende Verbindungen mit der Steuerung des ebenfalls eingesetzten „Moro"-Systems blieben aber aus. In einer anderen Untersuchung (1985) fand die Autorin für eine Experimentalgruppe mit einem „Selbstreflexionstraining" (nicht aber für die Kontrollgruppe) Zusammenhänge zwischen den „BIS"-Skalen (ohne die Skala zur „Merkfähigkeit") und verschiedenen Indizes der Steuerungsleistung im System „Moro". Es ergaben sich insbesondere für die „Bearbeitungsgeschwindigkeit" überzufällig enge Beziehungen zur Steuerungsleistung. Die „Bearbeitungsgeschwindigkeit" erwies sich auch in der Studie von Hörmann und Thomas (1989) - noch vor den ebenfalls signifikanten Prädiktoren „Verarbeitungskapazität" und der „Allgemeinen Intelligenz" - als nominell bedeutsamster Vorhersagefaktor der unter der experimentellen Bedingung der Transparenz erzielten Steuerungsleistung in der „Schneiderwerkstatt". Unter Intransparenzbedingungen ließen sich nur dann signifikant positive Korrelationen mit Intelligenzmaßen - und zwar innerhalb der Operationsklassen insbesondere mit der „Verarbeitungskapazität" und der „Merkfähigkeit" - ausmachen, wenn man diese Gruppe in zwei weitere Subgruppen mit hohem und niedrigem Systemwissen unterteilte. Diese Differenzierung überstrapaziert allerdings die mit 41 Personen ohnehin für den Mehrgruppenplan der Untersuchung (zu) knapp kalkulierte Stichprobengröße.

Die zuletzt genannten Studien weisen darauf hin, daß die Bedeutung der einzelnen Intelligenzdimensionen für das Problemlösen sowie ihr Zusammenwirken noch nicht abschließend ausgelotet ist. Obwohl die Funktion der „Verarbeitungskapazität" beim Problemlösen bislang theoretisch und empirisch am besten herausgearbeitet wurde, scheint es verfrüht, die Bedeutung anderer Intelligenzdimensionen für das Problemlösen, wie z.B. die Bedeutung der „Bearbeitungsgeschwindigkeit" (insbesondere bei Systemsteuerungen unter Zeitdruck) oder der „Merkfähigkeit" und die Interaktion dieser Komponenten sowie Faktoren höherer Ordnung wie der „All-

gemeinen Intelligenz" zu vernachlässigen. Es ist daher bedauerlich, daß in einigen Studien die Messung der Intelligenz auf die Skalen zur „Verarbeitungskapazität" aus dem BIS-Test verkürzt wurde, so etwa in den Untersuchungen von Hussy mit dem eher klassischen Problem „Superhirn" (1991a) oder mit der „Schneiderwerkstatt" (1991b). In der zuletzt genannten Untersuchung konnte der Autor für den ersten von fünf Durchgängen eine Korrelation zwischen der Steuerungsleistung und der „Verarbeitungskapazität" von r=.81 beobachten. Die Studie wurde allerdings aufgrund verschiedener Mängel wie z.B. der geringen Stichprobengröße und der fehlenden Wissensdiagnostik kritisiert (siehe Beckmann & Funke, 1991).

Eine auf 12 Aufgaben reduzierte Skala zur „Verarbeitungskapazität" des Tests zum BIS wurde in zwei Studien mit dem „Jogi"-Szenario eingesetzt (Schoppek, 1996). An beiden Studien nahmen jeweils 48 studentische Versuchspersonen teil. Mit Ausnahme einer Subgruppe mit einem spezifischen Training korrelierten in den übrigen Versuchsbedingungen die Intelligenzleistungen in der Dimension „Verarbeitungskapazität" im Streubereich von r=.32 und r=.74 mit verschiedenen Maßen des Steuerungserfolgs, insbesondere aber mit dem Maß „Endkapital". Durch die zusätzliche Berücksichtigung der ebenfalls erhobenen „Kontrollmeinung" bei der Vorhersage der Problemlösegütemaße konnte gegenüber der „Verarbeitungskapazität" keine zusätzliche Varianz aufgeklärt werden, die „Verarbeitungskapazität" war als alleiniger Prädiktor anzusehen.

Auch Intelligenzaufgaben aus anderen Tests lassen sich in das Berliner Intelligenzstrukturmodell klassifizieren (zur Invarianz des Modells über verschiedene Aufgabensätze hinweg siehe z.B. Jäger und Tesch-Römer, 1988). Schoppek (1991) korrelierte die Intelligenztestleistungen in verschiedenen Subtests des IST-70 mit dem über fünf Durchgänge gemittelten Steuerungserfolg beim Szenario „Feuer". Die höchsten Zusammenhänge ergaben sich für zwei Aufgaben, die in der Terminologie des Berliner Intelligenzstrukturmodells der „Verarbeitungskapazität" zuzuordnen sind („Würfelabwicklungen", r=.66 und Zahlenreihen, r=.54) sowie für eine Aufgabe zur „Merkfähigkeit" (r=.47).

Demgegenüber leisteten die IST-70 Aufgaben in einer Untersuchung von Putz-Osterloh und Köster (1988) keinen bedeutsamen Beitrag (r=-.16) zur Vorhersage der Problemlösegüte von 100 Personen in der „Schneiderwerkstatt" (als Problemlösegütekriterium wurden die Probanden anhand dreier Zielvariablen in drei „Gütekategorien" eingestuft). Die Steuerung der „Schneiderwerkstatt" erwies sich auch bei 32 Offiziersbewerbern als unabhängig von der (Raven-Test-)Intelligenzleistung (Putz-Osterloh & Schroiff, 1987)

K.J. Klauer sieht in seinen 1996 veröffentlichten Untersuchungsergebnissen einen experimentellen Nachweis dafür, daß induktives Denken (und somit klassische Intelligenztestleistungen) das Problemlösen begünstigen. Die Experimentalgruppen er-

hielten jeweils ein Training zum induktiven Denken. Nach dem Denktraining - bzw. im Fall der Kontrollgruppe nach der entsprechenden Kontrollintervention - bearbeiteten die Versuchspersonen das Problemlöseszenario „Hunger in der Sahelzone". Bei 84 Mädchen zeigte sich für eines von zwei Problemlösegütemaßen eine Überlegenheit der Gruppe, die ein Denktraining erhalten hatte. Allerdings war das Szenario für die Altersgruppe der Kinder möglicherweise etwas zu schwer (zur Überforderung siehe Kapitel 7). In einer konzeptionellen Replikation arbeitete Klauer (ebd.) mit 60 Kindern, die im Vergleich zur Untersuchungsgruppe des ersten Experiments (1) zwei Jahre älter waren und (2) als Jungen nach Ansicht des Autors Computer vorbehaltloser akzeptierten als Mädchen (siehe hierzu auch Abschnitt 10.1.3). Die Gruppe ohne das Denktraining erhielt diesmal kein Kontrolltraining, dafür aber während der Steuerung Informationen über das zu steuernde System. Kinder, die das - nachweislich wirksame - Denktraining erhalten hatten, erzielten im Durchschnitt um drei Viertel bis zu einer Standardabweichung bessere Leistungen bei der Systemsteuerung als die Gruppe ohne Denktraining. Demgegenüber wirkten sich die zusätzlichen Systeminformationen allein nicht auf die Steuerungsleistungen aus. Die Studie wurde allerdings aus methodischer Sicht (insbesondere aufgrund des fehlenden Trainings der Kontrollgruppe) kritisiert (siehe z.B. die Kritik von Hager und Hasselhorn (1996) sowie die Replik von K.J. Klauer (1996b)).

Für die in der Eignungsdiagnostik bereits eingesetzten Szenarien wurden vergleichsweise seltener Zusammenhangsuntersuchungen mit Intelligenztests durchgeführt. U. Funke (1992a) berichtet aus einer Studie mit 17 Personen eine Korrelation in Höhe von r=.40 zwischen einem Gesamtindex (Steuerung und Verhalten) der Problemlöseleistung bei „DISKo" und der Skala „Verarbeitunskapazität" aus dem BIS-Test. Für die „Textilfabrik" stellt Hasselmann (1993) die Ergebnisse einer Studie mit 21 Führungsnachwuchskräften vor. In dieser Untersuchung korrelierten die Ergebnisse im „IST-70" Untertest „Gemeinsamkeiten-Erkennen" signifikant (in Höhe von r=.45 bis r=.55) mit drei verschiedenen Problemlösegütemaßen der Steuerungsleistung. Borderline-signifikante Zusammenhänge mit der Steuerungsleistung zeichneten sich auch für den Subtest „Zahlenreihen" ab. Demgegenüber korrelierten die „IST-70" Tests „Satzergänzung" und „Wortauswahl" ebensowenig mit der Problemlösegüteleistung wie die Ergebnisse des Konzentrationstests „d2". Ein ähnliches Befundmuster berichtet Hasselmann (ebd.) auch für eine entsprechende Untersuchung mit 41 Studenten, wobei in dieser Studie der Subtest „Zahlenreihen" die größte Nähe zum Kapitalendwert in der „Textilfabrik" aufwies. Minderungskorrigierte Werte erzielten sogar eine Höhe von bis zu r=.93 für die Skala „Zahlenreihen" (ebd., S. 200). In der Untersuchung mit den Studenten wurden auch die Leistungen bei der Bearbeitung der Raven-Matrizen kontrolliert, auch diese Intelligenztestleistungen (r=.31) korrelierten mit der Steuerungsleistung. In der Studie

von Locher (1997) variierte der Zusammenhang zwischen der Steuerungsleistung bei der Bearbeitung des Szenarios „Heizölhandel" und der Leistung in der Skala „Verarbeitungskapazität" des BIS-4 Tests in Abhängigkeit von der Häufigkeit der Szenarienbearbeitung. Während sich für 40 Probanden kein Zusammenhang zwischen der Steuerungsleistung bei der Erstbearbeitung des „Heizölhandels" und der so gemessenen Intelligenz fand, zeigten sich die Retest-Steuerungsleistungen derjenigen Hälfte der Gruppe, die das Szenario ein zweitesmal bearbeitete, mit der „Verarbeitungskapazität" assoziert ($r = .46$).

Andere Autoren beschränken sich hinsichtlich der Zusammenhänge zwischen den Leistungen in eignungsdiagnostischen Szenarien und Intelligenztestleistungen auf unzureichende Andeutungen. So korrelierten laut Hartung und Schneider (1995, S. 233) in einer Studie mit 54 Teilnehmern die Ergebnisse im IST-70 zu $r = .52$ mit der Strategieauswertung im Szenario „Utopia", wobei aber zu beiden Maßen nähere Angaben fehlen. Laut Neubauer (1995, S. 167) konnten nicht näher genannte psychologische Testverfahren die Ergebnisse der Simulation „Manage" zu 75% aufklären. Dies entspricht einer Korrelation in einer Höhe von über $r = .86$.

9.1.2.3 Methodische Defizite der Untersuchungen zum Zusammenhang von Intelligenz und Problemlösen

Bei zahlreichen Untersuchungen zum Zusammenhang von Intelligenz und Problemlösen wurden einfache methodische Voraussetzungen verletzt. Neben den notorischen Schwächen der fehlenden Zielspezifkation der Problemsteuerung, der instruktionsunabhängigen post-hoc Bestimmung von Problemlösegütemaßen und der mangelnden Differenzierung auf Seiten der Intelligenzmessung kritisiert Süß (1996, S. 58f.) die Vernachläßigung der Reliabilität der Gütemaße sowie die Vernachlässigung der Symmetrie zwischen den Konstrukten im Sinne Wittmanns (1988). Süß bemängelt außerdem, daß weitere Prädiktoren der Problemlöseleistung (wie z.B. Wissen) häufig keine Berücksichtigung fanden. Unter diesen Umständen sind nach Ansicht des Autors bestenfalls mittlere Effektstärken zu erwarten. Süß (ebd.) hat nun ein kaum zu entwaffnendes Argument auf seiner Seite, wenn er daran erinnert, daß eine Stichprobengröße von 64 Personen notwendig ist, um einen mittleren Effekt bei einseitiger Testung mit einer statistischen Power von .80 aufzudecken. Tatsächlich wählten die meisten Autoren eine Stichprobengröße - und somit eine statistische Power- , die unterhalb dieser technischen Hürde blieb, so daß sich inhaltliche Interpretationen der ausbleibenden Effekte erübrigen.

Neben diesen Kritikpunkten ist außerdem die Praxis zu beklagen, anstelle von Intelligenztests eine - zumeist nicht einmal diskutierte - methodisch problematische Subskalenselektion vorzunehmen und die Applikation vereinzelter Aufgabentypen als

Intelligenzmessung auszugeben. Auch der Einsatz von Intelligenzverfahren mit umstrittener Qualität, wie z.B. dem IST-70, erleichtert nicht gerade die Interpretation der Daten. (Für eine kritische Würdigung des IST-70 siehe z.B. Brocke, Beauducel und Tasche, 1998; Schmidt-Atzert, Hommers und Heß, 1995)

9.1.2.4 Zusammenfassung: Intelligenz und Problemlösen

Die empirische Befundlage zum Zusammenhang von Intelligenz- und Problemlöseleistungen erscheint nur auf den ersten Blick uneinheitlich. Deutlicher wird das Bild, wenn man den Einfluß der oben genannten zusammenhangsmoderierenden Aspekte sowie die methodischen Voraussetzungen einiger Studien berücksichtigt. Von den inhaltlichen Aspekten kommt dabei insbesondere der Problemschwierigkeit und dem Wissen eine besondere Bedeutung für den Zusammenhang von Intelligenz- und Problemlöseleistungen bei.

Auch die Abhängigkeit der Zusammenhangsbefunde von den methodischen Voraussetzungen darf nicht unterschätzt werden: Läßt man einmal die Studien außen vor, die aufgrund einer oder mehrerer der in Abschnitt 9.1.2.3 genannten Mängel aus methodischen Gründen von vorne herein überhaupt keinen Zusammenhang zwischen Intelligenz- und Problemlöseleistungen aufdecken *konnten*, so ist das Bild weniger uneinheitlich als zunächst gedacht: Die Tatsache, daß sich in der überwiegenden Mehrzahl der methodisch soliden Studien ein empirischer Zusammenhang zwischen Intelligenz- und Problemlöseleistungen zeigt, spricht dafür, daß – entsprechend den Konstruktannahmen zur Intelligenz – in den Steuerungsleistungen ein überzufälliger Anteil an Intelligenzleistungen enthalten ist. Mit Werten im Bereich von $r=.30$ bis $r=.50$ sind Intelligenz- und Problemlöseleistungen aber deutlich niedriger miteinander korreliert als Leistungen in ein und derselben Intelligenzdimension, die mit verschiedenen Intelligenztests gemessen wurden. Allerdings ist die Höhe der Korrelationen vor dem Hintergrund der vergleichsweise geringen Reliabilität der Steuerungsleistungen (siehe Kapitel 8) und vor dem Hintergrund der oben (Abschnitt 9.1.1) referierten Befunde zur (fehlenden) Generalität der Problemlöseleistungen eher als hoch zu werten. Entgegen der Erwartung korrelieren Problemlöseszenarien mit Intelligenztests sogar tendenziell höher als untereinander.

Der empirische Zusammenhang zwischen Intelligenz- und Problemlöseleistungen entspricht in seiner Höhe der Korrelation zwischen den Leistungen in Skalen unterschiedlicher Intelligenzdimensionen oder aber auch den Trennschärfekoeffizienten einzelner Intelligenzaufgaben zur jeweiligen Gesamtskala. Diese Befunde sind vereinbar mit der Annahme, daß es sich bei Problemlöseszenarien um neue *Aufgabentypen* von Intelligenztests handelt, wobei mit diesen neuen Aufgabentypen eventuell ein Intelligenzaspekt erfaßt wird, der mit den bisherigen Aufgabentypen nicht voll-

ständig gemessen wird. Da zumindest für einige Problemlöseszenarien von einer noch hinreichenden Reliabilität ausgegangen werden kann (siehe Kapitel 8) könnten diese Aufgabentypen Indikatoren einer bislang unberücksichtigten Intelligenzfacette sein. Ebensogut ist es aber möglich, daß die systematische Varianz der Steuerungsleistung, die sich nicht auf herkömmliche Intelligenztestleistungen zurückführen läßt, auf bereits bekannte Konstrukte, z.B. auf Wissen, zurückgeht. Dies würde teilweise die Systemspezifität der systematischen Varianz der Steuerungsleistungen erklären. Die zuletzt genannte Überlegung wird im folgenden Abschnitt verfolgt.

9.1.3 Wissen und Problemlösen

9.1.3.1 Theoretische Überlegungen zum Zusammenhang von Wissen und Problemlösen

Sowohl bei abstrakten als auch bei semantisch eingekleideten Systemen müssen die Diagnostikanden zur Aufgabenbewältigung Wissen über das System erwerben, anwenden und aufgrund der Steuerungserfahrungen modifizieren. Bei semantisch eingekleideten Systemen spielt außerdem das Vorwissen und dessen (An-)Passung eine entscheidende Rolle für den Umgang mit dem Szenario. Im Kontext der Einbettung des Konstrukts Problemlösen im nomologischen Netz ist daher der Zusammenhang von Wissen und Problemlösen zu thematisieren. Dies gilt umso mehr, da das Wissen über das Problem und über Lösungswege als „interne Repräsentation" des Problems ein definierendes Attribut des Problemlösens ist. So wird für Kluwe (1990b, S. 121) ein Problem geradezu erst durch den Wissensmangel zu einem solchen. Auch für Dörner (1976) ist der Wissensmangel Ausgangspunkt der Problemlöseprozesse. Funke (z.B. 1992) beschreibt Problemlösen unter dem Aspekt des Wissenserwerbs als Konstruktion subjektiver Kausalmodelle.

Daß dem Wissen eine zentrale Bedeutung für das Problemlösen zukommt, ist weitgehend konsensuabel. Kontrovers diskutiert wird hingegen die Frage, welche Form von Wissen die Problemlöser zu welchem Zeitpunkt des Problemlöseprozesses mitbringen, erwerben, benötigen und oder anwenden. Süß et al. (1993a, S. 191) komprimieren die in der Literatur berichteten Wissenskonzepte auf die folgenden drei Unterscheidungen: (1.) deklaratives und prozedurales Wissen, (2.) Sach- und Handlungswissen sowie (3.) spezifisches und allgemeines Wissen. Die Kategorie des Sach- und Handlungswissens kann außerdem weiterhin nach dem „Präzisionsgrad" qualifiziert werden. Diese weitergehende Differenzierung läßt sich am Beispiel des Sachwissens über die Relation von zwei Variablen verdeutlichen. Funke (1990, S.151, mit anderer Terminologie: 1985 b, S.456) unterscheidet bei der *„Identifika-*

tion von Relationen" beispielsweise die folgenden drei Präzisionsgrade: das bloße Erkennen eines Zusammenhangs *(„Relationswissen")*, die Kenntnis der Richtung des Zusammenhangs *(„Vorzeichenwissen")* und die Kenntnis des exakten Gewichtungsfaktors *(„numerisches Wissen")*. Diese Dreiteilung des Präzisionsgrades (siehe auch die Dreier-Unterscheidung der Repräsentationsebenen bei Plötzner, Spada, Stumpf und Opwis 1990; Plötzner und Spada, 1992, S. 109) kann nach einem Vorschlag von Kersting (1991, S. 64 f.) zu einem Kontinuum mit den folgenden vier Stufen ausdifferenziert werden: „semi-qualitatives Wissen", „qualitatives Wissen", „semi-quantitatives" und „quantitatives Wissen". Dabei wird angenommen, daß es sich bei der Beziehung zwischen den Stufen von *quantitativ* zu *qualitativ* um eine einseitige Abhängigkeitsstruktur handelt. Eine Person, die über *„quantitatives"* Wissen verfügt, sollte demnach auch über alle „untergeordneten" Präzisionsgrade verfügen.

Die ersten beiden Wissensklassen – deklaratives und prozedurales Wissen einerseits und Sach- und Handlungswissen andererseits – werden oft zusammengefaßt, z.B. wird deklaratives Wissen mit dem Faktenwissen gleichgesetzt (z.B. Putz-Osterloh, 1988, S. 249). Tatsächlich können sowohl Sach- als auch Handlungswissen deklarativ sein. Die Unterscheidung zwischen Sach- und Handlungswissen betrifft nach Oberauer (1993b, S. 36) den *Inhalt* des Wissens, während die Differenz zwischen deklarativem und prozeduralem Wissen die *Rolle* des Wissens im Denken und Handeln betrifft. Mißverständnisse entstehen auch durch die vereinfachende Gleichsetzung von prozeduralem Wissen mit *implizitem Wissen*. Obwohl häufig prozedurales Wissen implizit ist, handelt es sich beim impliziten Wissen um eine *Teilmenge* des prozeduralen Wissens, nämlich um den Anteil prozeduralen Wissens, der einer Selbstbeobachtung und Explikation nicht zugänglich ist (Süß, Beauducel, Kersting & Oberauer, 1992).

Dem „impliziten Wissen" kommt nach Ansicht einiger Autoren eine herausragende Bedeutung für die Problembewältigung zu. Einige theoretische Überlegungen zur Bedeutung des impliziten Wissens für das Problemlösen haben ihren Ausgangspunkt allerdings in umstrittenen empirischen Befunden zur Dissoziation von explizierbarem Wissen und Problemlösen. In Experimenten der Oxforder Arbeitsgruppe (z.B. Berry & Broadbent 1984, 1987, 1988; Broadbent, FitzGerald & Broadbent, 1986) zeigte sich *kein Zusammenhang* zwischen der Steuerungsleistung bei den eingesetzten Kleinstsystemen einerseits und dem verbalisierten Wissen andererseits. Die nach Maßgabe der Wissensdiagnose erfolgreiche Induktion von Wissen verbesserte die nachfolgenden Steuerungsleistungen nicht. Die Steuerungserfahrung zeitigte zwar einen positiven Effekt auf die nachfolgende Steuerungsleistung, nicht aber auf das nachträglich verbalisiertes Wissen. Berry und Broadbent (1987, S.11f.) legten der Diskussion dieser Befunde die Annahme zugrunde, daß eine erfolgreiche Systemsteuerung ohne ein bestimmtes Wissen über die Verknüpfung bestimmter Sy-

stemvariablen unwahrscheinlich sei. Aus der empirischen Not wurde eine theoretische Tugend, indem das Ausbleiben der Zusammenhänge zwischen Wissen und Steuerungsleistungen die theoretische Begründung für die Einführung einer weiteren Wissensmodalität lieferte. Neben dem verbalisierbaren und somit erfassbaren *„expliziten"* Wissen - welches nichts zur Systemsteuerung beiträgt - müßte, so lautete vereinfacht wiedergegeben die Argumentation, noch eine weitere Modalität, das *„implizite Wissen"*, existieren, welches zwar die Steuerungsleistung, nicht aber unbedingt die Leistung im verbalen Wissenstest verbessern kann. Wer allein auf diese Argumentation setzt, blendet allerdings zahlreiche andere Interpretationsmöglichkeiten der Befunde - sofern diese überhaupt replizierbar und somit interpretationsbedürftig sind (zur Replizierbarkeit siehe z.B. die Übersichten von Berry, 1993; Haider, 1992) - aus. Als Grundproblem ist hier das von Buchner (1993, S. 4f.) für den Bereich des impliziten Lernens angeführte Exhaustivitätsproblem zu nennen, da der Nachweis impliziten Wissens im Prinzip voraussetzt, daß *alle* Versuche des Nachweises expliziten Wissens gescheitert sind.

Aber auch in sich ist die Interpretation der Experimente der Oxforder-Gruppe fraglich. So nimmt Kluwe (1990a, S. 245 f.) beispielsweise unter Verweis auf ingenieurpsychologische Studien an, daß Individuen komplexe Systeme auf der Basis *unvollständigen* Wissens handhaben. Haider (1991) konnte diese These empirisch bestätigen. In einer Replikationsstudie zeigte sie nicht nur, daß eine - im Vergleich zur Oxforder Arbeitsgruppe - andere Auswertung der Daten zum verbalen Wissen zu einem teilweise anderen Befund führt, sondern sie konnte vor allem nachweisen, daß das gegebene Problem mit einem einfachen Algorithmus auch *„ohne die Berücksichtigung bestimmter Systemverbindungen"* befriedigend gelöst werden kann (Haider, ebd, S. 72). Damit verliert die These zur Bedeutung des impliziten Wissens an Evidenz. Buchner, Funke und Berry (1995) unternahmen ebenfalls konzeptionelle Replikationen der Ausgangsstudien von Berry und Broadbent. Sie weisen im Kontext der Dissoziation von Wissen und Steuerungsleistungen darauf hin, daß gerade *schlechte* Systemsteuerer bei Aufgabenbearbeitungen ohne Explorationsphase in einem größeren Ausmaß mit verschiedenen Zustandsübergängen des Systems konfrontiert sind und dadurch mehr Systemwissen erwerben. Aufgrund weitergehender Analysen der Daten ihrer Replikationsstudien geben die Autoren außerdem zu bedenken, daß die für die Gesamtgruppe aufgezeigte Dissoziation nicht ausschließt, daß es Subgruppen gibt, für die Wissen und Steuerungsleistung stark assoziert sind.

Schließlich sollte man bei allen Untersuchungen zum Zusammenhang von Problemlösen und Wissen bedenken, daß vor einer inhaltlichen Interpretation der Daten zunächst deren Qualität geklärt sein sollte. Die Güte wissensdiagnostischer Instrumente wird in der Problemlöseforschung de facto kaum hinterfragt, Fortschritte der Wissensdiagnostik vollziehen sich eher auf konzeptioneller Ebene (z.B. Tergan,

1989a, 1989b) denn auf der Ebene der praktischen Messung. Der Erklärung bedarf daher möglicherweise nicht etwa ein Problemlösen ohne Wissen, sondern ein Simpleres: Untersuchungen zum Problemlösen ohne valide Wissensdiagnostik. Um diesen Aspekt besser einschätzen zu können, sollen die Möglichkeiten und Probleme der Wissensdiagnostik im folgenden im Rahmen eines kurzen Abschnitts abgehandelt werden. Erst wenn nachweislich psychometrisch befriedigende Verfahren zur Wissensdiagnose vorliegen, die den Besonderheiten des zu diagnostizierenden Gegenstands berücksichtigen, kann der Zusammenhang zwischen Wissen und Problemlösen empirisch geprüft werden. Theoretisch ist mit Süß et al. (1993a, S. 193) zu erwarten, daß das Ausmaß zutreffenden Vorwissens über den Realitätsbereich, der durch die Semantik des Problemlöseszenarios angesprochen wird, eine Determinante der Problemlöseleistungen ist. Der Einfluß dieses allgemeinen Vorwissens auf die Problemlöseleistung sollte aber mit zunehmender Systemerfahrung abnehmen, während der Einfluß des systemspezifischen Wissens - welches mit der Steuerungserfahrung erworben wird - mit der Systemerfahrung zunehmen sollte. Im übernächsten Abschnitt (9.1.3.3) werden einige empirische Befunde zum Zusammenhang von Wissen und Problemlösen vor dem Hintergrund dieser Überlegungen betrachtet.

9.1.3.2 Wissensdiagnostik

Potentielle Verfahren der Wissensdiagnostik werden z.B. in den Arbeiten von Arbinger (1991), Kluwe (1988) und Strohschneider (1990) sowie - speziell für die qualitative Wissensdiagnose - von Tergan (1988) vorgestellt. Kluwe unterscheidet vier *„Methoden zur Gewinnung von Daten über Wissen"*: Erstens die Methode des lauten Denkens, zweitens das Befragen, drittens das Kategorisieren und viertens die freie Wiedergabe. In der Problemlöseforschung wurde die Wissensdiagnose insbesondere über Methoden des Befragens realisiert.

Zahlreiche Verfahren zur Wissensdiagnostik sind mit Problemen behaftet, so bemängeln z.B. sowohl Kluwe (1988) als auch Strohschneider (1990) die mangelnde Objektivität und Validität der *„Methode des lauten Denkens"*. Dieser Einwand trifft auch die Methode der *„freien Wiedergabe"*. Andere Probleme der Wissensdiagnose sind grundsätzlicher Art und betreffen die Reaktivität der Messungen, die Veränderbarkeit des zu messenden Wissens, die damit einhergehenden Reliabilitätsprobleme und schließlich die Vielfalt möglicher Repräsentationsformen des Wissens. Zur Veranschaulichung der Reaktivität der Wissensmessung kann man sich beispielsweise vergegenwärtigen, daß die „Wissenstruktur", die man mit den Methoden des „Kategorisierens" (siehe z.B. die „Struktur-Lege-Technik" nach Scheele & Groeben, 1984) „diagnostiziert" hat, unter Umständen erst durch dieses diagnostische Verfahren forciert worden ist. Eine Möglichkeit, ein nicht-reaktives Meßverfahren zur

Diagnose des Strukturwissens zu erhalten, sieht Preußler (1996) in der Analyse von Priming-Effekten. In ihrem Ansatz hatte die Autorin allerdings Diagnosen über quantitatives Wissen ausgespart.

Die Reaktivität der Wissensmessungen ist experimentell zu kontrollieren, die Veränderbarkeit des Wissens erfordert mehrere Meßzeitpunkte, und die Vielfalt möglicher Repräsentationsformen des Wissens erfordert Methodenvielfalt auf Seiten der Wissensdiagnose.

Um Aufschluß über die Gültigkeit von Wissensmessungen im Kontext von Problemlöseverhalten zu erlangen, kommt der Methode der Kontentvalidität eine besondere Bedeutung zu. Voraussetzung für eine kontentvalide Testkonstruktion ist die Definierbarkeit von Itemuniversa. Diese „Definierbarkeit" ist oft angezweifelt worden (Loevinger 1965; Moser 1987; siehe aber Cronbach 1971, 454 f.). Tatsächlich ist es dieser Punkt, der die Verbreitung kontentvalider Testverfahren über die oftmals besonders übersichtlichen klassischen Anwendungsbereiche der Pädagogischen Psychologie (z. B.: „Grundrechnen") hinaus verhindert hat (siehe auch den Begriff der *„trivialen Validität"* bei Lienert 1967, S. 260). Für die Diagnose des deklarativen Sachwissens über computergestützte Problemlöseszenarien ist dieser Testkonstruktionsansatz jedoch indiziert, da die Grundgesamtheit des Testgegenstandes mit den Simulationsalgorithmen, die diesen Problemen zugrundeliegen, definiert werden kann. Kersting (1991) hat für das Problemlöseszenario „Schneiderwerkstatt" einen kontentvaliden Wissenstest zur Diagnose des deklarativen Sachwissens entwickelt, wobei sich das Vorgehen der Testkonstruktion an den für den deutschsprachigen Raum grundlegenden Arbeiten von K.J. Klauer (1983, 1984a, 1984b) orientierte. Die Annahme der Kontentvalidität wurde über den regelgeleiteten Konstruktionsprozeß hinaus durch die empirische Prüfung von solchen Hypothesen untermauert, die sich unmittelbar aus der Aufgabenanalyse ableiten ließen (Kersting, 1991; Kersting & Süß, 1995). Während es im ungünstigsten Fall bei der konventionellen Entwicklung von Wissenstests der *„idiosynkratischen Auffassung"* (Hornke & Habon, 1984, S.204) bzw. den *„Interessen oder der Intuition des Testkonstrukteurs"* (Feger, 1984, S.24) überlassen bleibt, *was* und *auf welche Art und Weise* in Form der vorgegebenen Items zum Gegenstand der Untersuchung gemacht wird, tritt bei der Entwicklung eines kontentvaliden Wissenstests die regelgeleitete Testkonstruktion an die Stelle der Subjektivität des Testkonstrukteurs.

9.1.3.3 Befunde zum Zusammenhang von Wissen und Problemlösen

Der Einfluß des Vorwissens auf das Problemlösverhalten wurde nur selten direkt untersucht. Die weiter oben (Abschnitt 4.3) berichteten Befunde zum Experten-Novizen Vergleich können diesbezüglich keine verbindliche Auskunft geben, da das

größere (Vor-)Wissen der Experten häufig nur postuliert, nicht aber gemessen wurde. Schlußfolgerungen auf die Bedeutung des Vorwissens für die Problemlöseleistungen können aber aus solchen Untersuchungen abgeleitet werden, in denen die semantische Einkleidung eines Systems variiert wurde. So wurde bereits erwähnt, daß das klassische Problem „Turm-von-Hanoi" in der ursprünglichen gegenständlichen Einkleidung leichter zu lösen ist als in der semantisch unvertrauten und willkürlich anmutenden „Monster"-Variante (siehe z.B. Klauer, 1993). Auch die Arbeit von Hesse (1982, 1985) zur Variation des semantischen Kontextes des „Dori" Problems (siehe Abschnitt 9.1.2.2) kann im Zusammenhang von Vorwissenseffekten interpretiert werden. Die semantische Variante wäre dieser Interpretation zufolge deshalb einfacher zu steuern, weil die Probanden hier - im Gegensatz zur „nichtsemantischen" Version - ihr Vorwissen einbringen konnten. Am deutlichsten wurde der Effekt des Vorwissens bislang in dem ebenfalls bereits angesprochenen Experiment von Funke (1992, S. 120 ff.) zur Variation der Vorwissensverträglichkeit (siehe Abschnitt 2.3.2.2) herausgearbeitet. Das „Altöl"- Szenario, welches dem Vorwissen der Probanden entsprach, wurde erfolgreicher gesteuert als das vorwissensunverträglich gestaltete Szenario.

In den Studien der Oxforder Arbeitsgruppe um Broadbent (die weiter oben im Kontext der Debatte um das implizite Wissen thematisiert wurden) zeigte sich eine Dissoziation von Wissen und Steuerungsleistungen. Auch in den Studien von Leutner (1989), Morris und Rouse (1985), Putz-Osterloh (1993b), Putz-Osterloh, Bott und Houben (1988), Preußler (1996), Renkl et al. (1994) sowie Strohschneider (1990) zeigte sich kein kohärenter Zusammenhang zwischen Wissen und Steuerungsleistungen. Demgegenüber berichten z.B. Beckmann (1994), Funke (1985b und 1992), Müller (1993), Putz-Osterloh (1985 und 1987), Reichert und Dörner (1988), Süß et al. (1993a), Vollmeyer, Burns und Holyoak (1996) sowie Wittmann et al. (1996) über positive Zusammenhänge zwischen Wissen und Problemlösen.

Die Ergebnisse der Berliner Untersuchung sollen hier etwas ausführlicher dargestellt werden, da im empirischen Teil der Arbeit einige Instrumente der Berliner Untersuchung erneut zum Einsatz kommen. In der Berliner Studie wurde zum einen allgemeines Vorwissen (Wirtschaftswissen) und zum anderen systemspezifisches Wissen erhoben. Das Wirtschaftswissen wurde vor der Systemsteuerung erhoben und war mit $r=.38$ ($N=148$) ein guter Prädiktor der nachfolgenden Steuerungsleistung (Süß et al., 1993a, S. 198). Das systemspezifische Wissen wurde erstmals nach der Programm-Einführung und somit nach den zwei „Übungsmonaten" erhoben. Damit kann nicht abschließend beurteilt werden, ob es sich um „Vorwissen" oder bereits um durch Beobachtung erworbenes Wissen handelt. Vier der sechs Wissenskalen korrelierten in Höhe von $r=.22$ bis $r=.37$ signifikant mit der folgenden Steuerungsleistung. Mit dem nach der ersten Steuerung erneut erhobenen sy-

stemspezifischen Wissen konnte die Steuerungsleistung in den beiden anschließenden weiteren Steuerungsdurchgängen zu r=.40 vorhergesagt werden. Diese Befunde zur deutlichen Assoziation von Wissen und Problemlöseleistungen konnten in der Wiederholungsuntersuchung repliziert werden.

9.1.3.4 Zusammenfassung: Wissen und Problemlösen

Zusammenfassend ergibt sich für die Verbindung zwischen Wissen und Problemlösen eine disparate Befundlage. Wie im Falle des Zusammenhangs zwischen Intelligenz und Problemlösen können auch hier zur Erläuterung sowohl inhaltliche als auch methodische Aspekte berücksichtigt werden. Hinsichtlich inhaltlicher Aspekte kann beispielsweise eine Anregung von Müller (1993, S. 111) angeführt werden. Der Autor überlegt - unterstützt durch ein kleines Experiment mit 20 Studenten - ob der Zusammenhang von Wissen und Steuerungsleistungen durch die Systempräsentation moderiert wird. Diesem Gedanken zufolge soll vor allem eine simultane Systempräsentation, die eine Analyse der Systemverläufe begünstigt, den Wissenserwerb und damit den Zusammenhang von Wissen und Problemlösen positiv beeinflussen. Demgegenüber soll bei einer sequentiellen Präsentation im Extremfall kein Wissen erworben werden. Dieser potentielle Moderatoreffekt kann allerdings nur einen Teil der widersprüchlichen Untersuchungsergebnisse aufklären.

Einflußreicher als dieser inhaltliche Aspekt könnten die methodischen Probleme sein. Tendenziell sind es häufiger diejenigen Untersuchungen, bei denen Wissen und Problemlösen unkorreliert blieben, die deutlichere methodische Schwächen aufweisen. Sowohl die Tatsache, daß in den meisten dieser Studien die psychometrische Güte der eingesetzten Instrumente zur Wissensdiagnose ungeklärt blieb als auch die hohe Spezifität der oft auf singuläre Wissensaspekte abzielenden Wissensdiagnosen sprechen gegen die Aussagekraft der Ergebnisse. Es ist davon auszugehen, daß die Problemlöseleistungen - wie in zahlreichen Untersuchungen gezeigt - nicht nur durch Intelligenz, sondern des weiteren auch durch Wissen determiniert sind. Es spricht vieles dafür, daß das Wissen für diejenige systematische Varianz der Steuerungsleistung verantwortlich ist, welche nicht auf herkömmliche Intelligenztestleistungen zurückgeführt werden kann. Problemlösen ließe sich somit als eine Interaktion von Intelligenzleistungen und Wissen beschreiben. Unter diesen Umständen müßten Intelligenz und Wissen gemeinsam die beste Vorhersage der Steuerungsleistungen ermöglichen. Während Intelligenz und Wissen in den bisherigen Abschnitten getrennt thematisiert wurden, beschäftigt sich der folgende Abschnitt mit der Vorhersage der Problemlöseleistungen durch eine Einheit von Intelligenz ***und*** Wissen.

9.1.4 Problemlösen als Integration von Intelligenz und Wissen

9.1.4.1 Theoretische Überlegungen zum Zusammenhang einer Einheit aus Intelligenz und Wissen mit dem Problemlösen

In den letzten Abschnitten zur Frage der Einordnung der Problemlösefähigkeit in das nomologische Netz wurde das Problemlösen einerseits mit der Intelligenz und andererseits mit dem Wissen in Zusammenhang gebracht. Intelligenz und Wissen können aber - abstrahiert man einmal die Konstrukte von ihren jeweiligen Repräsentationen in Testaufgaben - auch in ihren wechselseitigen Bezügen betrachtet werden, etwa im Sinne Cattells Theorie der fluiden und kristallinen Intelligenz (Cattell, 1957, 1963, 1971; Horn, 1980). Beim Problemlösen wirken Intelligenz und Wissen nicht getrennt, sondern in Interaktion miteinander: intellektuelle Fähigkeiten sind nach Ansicht von Süß et al. (1993a, S. 192f.) eine notwendige Voraussetzung für den Erwerb, die Anwendung und die Modifikation von Wissen. (Neben den intellektuellen Fähigkeiten hängt das Ausmaß des Wissens allerdings auch von anderen Gegebenheiten, etwa von den Lerngelegenheiten und der Motivation zum Nutzen der Lerngelegenheiten ab). Auf empirischer Ebene ist daher u.a. zu erwarten, daß die Problemlöseleistungen mit einer *Einheit* von Intelligenz und Wissen (operationalisiert z.B. als Aggregat der Intelligenz- und Wissensmessungen) im engeren Zusammenhang stehen als mit den jeweiligen Einzelmaßen für Intelligenz oder Wissen.

9.1.4.2 Empirische Befunde zum Zusammenhang einer Einheit aus Intelligenz und Wissen mit dem Problemlösen

Bislang wurden nur in den Arbeitsgruppen um Süß Intelligenz und Wissen als Einheit der Problemlösefähigkeit empirisch gegenübergestellt, obwohl auch die Daten anderer Arbeitsgruppen die Voraussetzungen für solche Analysen erfüllen. In den Berliner (Süß et al., 1991, 1993a) und Mannheimer (Wittmann et al., 1996) Studien zeigte sich zunächst, daß zumindest das deklarierbare Sachwissen substantiell mit der operativen Intelligenzkomponente „Verarbeitungskapazität" korreliert ist. Trotz dieser hohen Korrelation erwiesen sich in Regressionsanalysen sowohl Wissens- als auch Intelligenzmaße als *eigenständige* Prädiktoren der Steuerungsleistungen. In der Berliner Erstuntersuchung erlaubten Wissen und Intelligenz gemeinsam mit einer multiplen Korrelation in Höhe von $R = .60$ eine gute Vorhersage der ersten Steuerungsleistung; in der Wiederholungsuntersuchung betrug die entsprechende multiple Korrelation $R = .59$. (Indikator der Steuerungsleistungen: die über die beiden Steuerungsdurchgänge aggregierten Maße.)

Da die Steuerungsleistungen in diesem Ausmaß durch Intelligenz und Wissen

vorhergesagt werden können, stellt sich die Frage, ob mit der Szenarienbearbeitung überhaupt noch eine eigenständige Fähigkeit indiziert wird, die über die beiden bereits etablierten Fähigkeitskonstrukte hinaus geht. Zur Klärung dieser Frage partialisierte Süß (1996, S. 194 f.) in einer Analyse der Berliner Daten die Intelligenz- und Wissensvarianz aus der Steuerungsleistung aus. (Zum methodischen Vorgehen siehe unten, Abschnitt 15.4 und Tabelle 18.) Während die Steuerungsleistungen der beiden Meßzeitpunkte ursprünglich zu r=.43 miteinander korrelierten, standen die für die Erst- und die Wiederholungsuntersuchung gebildeten, um Intelligenz- und Wissensanteile „bereinigten" Residuen in keinem überzufälligen Zusammenhang mehr zueinander. Dies bedeutet, daß der Anteil der systematischen Varianz auf Intelligenz *und* Wissen zurückgeführt werden kann und die Annahme einer essentiellen Problemlösefähigkeit empirisch unbegründet ist.

9.1.5 Problemlösen und nicht-kognitive Personmerkmale

9.1.5.1 Theoretische Überlegungen zum Zusammenhang von Problemlösen und nicht-kognitiven Personmerkmalen

Es ist an verschiedenen Stellen in der Literatur auf die Notwendigkeit einer *ganzheitlichen*, einer *systemischen* Betrachtung der Problemlöseleistung hingewiesen worden. Dieser „Ganzheitlichkeit" zufolge ist das Problemlösen nicht nur im Kontext kognitiver, sondern auch im Kontext nicht-kognitiver Personmerkmale zu sehen. Insbesondere eine hohe emotionale Stabilität und heuristische Kompetenz (z.B. Stäudel, 1987, 1988), eine gewisse Resistenz gegenüber Reizüberflutung und Stress sowie eine positive Problem-Konfrontationsbereitschaft (Kreuzig, 1981) sollen ebenso wie die Selbstsicherheit das problemlösende Verhalten begünstigen. In dem folgenden Abschnitt werden einige empirische Ergebnisse zu diesen Punkten versammelt. Vorab ist aber darauf hinzuweisen, daß die konzeptionelle Trennung zwischen kognitiven und nicht-kognitiven Personmerkmalen nicht immer eindeutig ausfällt. So ordnet Funke (1990, S. 146) beispielsweise die heuristische und epistemische Kompetenz den kognitiven Merkmalen zu, da diese auf dem heuristischen und epistemischen Gedächtnis aufbauen. Andererseits postuliert Dörner (1988, S. 276) eine direkte Beziehung zwischen dem Wissen als kognitiven Merkmal und der emotionalen Lage (als emotionales Merkmal) sowie der Kompetenz als Ingredienz der emotionalen Lage.

9.1.5.2 Empirische Befunde zum Zusammenhang von Problemlösen und nicht-kognitiven Personmerkmalen

Untersuchungen zur Bedeutung nicht-kognitiver Personmerkmale für das Problemlösen konzentrierten sich bislang überwiegend auf die Merkmale Selbstreflexion, Emotion und Motivation sowie auf Angst und Selbstsicherheit als Persönlichkeitsmerkmale im engeren Sinne.

In der „Lohhausen" Studie (Dörner et al., 1983b) wurden Kennwerte für zahlreiche nicht-kognitive Personmerkmale erhoben. Lediglich die Selbstsicherheit, die Extraversion (mit einigen Einschränkungen) und die - nachträglich erhobenen - kognitiven Prozeßvariablen (erhoben mit dem Fragebogen für kognitive Prozeßvariablen, „FPK", siehe Kreuzig, 1981) erwiesen sich als bedeutsam für die Prognose der Leistungen der einzelnen „Bürgermeister". Während Hesse (1982) die Befunde für alle drei Merkmale nicht bestätigen konnte, zeigte sich die Bedeutung der postitiven Ausprägung im „FPK"-Fragebogen für das Problemlösen auch in der Studie von Kühle und Badke (1986). Diejenigen Problemlöser der Gesamtgruppe von 20 Personen, die eine vergleichsweise positive Steuerungsleistung erbrachten, zeichneten sich außerdem auch durch eine hohe heuristische Kompetenz und eine hohe emotionale Stabilität aus. Die Befunde von Kühle und Badke (1986) hinsichtlich der heuristische Kompetenz und der emotionale Stabilität decken sich teilweise mit den Ergebnissen von Stäudel (1987). Stäudel ließ 43 Personen das Szenario „Moro" unter erschwerten Voreinstellungen steuern und eine Reihe von Persönlichkeitsfragebogen ausfüllen. Die Personen, die sich als kompetenter und weniger emotional belastet beschrieben, erzielten hinsichtlich zahlreicher Gütevariablen bessere Werte. Nach Ansicht von Dörner (1985, S. 164ff.) wirkt sich die Kompetenz über Emotionen vermittelt auf das Problemlöseverhalten aus. Diese Annahme einer leistungsfördernden Wirkung der aktuellen Kompetenz auf das Problemlöseverhalten konnte aber in einer längstschnittlich angelegten und pfadanalytisch ausgewerteten Untersuchung von Köller, Strauß und Sievers (1995) nicht bestätigt werden. Die Tauglichkeit des Kompetenzfragebogens als Meßinstrument und vor allem als Indikator für heuristische Kompetenz muß nach einer kritischen Analyse von Köller und Strauß (1994) grundsätzlich in Frage gestellt werden. In der Reanalyse der Autoren zeigten sich bei 94 Probanden darüber hinaus keine überzufälligen Korrelationen der Skalen des Kompetenzfragebogens mit der Steuerungsleistung im „Heizölhandel".

Die Kombination der Problemlösefähigkeit mit der Selbstsicherheit, die sich in der „Lohhausen"-Studie zeigte, ließ sich in Studien von Hesse (1982), Putz-Osterloh (1985) und Rhenius (1994) nicht replizieren. In diesen Untersuchungen wurde die Selbssicherheit mit dem gleichen Fragebogen erfaßt, der auch in der „Lohhausen" Studie zum Einsatz kam. Rhenius (ebd.) zweifelt daran, daß dieser Un-

sicherheitsfragebogen (Ullrich de Muynck und Ullrich, 1977, zitiert nach Rhenius, 1994) aus dem Bereich der experimentellen Therapieforschung ein geeignetes Instrument für die Problemlöseforschung darstellt. In seiner Studie zeigte sich der erwartete Zusammenhang zwischen Selbstsicherheit und Steuerungserfolg nur für Versuchspersonen mit unterdurchschnittlichen Selbstsicherheitswerten - allerding nahmen insgesamt nur 15 Personen an der zitierten Untersuchung teil.

Hesse et al. (1983) konnte den Einfluß motivationaler und emotionaler Faktoren auf das Problemlöseverhalten experimentell nachweisen. Die Autoren manipulierten durch eine Variation des semantischen Kontextes die „persönliche Betroffenheit" der Probanden, die eine Epidemie zu bekämpfen hatten. Während es für eine Gruppe bei der Steuerung des Systems um die Bekämpfung eines „harmlosen" leichten Grippevirus ging, hatte es die andere Gruppe bei der Steuerung des formal gleichen Systems vorgeblich mit der Bekämpfung einer gefährlichen Epidemie zu tun. Insbesondere Erfolgsmotivierte innerhalb der Gruppe der stark „betroffenen" Problemlöser steuerten das System ausdauernder und besser.

Der Einfluß nicht-kognitiver Personmerkmale auf das Problemlöseverhalten soll sich auch zeigen, indem man Probanden, die unter Lärmstreß arbeiten mußten, mit solchen Probanden vergleicht, die ungestreßt arbeiten konnten. In einem entsprechenden Experiment von Dörner und Pfeifer (1991, 1992) mit insgesamt 40 Personen zeigten sich allerdings keine Leistungsunterschiede zwischen den beiden Gruppen, lediglich eine Tendenz zur Problemdekompensation, Überreaktion und Schwerpunktbildung soll bei den gestreßten Teilnehmernzu erkennen gewesen sein.

Hussy und Granzow (1987) unternahmen ebenfalls eine experimentelle Studie, ihr Augenmerk galt dem Einfluß des Personmerkmals „Reflexivität-Impulsivität des Verarbeitungsstils" in Kombination mit der Bedeutung der Informationsdarbietung. Vor allem für impulsive Personen zeigte sich, daß die Nutzbarkeit von Informationen eine Funktion der Informationsart und des Präsentationszeitpunktes ist.

Über den in Bezug auf die Bedeutung der nicht-kognitiven Personvariablen korrelativen Ansatz der Berliner Untersuchung berichtet Süß (1996, S. 186f.). Zwischen der Leistungsorientierung, Selbstwirksamkeitserwartungen und Teilnahmemotivation einerseits und der Steuerungsleistung bei der „Schneiderwerkstatt" andererseits ergaben sich moderate Zusammenhänge. In der Mannheimer Untersuchung (Wittmann et al., 1996) zu den Determinanten des Problemlösens wurden zum Teil die gleichen Persönlichkeitsskalen eingesetzt wie in Berlin. Diesmal standen die Skalen (z.B. Leistungsmotivation, Belastbarkeit, Gewissenhaftigkeit, soziale Orientierung, Selbstwirksamkeitserwartungen, biografischer Fragebogen) aber in keinem überzufälligen Zusammenhang mit der der Steuerungsleistung.

Hasselmann (1993) berichtet moderat negative Zusammenhänge zwischen einzelnen Testwerten des „16 PF" -Persönlichkeitsfragebogens und der Steuerungs-

leistung in der „Textilfabrik", wobei insbesondere Beeinträchtigungen der emotionalen Stabilität dem Problemlöseerfolg tendenziell abträglich sind.

Für die Arbeiten zum Problemlösen im Kontext von *Persönlichkeitsmerkmalen im engeren Sinne* können z.B. die Studien von E. Müller (1991) sowie Stöber (1996) genannt werden, die mit Hilfe der Simulation „Risiko" Unterschiede in der Strategie Hoch- und Niedrigängstlicher untersuchten. In der Studie von Stöber steuerten 60 Personen das Szenario „Risiko". Dabei wählten die aufgrund der Ergebnisse im STAI-Trait-Fragebogen als „hochängstlich" klassifizierten Personen eine - im Vergleich zu den „Niedrigängstlichen" - engere Perspektive, indem sie sich primär um zentrale Probleme und um Folgeprobleme kümmerten.

Schließlich sind auch die Berichte über Zusammenhänge zwischen Problemlöseverhalten und Assessment-Center Urteilen im Kontext der Bedeutung nicht-kognitiver Personmerkmale für das Problemlösen auszuwerten. Dies gilt zumindest dann, wenn es um Assessment-Center Urteile über primär nicht kognitive-Fähigkeiten, wie z.B. das Führungspotential, geht. Laut Hasselmann (1993, S. 181) sollten die Steuerungsleistungen bei der „Textilfabrik" z.B. insbesondere mit Indikatoren des Führungspotentials kovariieren. Dieser Zusammenhang stellte sich in seiner Studie dann auch ein - überraschenderweises war diese Verbindung aber am deutlichsten für jenes Problemlösegütemaß, dem die geringste Reliabilität attestiert worden war, nämlich dem Kapitalendwert. Für die deutlich reliableren Trendmaße ergaben sich diesbezüglich nominell geringere Korrelationen. Außerdem zeigten sich Zusammenhänge zwischen der Steuerungsleistung und dem Assessment-Center-Urteil über die Belastbarkeit der Führungsnachwuchskräfte. Die Urteile über verbale Ausdrucksfähigkeit und den Teamgeist korrelierten hingegen nicht mit der Steuerungsleistung.

In einer Studie von Putz-Osterloh und Haupts (1989) sowie Putz-Osterloh und Schroiff (1987) standen die Assessment-Center-Urteile und die Indikatoren des Problemlösens in keiner klaren empirischen Relation zueinander. Vierzehn unterschiedliche Persönlichkeitsmerkmale von 32 Bewerber wurden in dem Assessment-Center ähnlichen Verfahren (in dem auch Testverfahren zur Anwendung kamen) von drei Beurteilern der Offiziersbewerberzentrale im Konsens-Verfahren eingeschätzt. Außerdem erhielten die Bewerber ein Gesamt-Eignungsurteil. Die Autorinnen der Studie wählten aus diesen 14 Merkmalen vier aus und korrelierten diese Einschätzungen mit der Steuerungsleistung bei der „Schneiderwerkstatt" sowie mit vier inhaltlich entsprechenden Strategiemerkmalen, die die Autorinnen anhand des Steuerungsverhaltens (Daten des Rechnerprotokolls und aus den Protokollen des lauten Denkens) bei der Bearbeitung des Szenarios gebildet hatten. Von den so gebildeten „Strategiemerkmalen" standen die Merkmale „Entscheidungsfähigkeit" und „Organisationsfähigkeit" in keinem systematischen Zusammenhang mit den Merkmalseinschätzungen aus dem Assessment-Center, auch die Steuerungsleistung ließ sich

nicht mit den Verhaltenseinschätzungen der Offiziersbewerberzentrale vorhersagen.

Vermutlich stehen auch die Leistungen im Szenario „Manage" in keinem systematischen Zusammenhang zu Assessment-Center Urteilen. Diesbezüglich listet Kreuzig (1995b, S. 392f.) Einzelfall um Einzelfall auf, bei denen Assessment-Center Ergebnis und Steuerungserfolg nicht übereinstimmten. Die zur Beurteilung der Fälle notwendigen Angaben zu dem Assessment-Center Verfahren fehlen aber ebenso wie Angaben zur Gesamtstichprobe oder gruppenstatistische Werte.

9.1.5.3 Zusammenfassung: Problemlösen und nicht-kognitive Personmerkmale

Zum Zusammenhang von Problemlösen mit nicht-kognitiven Personmerkmalen liegen bislang vergleichsweise wenige Studien mit widersprüchlichen Ergebnissen vor. Selbst in den Studien, in denen sich ein Effekt für nicht-kognitive Personmerkmale zeigte, handelte es sich um kleine Effekte, die bislang nicht konsistent repliziert werden konnten. Ähnlich lockere Assoziationen existieren auch zwischen Intelligenztestleistungen und nicht-kognitiven Personmerkmalen (wie z.B. Leistungsmotivation, Selbstwirksamkeit und Testängstlichkeit), ohne daß deshalb das Konstrukt „Intelligenz" im nicht-kognitiven Bereich verankert werden muß.

Gegen eine substantielle Verbindung zwischen der Problemlösefähigkeit und nicht-kognitiven Personmerkmalen spricht auch, daß in einigen der aufgeführten Studien der Effekt der nicht-kognitiven Personmerkmale möglicherweise auf die besondere Untersuchungssituation zurückgeführt werden muß. In der Mehrzahl der Studien, in denen sich eine Assoziation zwischen Problemlösen und nicht-kognitiven Personmerkmalen abzeichnete, wurde die Systemsteuerung nämlich über einen *Versuchsleiter* vermittelt. Neben den kognitiven Anforderungen wird dem Problemlöser in einer solchen Situation zusätzlich die *Kommunikation* seines Problemlöseverhaltens abverlangt, womit zweifelsohne nicht-kognitive Merkmale zum Tragen kommen. Der Problemlöser muß sich überwinden, Fragen zu stellen und/oder Fragen zu wiederholen, er muß Anweisungen erteilen, die Reaktionen des Versuchsleiters verarbeiten, er steht in all seinen Maßnahmen unter sozialer Kontrolle usw. Die bei der über einen Versuchsleiter vermittelten Systemsteuerung auftretenden Effekte nicht-kognitiver Personmerkmale können nicht ursächlich mit dem Konzept „Problemlösefähigkeit" oder den Szenarien als Meßinstrumente verknüpft werden.

Insgesamt wird daher hier die These vertreten, daß der Einfluß von nicht-kognitiven Personmerkmalen auf die Problemlöseleistungen nicht größer ist als auf Intelligenztestleistungen, und daß Problemlöseszenarien *primär* kognitive Anforderungen stellen. Die vorliegende Arbeit betont den kognitiven Aspekt des Problemlösens und zielt darauf ab, eine Einschätzung darüber abzugeben, welches Potential computergestützten Problemlöseszenarien als Alternative zur Intelligenzdiagnostik zukommt.

9.2 Kriteriumsvalidierung

Über die Brauchbarkeit diagnostischer Verfahren entscheidet u.a. die Enge der Relation zwischen den Verfahrensergebnissen einerseits und den Bewährungskriterien andererseits. Abschnitt 9.2.3 gibt einen Überblick über einige empirische Studien zur Kriteriumsvalidität von Problemlöseszenarien. Bei der Interpretation dieser Ergebnisse sowie bei der Interpretation der weiter unten – in den Kapiteln 16 und 17 – berichteten Befunde sind zwei Aspekte des Ansatzes der Kriteriumsvalidität zu beachten, die vorab dargestellt werden. Zunächst gilt es zu berücksichtigen, daß die Relation zwischen Prädiktor und Kriterium zahlreichen inhaltlichen und methodischen Einflußfaktoren unterliegt, die im folgenden Abschnitt (9.2.1) am Beispiel der Eignungsdiagnostik erläutert werden. Außerdem wird dem Bericht über die empirische Befundlage zur Kriteriumsvalidität von Problemlöseszenarien noch ein Hinweis auf die allgemeine und die für den Polizeiberuf spezifische Kriteriumsvalidität von Intelligenztests vorangestellt (Abschnitt 9.2.2). Will man den Nutzen neuer Verfahren einschätzen, empfiehlt sich nämlich insbesondere ein Vergleich der Vorhersagegenauigkeit des neuen Verfahrens (Problemlöseszenarien) mit der Treffsicherheit alternativer oder konkurrierender Verfahren (Intelligenztests).

9.2.1 Zur Abhängigkeit der Kriteriumsvalidität von inhaltlichen und methodischen Einflußfaktoren (dargestellt am Beispiel der Eignungsdiagnostik)

Die Kriteriumsvalidität eignungsdiagnostischer Verfahren hängt nicht nur von den Eigenschaften des zu validierenden Verfahrens und den Umständen seiner Anwendung ab (siehe z.B. Jäger, 1986, S. 281). Die Validität wird darüber hinaus im wesentlichen durch drei weitere Faktoren beeinflußt, nämlich durch (1) die Zielgruppe der Validierung, (2) durch das gewählte Kriterium und (3) durch die Beziehung zwischen Prädiktoren und Kriterium (siehe Tabelle 3).

Hinsichtlich der ***Zielgruppe*** sind moderiende Effekte u.a. für die *Art* der (1a) *Berufe* und *Ausbildungen* (z.B. einfache Tätigkeiten versus Führungspositionen; praktische Ausbildung versus theoretische Ausbildung) sowie (1b) für den *Grad der Vorselektion* der jeweiligen Gruppen zu erwarten.

Die Kriteriumsvalidität wird außerdem durch die ***Wahl des Kriteriums*** bestimmt. Hier gilt es vor allem die Effekte der (2a) der Art und der (2b) Qualität des Kriteriums zu beachten.

Hinsichtlich der *Art des Kriteriums* macht es z.B. einen Unterschied, ob Ausbildungs- oder Berufsleistungen herangezogen werden. (Dabei läßt sich differenzieren, zu welchem Zeitpunkt des Ausbildungs- oder Berufsabschnitts die Leistungen ge-

messen werden). Ebenso wirkt es sich aus, ob spezifische Verhaltensweisen oder ein „Gesamtverhalten", durchschnittliche oder Maximalleistungen, vorhergesagt werden sollen. Zu berücksichtigen ist schließlich, ob der Validierung ein ein- oder ein mehrdimensionales Kriterium zugrundegelegt, ob mit einem singulären Kriterium oder mit mehreren (in welcher Form kombinierten?) Kriterien gearbeitet wird.

Bei der *Qualität der Kriterien* geht es einerseits um deren *metrische* Qualitäten (z.B. Zuverlässigkeit und Diskriminationsfähigkeit). Andererseits geht es um deren *inhaltliche* Qualtität. Unter inhaltlichen Qualitätsgesichtspunkten können z.B. subjektive und weniger subjektive Kriterien unterschieden werden. Innerhalb der Klasse der weniger subjektiven Kriterien (gängigerweise - aber unpräzise - als „objektive Kriterien" bezeichnet) lassen sich direkte Kriterien (z.B. Produktionsstückzahl, Fehlerquote, Unfälle) und indirekte Kriterien (Gehalt, Beförderung, Fehlzeiten) abgrenzen. Aber auch ein und dasselbe Kriterium kann mit wechselnder inhaltlicher Güte (z.B. Qualität der Vorgesetztenbeurteilung) erfaßt werden. Vor allem geht es bei der Qualität der Kriterien um deren inhaltliche Gültigkeit oder Relevanz. Diesbezüglich sind zwei Quellen potentieller Irrelevanz zu minimieren. Zum einen ist darauf zu achten, nach Möglichkeit solche Anteile an der gemessenen Kriteriumsvarianz zu minimieren, die nicht im Kriteriums*konstrukt* enthalten sind (Kontamination). Zum anderen ist die *Defizienz* der Kriteriumsmessung zu minimieren. Eine Kriteriumsmessung ist in dem Maße defizient, in dem zum Kriteriumskonstrukt gehörende Varianzquellen in der Kriteriumsmessung unberücksichtigt bleiben.

Schließlich wird die Kriteriumsvalidität auch von der ***Beziehung*** zwischen Prädiktor und Kriterium beeinflußt. Hierbei spielt (3a) der *zeitliche Abstand* zwischen Prognose und Kriteriumsmessung ebenso eine Rolle wie die (3b) *Vergleichbarkeit* der beiden Maße. Unter dem Aspekt der Vergleichbarkeit ist z.B. der von Wittmann (1988, Wittmann & Matt, 1986) herausgearbeitete Gesichtspunkt der Symmetrie zwischen Prädiktor und Kriterium zu berücksichtigen. Schließlich spielt natürlich auch die (3c) *Art der Beziehung* zwischen Prädiktor und Kriterium (z.B. linear/nonlinear) sowie (3d) die *Art der Analyse* dieser Beziehungen eine Rolle (z.B. der Einsatz von einfachen oder multiplen Analyseverfahren [letzteres mit oder ohne Kreuzvalidierung], die Berücksichtigung von Moderator- und Suppressor-Variablen).

Einige der hier genannten verfahrensexternen Einflußfaktoren auf die Kriteriumsvalidität eignungsdiagnostischer Verfahren sind mittlerweile häufig untersucht worden, nur wenige Beispiele sollen anhand der Kriteriumsvalidität von Intelligenztests die Richtung der Effekte andeuten. So wurde für die Kategorie der Auswahl des Kriteriums z.B. immer wieder berichtet, daß die Validität von Intelligenztests höher ausfällt, wenn Ausbildungsleistungen vorhergesagt werden als wenn berufliche Leistungen als Kriterium herangezogen werden (z.B. Hunter & Hunter, 1984). Mit Selektionseffekten muß insbesondere bei anspruchsvollen Berufstätigkeiten gerechnet

werden. So diskutieren Funke, Krauss, Schuler und Stapf (1987, S. 419) das Ergebnis einer Metaanalyse, derzufolge Intelligenztests kaum zur Vorhersage wissenschaftlicher Leistungen im Bereich Forschung und Entwicklung beitragen, beispielsweise vor dem Hintergrund möglicher Selektionseffekte. Als ein Beispiel für den Einfluß des jeweils gewählten Kriteriums kann der Befund angeführt werden, daß unter allen Standard- Validierungskriterien die gängstige, die Vorgesetztenbeurteilung - neben der Flutktuation - durch Intelligenztests am schlechtesten prognostizierbar ist (siehe z.B. Schmitt, Gooding, Noe & Kirsch, 1984). Die Vorhersagbarkeit der Kriterien durch Intelligenztests nimmt mit deren Subjektivitätsgrad ab (ebd.). Anderen Faktoren kommt hingegen ein geringerer Einfluß zu, als ursprünglich vermutet. Die lange Zeit über präferierte Auffassung, die Kriteriumsvalidität kognitiver Testverfahren sei zu einem hohen Grade (berufs-)situationsspezifisch, konnte beispielsweise im Laufe der Zeit durch die Ergebnisse von Meta-Analysen widerlegt werden. Intelligenztests können bei praktisch allen bislang untersuchten Berufen zur Leistungsprognose beitragen. Auch die Annahme, mit zunehmender Komplexität der beruflichen Anforderungen, etwa bei Führungsaufgaben, würde der mit Intelligenztests vorhersagbare Varianzanteil geringer (siehe etwa Jäger, 1986, S. 280) findet in Meta-Analysen (z.B. Hunter und Hunter, 1984) keine Unterstützung.

Hinsichtlich des Einflußes der Qualität von Kriterien kann beispielhaft auf einen Befund von Schuler et al. (1995) hingewiesen werden. In der Studie zeigte sich, daß die Validität der Auswahlverfahren in bezug auf die Vorgesetztenbeurteilung in Abhängigkeit von der Dauer der Zusammenarbeit zwischen den beurteilenden Vorgesetzten und dem zu beurteilenden Mitarbeiter variierte. Die Vorgesetztenbeurteilungen waren nur dann als Kriterien zu gebrauchen, wenn die Vorgesetzten ihre Mitarbeiter mindestens zwei Jahre kannten. Vergleichbare Ergebnisse wurden von Diemand, Becker und Schuler (1997) berichtet. Auf einer allgemeineren Ebene kann man einen Bezug zwischen dieser Beobachtung und den Befunden von Hossiep (1995) herstellen. Diesen Ergebnissen zufolge entfaltet sich die berufsbezogene Prognosekraft von eignungsdiagnostischen Testverfahren frühestens mittelfristig.

9.2.2 Zur Bedeutung von Validitätskoeffiziente und zur Höhe der Kriteriumsvalidität von Intelligenztests

Um die weiter unten berichteten Kriteriumsvaliditäten für Problemlöseszenarien und die im Empirie-Teil (Abschnitt 17) ermittelten Bewährungsdaten einschätzen zu können, muß man sich mit Jäger (1986, S. 281) vergegenwärtigen, daß sich die angemessene Bedeutung von Validitätskoeffizienten nicht einfach aus ihrer numerischen Höhe und deren Abstand von perfekten Vorhersagen (1,00) ergibt. *„Maß-*

Tab. 3: Verfahrensexogene Einflußfaktoren auf die Kriteriumsvalidität eignungsdiagnostischer Verfahren

1.	Zielgruppe	1a)	Art d. Berufe u. *Ausbildungen* (z.B. einfache Tätigkeiten vs. Führungspositionen; prakt. vs. theoret. Ausbildung)			
		1b)	Grad der Vorselektion			
2.	Wahl des Kriteriums	2a)	Art des Kriteriums	○ Ausbildungs- oder Berufsleistungen ○ jeweiliger zeitlicher *Abschnitt* der Ausbildungs- oder Berufsleistungen ○ spezif. Verhalten od. „Gesamtverh." ○ durchschnittl. oder Maximalleistungen ○ ein- oder mehrdimensionales Kriterium ○ singuläres Kriterium/mehrere Kriterien ○ bei multiplen Krit.: Kombinationsform		
		2b)	Qualität des Kriteriums	metrische Qualität	z.B.	Zuverlässigkeit; Diskriminationsfähigkeit
				inhaltliche Qualität (Frage der Relevanz)	subjektiv	Qualität (z.B. Dauer d. Mitarbeiterkenntnis bei Vorgesetztenbeurteilung)
					weniger subjektiv	direkt (z.B. Produktionsstückzahl, Fehlerquote, Unfälle)
						indirekt (z.B. Gehalt, Beförderung, Fehlzeiten)
3.	Beziehung zwischen Prädiktor und Kriterium	(3a)	zeitl. Abstand zwischen Prädiktor und Kriterium			
		(3b)	Vergleichbarkeit der beiden Maße (z.B. Symmetrie)			
		(3c)	Art der Beziehung zwischen Prädiktor und Kriterium (z.B. linear oder non-linear)			
		(3d)	Art der Analyse der Beziehungen zw. Prädiktor und Kriterium (z.B. der Einsatz von einfachen oder multiplen Analyseverfahren; Kreuzvalidierungen)			

gebend für die praktische Verwendbarkeit eines Tests sind zwei andere Bewertungsmaßstäbe, nämlich die Signifikanz des Abstandes seiner Validität von der eines Würfelbechers (0,00) und die Abstände von den Validitäten alternativ verfügbarer Instrumente. " Problemlöseszenarien gelten als Alternative zu Intelligenztests (siehe Kapitel 3 sowie die Ausführungen zur Konstruktvalidität weiter oben). Entsprechend müssen die Kriteriumsvaliditäten von Intelligenztests als Vergleichsmaßstab für entsprechende Werte von Problemlöseszenarien angesehen werden. Zur Orientierung soll im folgenden zunächst ein allgemeiner und anschließend ein berufsspezifischer Hinweis auf die Kriteriumsvalidität von Intelligenztests gegeben werden, wobei die berufsspezifischen Validitätsinformationen sich auf die im Empirie-Teil der Arbeit thematisierte Berufsgruppe „Polizei" beziehen. Ein solcher Hinweis auf einige in der Fachliteratur dokumentierte Erkenntnisse kann und soll den unmittelbaren empirischen Vergleich der Kriteriumsvaliditäten der beiden Verfahren nicht ersetzen.

In mehreren grundlegenden Arbeiten wurde die durchschnittliche Kriteriumsvalidität von Intelligenztests mit Hilfe metaanalytischer Validitätsgeneralisierungen über zahlreiche Studien (meist US-amerikanischer Herkunft[10]) hinweg geschätzt. Nach den dabei möglichen Artefaktkorrekturen (z.B. für Stichprobenfehler, Reliabilitätsunterschiede in den einzelnen Studien, Streuungseinschränkungen) resultierten bei Hunter und Hunter (1984) ein - über mehr als tausend Einzelstudien gemittelter - Koeffizient von r=.54 für den Ausbildungs- und r=.45 für den Berufserfolg.

Diese Werte ergaben sich als Mittelung über alle analysierten Ausbildungsgänge und Berufsbilder. Wie sieht es aber mit der Kriteriumsvalidität in der für die vorliegende Studie gewählten Berufsgruppe „Polizei" aus? (Zur Begründung der Wahl der Untersuchungsgruppe siehe Abschnitt 12.1.)

Orientiert man sich an der Fachliteratur, so herrscht bei der Polizei ein Mangel an Evaluation der Personalauswahlpraxis. Zumindest 1982 klagte der scheidende Herausgeber des „Journal of Applied Psychology", Campbell (zitiert nach Hirsh, Northrop & Schmidt, 1986, S. 400), daß innerhalb der Forschungsliteratur zur Personalauswahl gerade Validitätsstudien für den Bereich Polizei und Feuerwehr fehlen. Dabei müßten zumindest für Intelligenztests eine sehr hohe Zahl an Prädiktordaten vorliegen. Nach einer Übersicht von Ash, Slora und Britton (1990, zitiert nach Schmidt, Ones & Hunter, 1992, S. 633) verwendeten 92 % der untersuchten amerikanischen Bundes- und Stadtpolizeiorganisationen bei der Personalauswahl u.a. kognitive Fähigkeitstests.

Einen Großteil der vorliegenden Ergebnisse der (überwiegend amerikanischen)

---

[10] Die Herkunft der Studien ist bei der Interpretation der Ergebnisse in Rechnung zu stellen, da z.B. die Berufsausübung und insbesondere aber die Berufsausbildung in Amerika sich deutlich von ihrem deutschen Pendant unterscheiden.

Bewährungskontrollen aus dem Bereich „law enforcement" fassten Hirsch, Northrop und Schmidt (1986) metaanalytisch zusammen. Für unterschiedliche kognitive Testverfahren ergab sich über verschiedene Studien mit insgesamt 12897 Polizisten („police officers and detectives in public service") hinweg eine durchschnittliche beobachtete Validität von r=.34 (Standardabweichung |SD| =.10) für den Ausbildungserfolg. Demgegenüber wurde für den Berufserfolg lediglich ein entsprechender Koeffizient von r=.09 (SD=.12) ermittelt (Gesamt N über alle Studien=14991). Für beide Kriterien mußte die Generalisierbarkeit der Validitäten für einige der Testkategorien (z.B. für die Aufgaben zum gedanklichen Umgang mit Zahlen und Größenordnungen) in Frage gestellt werden.

Die geringe Vorhersagbarkeit des polizeilichen Berufserfolgs hatte sich bereits in der Übersicht bei Ghiselli (1973, S. 471f.) für eine etwas breiter angelegte Kategorie („protective occupations") gezeigt. Hirsch, Northrop und Schmidt (1986) diskutieren in diesem Zusammenhang zwei mögliche Einflußfaktoren: Erstens könnte die relativ große Unabhängigkeit sowie der große Anteil an „unbeaufsichtigtem" Außendienst dazu führen, daß die tatsächliche Berufsleistung von Polizisten durch „Außenstehende" (z.B. Vorgesetzte) nicht so gut eingeschätzt werden kann, so daß die verwendeten Kriterien nur unzureichende Indikatoren des Berufserfolgs darstellen. Zweitens könnten gerade im Polizeiberuf nicht-kognitive Variablen eine hohe Bedeutung für den Berufserfolg erlangen.

Der Arbeit von Hirsch, Northrop und Schmidt (1986) läßt sich nicht entnehmen, ob es sich bei den eingesetzten Verfahren um „herkömmliche" kognitive Tests oder um Tests handelt, die speziell für die Personalauswahl im Bereich „Polizei" entwickelt wurden und daher möglicherweise bessere Validitäten aufweisen. Eine solche spezielle Testentwicklung des „Educational Testing Service" führte zur „Multijurisdictional Police Officer Examination" („MPOE"). Dieser Test intendiert die Messung unterschiedlicher kognitiver Fähigkeiten, der Gesamttestwert kann als ein Maß der allgemeinen Intelligenz angesehen werden. Ford und Kraiger (1993) berichten über die Ergebnisse von sechs Studien zur konkurrenten und einer Studie zur prädiktiven Kriteriumsvalidität dieses Tests. Für Beurteilungen der beruflichen Leistungen variierten die korrigierten Validitätskoeffizienten der konkurrenten Studien zwischen r=.08 und r=.32, während in der prädiktive Studie (Vorhersagezeitraum: zwischen sechs und 24 Monate) für 122 Personen ein unkorrigierter Validitätskoeffizient von r=.23 erzielt wurde. Über alle Studien (Gesamt N=913) errechnen die Autoren einen gewichteten Durchschnittskoeffizienten von r=.20. Nach dem Vorbild des „MPOE" entwickelte der Calgary Police Service den „Police Applicant Test" („PAT"). Gruber (1986) beziffert die Treffsicherheit dieses Verfahrens bei der Vorhersage (zwei-Jahreszeitraum) einer Vorgesetzteneinstufung von 63 Polizisten auf einen unkorrigierten Validitätskoeffizienten von r=.19.

Von den polizeispezifischen Bewährungskontrollen aus dem deutschen Sprachraum soll kurz auf die Studie von Wolf (1990) hingewiesen werden, da es hinsichtlich des dort applizierten Intelligenztests eine kleine gemeinsame Schnittmenge mit den bei einer Subgruppe der Stichprobe der vorliegenden Studie verwendeten Testaufgaben gibt (Intelligenztest der Deutschen Gesellschaft für Personalwesen e.V. (DGP), siehe Abschnitt 12.2.3.2). Bei 249 Beamten des gehobenen Polizeivollzugsdienstes korrelierten von den elf Intelligenztestaufgaben im engeren Sinne lediglich eine zu r=.12 überzufällig mit der Ausbildungsleistung („Endergebnis Laufbahnprüfung"). Neben diesen Denkaufgaben im engeren Sinne wurden ein Merkfähigkeitstest sowie Aufgaben zur Überprüfung des Arbeitsverhaltens (Konzentration und Sorgfalt) und der Kenntnisse in verschiedenen Wissensdomänen eingesetzt. Ein von Psychologen abgegebenes „klinisches Urteil" über die Gesamtleistungen in allen Tests stand in keinem systematischen Zusammenhang zum Ausbildungserfolg. Mit dem von Wolf analysierten DGP-Test wurde in den eignungsdiagnostischen Untersuchungen der DGP der früher eingesetzte WILDE-Test (Jäger und Althoff, 1994) abgelöst. Für den WILDE-Test (Gesamtwert) konnten in einer Studie von Althoff (1977) für eine Gruppe von 76 Kriminalkommissaren konkurrente und retrograde Kriteriumsvaliditäten im Streubereich von r=.27 bis r=.46 nachgewiesen werden. Ebenfalls an Ausbildungsnoten validierte Greif (1972) den WILDE-Test. Die Ausbildungsnoten von 314 Bereitschaftspolizisten korrelierten zwischen r=.19 bis r=.43 mit den WILDE-Gesamtwert.

9.2.3 Empirische Studien zur Kriteriumsvalidität von Problemlöseszenarien

Schließt man sich der Ansicht von Hossiep (1995, S. 73) an, daß der Stand veröffentlichter Untersuchungsergebnisse zur Validität eignungsdiagnostischer Verfahren im Sinne von Kontrollen beruflicher Bewährung in der Bundesrepublik Deutschland ganz allgemein (über verschiedene Auswahlverfahren hinweg) als *„katastrophal"* bezeichnet werden kann[11] (siehe auch Althoff, 1984, S. 144), so fehlt einem für die treffliche Kennzeichnung der weitaus ungünstigeren Situation der Kriteriumsvalidierung von computergestützten Problemlöseszenarien das Vokabular.

Möglicherweise aus dem Mangel an Studien zur Kriteriumsvalidität heraus, wer-

[11] Nach Hossiep (1995, S. 73f.) wurden in den sechs einschlägigen deutschsprachigen Fachzeitschriften innerhalb eines Zeitraums von 44 Jahren lediglich 82 Arbeiten zum Thema empirische Validitätskontrollen berufseignungsdiagnostischer Verfahren veröffentlicht. Allein 13 dieser 82 Arbeiten stammen von ein- und demselben Autor, der ausschließlich Befunde aus dem Eschweiler Bergwerks-Verein berichtet.

den an verschiedener Stelle auch Untersuchungen zu anderen Fragestellungen der Thematik von *Kriteriums*validierungen zugeschlagen. So etwa, wenn die in Kapitel 4.3 dargestellten Experten-Novizen Vergleiche als *„empirische Belege für die externe Validität"* (Putz-Osterloh, 1991a f, S. 101) interpretiert werden. Diese Einordnung der Experten-Novizen Studien ist in vielerlei Hinsicht suspekt. Neben den im Abschnitt 4.3 beschriebenen spezifischen experimentellen Schwächen einzelner Untersuchungen weist der Experten-Novizen Vergleich einige grundsätzliche Schwachpunkte auf. Der mit der Wahl der Extremgruppen erhofften indirekten Induktion von Expertenwissen fehlt (1) jegliche Manipulationskontrolle. Voraussetzung des Expertentums wäre außerdem die „ökologische Validität" der Szenarien, die aber, wie oben (Kapitel 4) bereits ausgeführt, (2) völlig ungeprüft bleibt. (3) Möglicherweise leistungsrelevante Aspekte wie Alter, Motivation usw. blieben in den Untersuchungen unkontrolliert. Abgesehen von der (4) grundsätzlichen methodischen Problematik von Extremgruppenvergleichen, gilt es vor allem (5) zu bedenken, daß es trotz dieser Untersuchungen nach wie vor keinen Hinweis darauf gibt, *„daß die untersuchten Verfahren nicht nur zwei bereits als extrem disparat bekannte Gruppen differenzieren, sondern die für eignungsdiagnostische Anwendungen typische Forderung erfüllen,* ***innerhalb einer homogenen Bewerbergruppe*** *zu differenzieren und reale Problemleistungsdifferenzen prognostizieren zu können. Hieraus ergibt sich die Notwendigkeit zur Durchführung kriteriumsorientierter Validierungen mit Außenkriterien für Problemlösekompetenz an Gruppen realistischer Homogenität."* (U. Funke, 1991, S. 116)

Dem Experten-Novizen Vergleich liegt die „falsche" Schlußrichtung zugrunde: man schließt aufgrund der Ereignisse (z.B.: Gruppenzugehörigkeit: Expertentum) auf die diagnostische Information während man beim aktuarischen Ansatz in der Diagnostik (siehe Wiggins, 1973, S. 81 f.) aufgrund der diagnostischen Information auf die Ereignisse schließen will. Die Ergebnisse zu Experten-Novizen Befunden werden daher nicht im Rahmen der Befunde zur Kriteriumsvalidität berichtet.

Auch Interkorrelationen von Steuerungsleistungen als eignungsdiagnostische Prädiktoren mit anderen eignungsdiagnostischen Prädiktoren - wie z.B. dem Eignungsurteil aus einem Assessment-Center - können nicht als Beleg für den Zusammenhang zwischen Prädiktor und Kriterien der tatsächlichen beruflichen oder schulischen Bewährung interpretiert werden. Aus der Korrelation zweier Verfahren, ergibt sich - sofern die Korrelation nicht außergewöhnlich hoch ausfällt - auch dann kein zwingender Hinweis auf die Validität des einen Verfahrens, wenn sich für das andere Verfahren die Validität nachweisen läßt. Entsprechende Berichte über die Assoziation oder die Dissoziation von Problemlösegütemaßen und Assessment-Center Beurteilungen - wie sie weiter oben, Abschnitt 9.1.5.2, referiert wurden - werden daher in der vorliegenden Arbeit im Kontext der Kriteriumsvalidierung nicht

berücksichtigt. Ebenfalls unberücksichtigt bleiben zwangsläufig Studien zur Kriteriumsvalidierung von computergestützten Problemlöseszenarien, deren Existenz zwar behauptet, die aber nicht den wissenschaftlichen Standards entsprechend dokumentiert sind. So deuten etwa Hartung und Schneider (1995, S. 233) die Existenz einer Studie zur (vermutlich: konkurrenten) Kriteriumsvalidität des Szenarios „Utopia" an 54 AC-Teilnehmern eines Elektrokonzerns an, bei dem sich eine bedeutsame Korrelation zwischen den Verhaltensmaßen und einem „externen Kriterium Vorgesetztenurteil" ergeben habe. Angaben, die zur Beurteilung der Studie unentbehrlich sind (z.B. über den genauen Untersuchungsablauf, über das externe Kriterium usw.) werden aber nicht getroffen, auch ein Hinweis auf weiterführende Literatur findet sich nicht.

9.2.3.1 Retrograder Validierungsansatz

Eine retrograde Kriteriumsvalidierung des Szenarios „Feuer" wurde von Schoppek (1991) mit 22 Studenten der Wirtschaftswissenschaften durchgeführt. Die Studenten hatten kurz vor der Untersuchung ihre Diplom-Vorprüfung absolviert. Sowohl die Steuerungsleistung beim System „Feuer" als auch die Ergebnisse des „IST-70" Tests korrelierten mit den Prüfungsnoten ($r=.53$ für das computergestützte Problemlöseszenario und $r=.52$ für die Intelligenztestaufgaben). Multiple Auswertungskonzepte erbrachten keine bedeutsame Steigerung der Vorhersageleistung, so daß die Steuerungsleistung und die Intelligenzleistung nach Ansicht des Autors eher ähnliche Varianzanteile der Prüfungsleistung aufklärten.

Putz-Osterloh und Köster (1988) korrelierten - ebenfalls einem retrograden Ansatz der Kriteriumsvalidierung folgend - die Steuerungs- und Strategieleistungen beim System „Schneiderwerkstatt" mit den Schulnoten der 100 Untersuchungsteilnehmer. Dabei zeigte sich keinerlei Zusammenhang.

9.2.3.2 Konkurrenter Validierungsansatz

Zur konkurrenten Kriteriumsvalidierung tragen die Studien von Obermann (1991) und U. Funke (Schuler et al., 1995; U. Funke, 1993, 1995a und 1995b) bei.

Obermann (1991) gibt in der Handanweisung zum Szenario „Airport" knappe Hinweise auf eine Studie zur konkurrenten Kriteriumsvalidierung mit 23 Personen (Gruppenleiter Finanzdienstleistungen). In einer „Eigen-" und „Fremdeinschätzung" wurden gerade einmal drei Items bearbeitet. Zunächst wurde mit einem Item im anforderungsanalytischen Sinne nach der „Häufigkeit der Erfahrung mit komplexen Problemen" gefragt. Dieses Item taugt nicht zur Kriteriumsvalidierung. Mit dem zweiten Item sollte die Fähigkeit zum komplexen Problemlösen ipsativ mit anderen

Fähigkeiten (z.B. Sozial- und Kontaktverhalten, Streßresistenz) verglichen werden. Auch dieses Item kann nur bedingt als Kriterienmaß verwendet werden, so daß für die eigentliche Kriteriumsvalidierung nur noch ein single-item zur Ausprägung des Problemlöseverhaltens im Vergleich zu anderen Mitarbeitern übrig blieb. Die mit Hilfe dieses Items vorgenommene Fremdeinschätzung korrelierte zu .55 mit der Gesamtleistung der Systemsteuerung.

U. Funke setzte das computergestützte Problemlöseszenario „DISKo" im Rahmen einer Studie zur Personalauswahl in Forschung und Entwicklung (Schuler et al., 1995) ein. In dieser Untersuchung wurde u.a. die konkurrente Validität verschiedener Personalauswahlverfahren anhand eines arbeitsanalytisch fundierten Leistungsbeurteilungsverfahrens ermittelt. Zur Validierung des Problemlöseszenarios wurde auf Basis der Vorgesetztenurteile neben der „allgemeinen Leistung" (arithmetisches Mittel der 53 Einzelskalen des Beurteilungsverfahrens) auch ein aus vier Einzelskalen gemitteltes spezifisches Kriterium „fachliches Problemlösen" gebildet. Dabei handelte es sich um die Einzeldimensionen „Problemstrukturierung", „Lösung (Ergebnisevaluation und -interpretation)", „Untersuchen und püfen" sowie „Produkte (Neu- oder Weiterentwicklungen, Patente etc.)". Zur Bestimmung der Reliabilität des Kriteriums wurde ca. 30% (33 Personen) der Vorgesetzten nach ca. zwei Monaten eine erneute Leistungsbeurteilung abverlangt. Der so bestimmte Reliabilitätskoeffizient für das Gesamtkriterium betrug r=.93. Für die vier Einzelskalen, die zu dem Kriterium „fachliches Problemlösen" zusammengefaßt wurden, betrug der entsprechende Reliabilitätskoeffizient im Mittel r=.54. Bei der Gesamtgruppe von 117 Personen ergab sich eine nicht bedeutsam vom Zufall abweichende Korrelation zwischen dem Gesamtergebnis des Szenarios (Aggregat aus Steuerungsleistung und Strategieindex) einerseits mit dem spezifischen Kriterium andererseits, in Höhe von r=.12 bzw. mit dem allgemeinen Kriterium in Höhe von r=.09 (Schuler et. al, 1995, S. 113). Allerdings arbeiteten die Teilnehmer der Studie in sehr unterschiedlichen Unternehmen und Abteilungen aus dem Bereich Forschung und Entwicklung, so daß es sich hinsichtlich der Arbeitsplatzanforderungen um eine heterogene Gruppe handelte. Zur Demonstration der Gruppenspezifität der berichteten Validitäten wurden zwei Gruppen gebildet, die im Berufsalltag im unterschiedlichen Maße kognitiv gefordert waren. Für die Gruppe mit hohen kognitiven Anforderungen wurde mit einer Korrelation in Höhe von r=.43 für das Gesamtergebnis der Szenarienbearbeitung eine deutliche höhere konkurrente Kriteriumsvalidität (spezifisches Kriterium) erzielt, wobei nur diejenigen Personen in die Analyse aufgenommen wurden, deren Vorgesetzte aufgrund längerer Zusammenarbeit die Problemlöseleistung zuverlässig beurteilen konnten. (Eine Angabe zur Gruppengröße wird nur für eine Gesamtübersicht der spezifischen Validität aller verwandten Personalauswahlverfahren gegeben und mit *„n=37-65"* (ebd. S. 166) beziffert.) Auf

die (z.B. regressionsanalytische) Kombination des computergestützten Problemlöseszenarios mit den übrigen Personalauswahlverfahren und auf die Bestimmung der Effekte dieser Kombination wurde aus Gründen der hohen Anforderungsspezifität des Szenarios für die Gesamtstichprobe verzichtet (ebd., S. 171).

Über diese Studie berichtet der Autor auch an anderer Stelle (U. Funke, 1993, 1995a und 1995b), wobei die dort berichteten Ergebnisse aufgrund der Verwendung eines geringfügig anderen (wiederum: spezifischen) Leistungskriteriums von den oben dargestellten Werten abweichen. Diesmal wurden die folgenden Einzeldimensionen zu einer Kriteriendimension „Problemlösen" zusammengefaßt: „Produzierte Ideen", „wissenschaftliche Kenntnisse", „Innovation", „Untersuchen und Prüfen", „technischer Service", „Problembearbeitung" und „Problemstrukturierung". Außerdem wurde eine etwas anders zusammgesetzte Teilgruppe der Vorgesetzten mit einem zuverlässigen Urteil (mindestens zwei Jahre Mitarbeiterkenntnis) für die Auswertung berücksichtigt. Diesen Berichten zufolge korrelierte die Steuerungsleistung (Aggregat aus Gesamtkapital und Gesamtkapitalanstieg) zu r=.36 (.39), ein (aus fünf Verhaltensmaßen gemittelter) Verhaltensindex zu r=.39 (.42) und die aus diesen beiden Werten gemittelte Gesamtsumme zu r=.44 (.47) mit dem Kriterium. Die Werte in Klammern geben jeweils die für die Unreliabilität der Kriterien attenuationskorrigierten entsprechenden Validitäten an.

9.2.3.3 Prädiktiver Validierungsansatz

Zum Zeitpunkt der Erstellung dieser Arbeit lag nur eine Studie zur prädiktiven Validierung von computergestützten Problemlöseszenarien vor. Hasselmann (1993) untersuchte 21 Mitarbeiter einer Großbank. Dabei handelte es sich um eine vorselektierte Gruppe, die aufgrund ihrer besonderen Leistungen (1.) von ihren Vorgesetzten für das Führungsnachwuchsprogramm des Unternehmens vorgeschlagen worden waren und (2.) ein Assessment-Center erfolgreich bestanden hatten. Im Rahmen dieses Assessment-Centers steuerten die Führungskräfte das System „Textilfabrik". Zur prädiktiven Validierung wurden nach einer Zeit von 19 bis 29 Monaten (Durchschnitt 24 Monate) verschiedene Kriterien der beruflichen Bewährung erhoben. Der Autor gliedert die erhobenen Kriterien in die Kategorien „harte Daten" (z.B. aktuelles Gehalt, aktuelle Position) und „weiche Daten". Zur Gruppe der „weichen" Daten zählt Hasselmann u.a. Potentialeinschätzungen des für die jeweilige Führungskraft zuständigen Betreuers im Personalbereich. Diese Personalbetreuer nahmen auch eine Einschätzung der aktuellen Leistung und des aktuellen Nutzens des Mitarbeiters vor und nannten einen Rangplatz, den der jeweilige Mitarbeiter innerhalb der Gruppe der Führungsnachwuchskräfte ihrer Ansicht nach aktuell belegt. Die Instrumente zur Leistungs- und Nutzensbeurteilung werden in der Studie

von Hasselmann nicht weiter vorgestellt, deren psychometrische Qualitäten werden nicht beschrieben. Problematisch erscheinen die erhobenen Daten zur voraussichtlichen Position des Mitarbeiters in der näheren (zwei Jahre nach der Kriteriumserhebung) und ferneren (vier Jahre nach der Kriteriumserhebung) Zukunft, da hier Prognosen als Kriterium zur Bestimmung der prognostischen Aussagekraft von Steuerungsleistungen verwendet werden.

Mit Korrelationen im Streubereich von r=.42 bis r=.61 ergaben sich die deutlichsten Zusammenhänge zwischen den unterschiedlichen Gütemaßen der Systemsteuerung als Prädiktoren und den Kriterien der formellen beruflichen Bewährung (aktuelle Position und aktuelles Gehalt) sowie den Kriterien der „voraussichtlichen Position in zwei oder fünf Jahren" bzw. den Einschätzungen des allgemeinen und führungsspezifschen Potentials. Nominell etwas geringere Korrelationen in Höhe von r=-.32 bis r=-.39 ergaben sich für das Kriterium „Rangplatz". Hervorzuheben ist, daß die „eigentlich" naheliegenden Kriterien, nämlich die Einschätzung der aktuellen Leistung und des aktuellen Nutzens des Mitarbeiters nicht durch die Steuerungsleistung vorhergesagt werden konnten (r=.09 bis r=.16). Auch die Einschätzung des Potentials zur Übernahme von Positionen mit hoher Fachverantwortung sowie die Entwicklungsfortschritte, die aus Sicht der Beurteiler von den Führungsnachwuchskräften seit der Prädiktorenerhebung erzielt wurden, korrelierten nicht statistisch bedeutsam mit der Leistung in der „Textilfabrik". Die für die „Textilfabrik" ermittelten prädiktiven Validitäten für ausgewählte Kriterien liegen auf einem Niveau mit den ebenfalls berichteten prädiktiven Validitäten der Verhaltensbeurteilungen aus dem Assessment Center. Auch für eine Postkorbübung sowie für die eingesetzten Untertests zum logischen Denken aus dem „IST-70" konnte der Nachweis der prädiktiven Validität erbracht werden, die Analogie-Aufgaben des „IST-70" erzielten signifikante Korrelationen in Höhe von r=.38 bis r=.78 mit sechs der erhobenen Kriterienmaße. Lediglich der Konzentrationstest „d2" und der sprachliche „IST-70" Subtest „Satzergänzung" erwiesen sich als weitgehend unkorreliert mit den im Schnitt zwei Jahre später erhobenen Kriterien.

In einer späteren Darstellung der Untersuchung (Hasselmann, 1995) berichtet der Autor, daß nicht nur die „Textilfabrik", sondern auch ein zweites Szenario, nämlich der „Brennstoffvertrieb" im Rahmen der Studie zur Kriteriumsvalidierung bearbeitet wurde - die mit diesem Szenario erzielten Vorhersagen fanden in den früheren Darstellungen der Studie (Hasselmann, 1991, 1993) keine Erwähnung. Für den „Brennstoffvertrieb" liegen dem 1995 publizierten Hinweis zufolge Ergebnisse von 25 weiteren (anderen) Personen vor. Hinsichtlich der prädiktiven Validierung der über ein Trendmaß („Trendfu") bestimmten Problemlöseleistung für den „Brennstoffvertrieb" ergab sich ein anderes Muster als für das entsprechende Trendmaß für die „Textilfabrik". Fanden sich für die „Textilfabrik" Zusammenhänge zwischen dem

Steuerungserfolg und Aspekten des Führungspotentials, so korrelierte der Steuerungserfolg beim „Brennstoffvertrieb" mit r=.52 am höchsten mit der Einschätzung des Potentials zur Übernahme von Positionen mit ausgeprägter Fachverantwortung. Gerade dieses Kriterium hatte sich als mehr oder minder unabhängig (r=.15) von der Steuerungsleistung in der „Textilfabrik" erwiesen. Auch die Einschätzung des aktuellen Nutzens des Mitarbeiters konnte zu r=.37 durch die Steuerungsleistung im „Brennstoffvertrieb" - nicht aber in der „Textilfabrik" (r=.09) - vorhergesagt werden. *„Inwieweit diese unterschiedlichen Zusammenhänge inhaltlich begründet sind, oder es sich um Artefakte handelt (...)"* bleibt nach Ansicht von Hasselmann (1995, S. 250) *„(...) noch aufzuklären."* Diese unterschiedlichen Befundmuster wecken erneut Zweifel an der Generalität der gemessenen „Problemlösefähigkeit" (siehe oben, Abschnitt 9.1.1).

Einige spannende Fragen wurden von Hasselmann nicht untersucht. So wäre es zunächst interessant gewesen zu prüfen, inwieweit die Steuerungsleistung gegenüber den übrigen Prädiktoren der Assessment-Center Urteile sowie vor allem gegenüber den Intelligenzaufgaben einen *inkrementellen* Beitrag zur prädiktiven Validierung leistet. Durch eine solche Prüfung wäre es möglich gewesen, die Nützlichkeit (Lienert, 1967, S. 19) des computergestützten Problemlöseszenarios für die Management-Diagnostik zu beurteilen. Zur Klärung der Nützlichkeitsfrage - im Sinne eines empirischen Vergleichs der Kriteriumsvalidität der Szenarien mit der Kriteriumsvalidität bereits vorhandenen Verfahren mit überlappendem Geltungsanspruch - trägt die Hasselmann Studie daher nichts bei.

Auf der Seite der Kriterien hätte sich eine Aggregierung angeboten, um die Reliabilität der Urteile zu erhöhen.

9.2.4 Zusammenfassung zur Kriteriumsvalidität

Die vorliegenden Studien zur Kriteriumsvalidierung sind quantitativ und qualitativ unzureichend. Lediglich in der Studie von U. Funke wurde die Qualität der Kriterien diskutiert und geprüft, und nur in dieser Studie wurde mit einer zahlenmäßig umfassenderen Stichprobe gearbeitet. Die Arbeit von U. Funke erlaubt aufgrund der gleichzeitigen Erhebung von Prädiktor und Kriterium aber keine Aussage zur prädiktiven Validität des Szenarios. In keiner Studie wurde die Generalisierbarkeit der Kriteriumsvalidität über verschiedene Szenarien hinweg geprüft, unbeantwortet blieb auch die zentrale Frage, wieviel *inkrementelle* Validität computergestützte Problemlöseszenarien gegenüber vorhandenen diagnostischen Verfahren liefern.

Diese ohnehin ungünstige Situation der Kriteriumsvalidierung von Problemlöseszenarien gewinnt an Dramatik, wenn man die hohe Systemspezifität der Steue-

rungsleistungen in Rechnung stellt (Abschnitt 9.1.1). Für den praktischen Einsatz in der Diagnostik bedeutet dies, daß die hier berichteten Ergebnisse zur Kriteriumsvalidität von computergestützten Problemlöseszenarien in einem weitaus geringerem Maße auf andere Szenarien verallgemeinert werden können als dies z.B. hinsichtlich der Befunde für Intelligenztests der Fall ist.

9.3 Generaldiskussion zur Validität von computergestützten Problemlöseszenarien, Schlußfolgerungen und Ausblick

Das von computergestützten Problemlöseszenarien intendierte stabile Fähigkeitskonstrukt wird von den einzelnen Steuerungsleistungen in spezifischen Situationen nur äußerst unzureichend reflektiert. Die Befunde zur (ungenügenden) Generalität zeigen, daß mit den einzelnen Szenarien nur im geringen Ausmaß aufgabenübergreifende Fähigkeiten erfaßt werden, der weitaus größte Anteil der Leistungsvarianz muß demnach auf situative, aufgabenspezifische sowie auf Fehlervarianz zurückgeführt werden. Die europäische Problemlöseforschung führt - wie Sternberg (1995, S. 303) beklagt - eher zu aufgabenbezogenen als zu generalisierbaren theoretischen Annahmen über das Problemlösen.

Berücksichtigt man, daß der Anteil systematischer Varianz an Steuerungsleistungen offensichtlich klein ist, kommt den an und für sich nur mittleren Effekten von Intelligenz und Wissen auf das Problemlösen eine herausragende Bedeutung zu. In der Berliner und in der Mannheimer Untersuchung zeigte sich, daß sich die gesamte stabile Varianz der Steuerungsleistungen bei einem Szenario (Berliner Untersuchung, siehe Süß, 1996) bzw. bei mehreren Szenarien (Mannheimer Untersuchung, siehe Wittmann et al., 1996) durch Messungen zu bereits etablierte Fähigkeitskonstrukten aufklären ließ, so daß die Annahme einer gesonderten Problemlösefähigkeit beim gegenwärtigen Stand der Forschung unnötig ist.

Die mit computergestützten Problemlöseszenarien gewonnenen Meßwerte besitzen der Literaturübersicht zufolge keine spezifische Bedeutung außerhalb der Bedeutung von Intelligenz und Wissen. Tritt das Problemlöseszenario als diagnostisches Instrument neben Intelligenztests und Wissensfragen, reduziert sich der Gewinn an Informationen über aufgabenübergreifende Fähigkeiten um das Maß der Überlappung. Im Vergleich zu einem isolierten Einsatz von entweder Testaufgaben zur (fluiden) Intelligenz oder Wissensfragen erfordern Problemlöseszenarien aber *zugleich* intellektuelle und Wissensleistungen. Problemlöseaufgaben bieten daher theoretisch die Gelegenheit, zu einem kleinen Anteil simultan Intelligenz *und* Wissen zu diagnostizieren, wobei allerdings diese beiden sinnvoll differenzierbaren

Konstrukte in den Problemlösegütemaßen zu einem nicht mehr trennbaren Konglomerat verschmolzen werden. Durch den Befund der hohen Aufgabenspezifität der Problemlöseleistungen wird das Ansinnen, bei Personalentscheidungen die Ergebnisse der Steuerung von Problemlöseszenarien zu berücksichtigen, massiv in Frage gestellt. Die mit Problemlöseszenarien erstellte Diagnose variiert erheblich in Abhängigkeit von dem jeweils eingesetzten Szenario. Je nachdem, welches Szenario gerade Verwendung findet, würde in einer derart diagnostisch gestützten Personalauswahl mal dieser und mal jener Kandidat eine positive Entscheidung erhalten. Dies wäre nur dann sinnvoll, wenn der Unterschied in den spezifischen Anforderungen der Szenarien mit den unterschiedlichen Anforderungen verschiedener Berufsbilder korrespondieren würde. Tatsächlich hat man sich bislang weder um eine Analyse der Anforderungen einzelner Szenarien (siehe oben, Abschnitt 2.3.2.3) noch um eine Analyse der Korrespondenz zwischen Szenarien und Berufsanforderungen (siehe Abschnitt 3.1 und 4.2) bemüht, so daß der Befund der Aufgabenspezifität der Problemlöseleistungen in keinem positiven Licht erscheinen kann.

Die hohe Systemspezifität verschiedener Problemlöseleistungen vereitelt auch den Ansatz der Validitätsgeneralisierung. Somit bleiben die ohnehin beschränkten Aussagen über die Kriteriumsvalidität, die sich aus den seltenen (mit nur einer Ausnahme konkurrent oder retrograd angelegten) Studien zu dieser Frage ableiten lassen, in ihrer Gültigkeit auf das jeweils eingesetzte Szenario begrenzt.

Die Ausführungen zur Konstruktvalidierung legen es nahe, die Kriteriumsvalidität von Problemlöseszenarien unmittelbar mit der Kriteriumsvalidität von Intelligenz- und Wissenstests zu vergleichen und somit die Nützlichkeit der Problemlöseszenarien im Sinne Lienerts (1967, S. 19) überhaupt erst einer Prüfung zugänglich werden zu lassen. Validitätsvergleiche zwischen Verfahren lassen sich nur bedingt anhand von Studien vornehmen, in denen entweder das eine oder das andere Verfahren eingesetzt wurde. Validitätsvergleiche sind nach Ansicht von Schuler (1996, S. 168) vielmehr erst dann sinnvoll, wenn die Verfahren auf gleiche Weise an der gleichen Zielgruppe durchgeführt wurden und deren Effekt am gleichen Kriterium gemessen werden. Wünschenswert sind im eignungsdiagnostischen Kontext vor allem solche Verfahren, die gegenüber vorhandenen Verfahren eine inkrementelle Validität liefern. Diese besondere Bedeutung der *„Steigerung der Validität gegenüber den bisher eingesetzten Verfahren"* (Kreuzig, 1995b, S. 399) ist auch den Anbietern eignungsdiagnostisch orientierter computergestützter Problemlöseszenarien bewußt. Zu dieser wichtigen Frage konnte der Anwender aber bislang lediglich folgendes erfahren: *„MANAGE! braucht den Vergleich mit anderen Verfahren nicht zu scheuen"* (Kreuzig, ebd.). Mag Kreuzig die Scheu vor dem Vergleich mit diesem Zitat auch bereits verbal abgelegt haben, empirisch überwunden ist sie bislang nicht.

Ausgehend von diesen Überlegungen lag es nahe, eine Studie zur prädiktiven Kriteriumsvalidierung anzustellen, in der – zur Klärung der Frage der Systemspezifität – mehr als nur ein Problemlöseszenario eingesetzt wird, und in der gleichzeitig etablierte diagnostische Verfahren zu verwandten oder gar identischen Fähigkeiten, nämlich Intelligenz und Wissen, zum Einsatz kommen. Über diese Studie wird im empirischen Teil der vorliegenden Arbeit berichtet. Vorab sollen im folgenden Kapitel aber noch einige Aspekte der Evaluation diagnostischer Verfahren thematisiert werden, die bislang vernachlässigt wurden.

10. „Fairneß", „Verfälschbarkeit", „Normierung" und „Ökonomie" als besondere Gesichtspunkte beim Einsatz computergestützter Problemlöseszenarien als psychodiagnostische Verfahren

Das Testkuratorium (1986) hat eine Reihe von Gesichtspunkten benannt, die bei der Evaluation psychodiagnostischer Verfahren berücksichtigt werden sollten. Auf die meisten Beurteilungsaspekte wurde im Rahmen der vorliegenden Auseinandersetzung mit einer möglichen diagnostischen Verwendung computergestützter Problemlöseszenarien bereits eingegangen. Einige bislang vernachlässigte Aspekte sollen im folgenden Kapitel noch hervorgehoben werden, nämlich die Aspekte der „Fairneß" (insbesondere vor dem Hintergrund der (geschlechts- und möglicherweise auch altersspezifisch) unterschiedlich ausgeprägten Computererfahrung), „Verfälschbarkeit", „Normierung" und „Ökonomie".

10.1 Zur Fairneß einer Diagnostik mit computergestützten Problemlöseszenarien

10.1.1 Zum Begriff der „Fairneß" und der Bedeutung von gruppenspezifischen Leistungsunterschieden für die Fairneß der diagnostischen Entscheidung

Der Aspekt der Fairneß betrifft nach der Definition des Testkuratoriums (1986, S. 360) das Ausmaß einer eventuell bestehenden systematischen Diskriminierung bestimmter Testpersonen, z.B. aufgrund ihrer ethischen, soziokulturellen oder geschlechtsspezifischen Gruppenzugehörigkeit, bei der Abschätzung von Kriteriumswerten. Da zur Zeit keine Studien zur Kriteriumsvalidität von Problemlöseszenarien vorliegen, die eine gruppenspezifische Auswertung erlauben würden, läßt sich über die Frage der Fairneß dieser Instrumente zunächst nicht abschließend befinden. Auch wenn das Fairneß-Problem, wie Wottawa und Amelang (1980, S.202) zu Recht betonen, darin liegt *„ob bestimmte Tests zu subgruppenspezifischen Fehleinschätzungen der Kriteriumswerte führen, und nicht darin, ob es Unterschiede der*

Testmittelwerte an sich gibt" (siehe auch Vernon, 1979, S.318), wirken sich gruppenspezifische Leistungsmittelwerte aber gleichwohl auf bestimmte Fairneßaspekte aus und sind somit als Indikatoren *potentieller* Fairneßprobleme zu werten. Aufgrund fehlerbehafteter Messungen auf seiten des Kriteriums und des Prädiktors gibt es grundsätzlich keine perfekten diagnostischen Prognosen, d.h. es wird immer auch Fehlentscheidungen geben. Eine Fehlentscheidung an sich ist unter keinem Gesichtspunkt „fair", unter Fairneßgesichtspunkten ist aber zu diskutieren, welche Gruppenmitglieder welcher Fehler mit welcher Häufigkeit trifft. An dieser Stelle der Fairneß-Debatte können nun gruppenspezifische Mittelwertsunterschiede Bedeutung erlangen. Das im Rahmen der Fairneß-Debatte zumeist angewandte *„(Regressions)Modell der fairen Vorhersage"* nach Cleary[12] richtet die „Fairneß" an der Gruppe der Selegierten aus und übergeht, daß dabei zugleich eine Unfairneß hinsichtlich der Abgelehnten bestehen kann. Selbst bei identischen Kriteriums-Prädiktor-Verhältnissen können sich nämlich testleistungsheterogene Gruppen unterschiedlich auf die beiden Fehler*typen* der Vorhersage verteilen (siehe Wigdor & Sackett, 1993). Der dabei wirksam werdenden Effekt von Gruppen-Leistungsunterschieden auf die diagnostischen Entscheidungen kann am Beispiel einer von Kersting (1995, S. 37 f.) im Rahmen der Untersuchung von innerdeutschen Leistungsunterschiede in Berufseignungstests angestellten Modellrechnung veranschaulicht werden. Während der testleistungsstärkeren Westgruppe relativ häufiger der „angenehmere" Fehler einer Überschätzung zufällt (Zulassung trotz mangelnder Eignung), wird geeigneten Bewerbern aus den neuen Bundesländern aufgrund ihrer relativen Testdefizite vergleichsweise häufiger zu Unrecht die Zielposition verwehrt (falsch Negative). Diese „Verschiebung" und das damit einhergehende Fairneßproblem ergibt sich, obwohl der Test - den Annahmen der Modellrechnung zufolge - für beide Gruppen eine identische Kriteriumsvalidität aufweist. Gruppenunterschiede in diagnostischen Verfahren konstituieren unter spezifischen Fairneßgesichtspunkten also selbst bei *identischen* Steigungen der Regressionslinien zwischen Test und Kriterium für beide Populationen ein Problem bei der Entscheidungsfindung. Neben der Frage nach der (singulären oder differentiellen) prognostischen Validität von Problemlöseszenarien, die zur Zeit nicht geprüft werden kann, sollte daher untersucht werden, ob gruppenspezifische Leistungsunterschiede bei der Bearbeitung von Problemlöseszenarien auftreten. Erwähnt wurden bereits Wissenseffekte auf die Problemlöseleistung, die dann ein Fairneßproblem begründen können, wenn das für die Problemlösung hilfreiche Wissen selbst nicht eignungsrelevant ist und die getesteten Kandidaten hin-

---

[12] Ein Auswahlverfahren ist dem *„(Regressions)Modell der fairen Vorhersage"* zufolge *„dann fair, wenn das dafür verwendete Vorhersageinstrument (Test) für das Kriterium in keiner der beiden zu vergleichenden Gruppen eine systematische Über- und Unterschätzung ihrer Kriteriumswerte erbringt"* (Bartussek, 1982, S. 3).

sichtlich dieses Wissens eine leistungsheterogene Gruppe darstellen. Im folgenden soll ein anderer Effekt, nämlich der Effekt unterschiedlicher Computer-Erfahrung auf die Problemlöseleistungen thematisiert werden.

10.1.2 Zum Einfluß der Computererfahrung und der Einstellung zur Arbeit mit Computern auf die Problemlöseleistung

Sofern die Szenarienbearbeitung nicht über einen Versuchsleiter vermittelt wird - was mit unkontrollierbaren Versuchsleitereffekten verbunden ist - ist mit Problemlöseszenarien die Verwendung des Computers zur Test*vorgabe* verbunden. Durch die damit notwendige Interaktion mit dem PC ergibt sich für die Diagnostikanden zusätzlich zur eigentlichen Problemlösung eine weitere Aufgabenstellung: der Umgang mit dem Computer. Kleinmann und Strauß (1995, S.127) grenzen diese Aufgabe der Steuerung der Softwaresysteme als *Interaktionsproblem* gegenüber der eigentlichen Aufgabe als *Sachproblem* ab. Die Autoren akzentuieren, daß sich das diagnostische Ziel, etwas über die Kompetenz zur Bearbeitung des Sachproblems zu erfahren, nur dann realisieren läßt, wenn man den Einfluß der Vorerfahrung mit Computern und dem Einfluß der eventuell mangelhaften Softwareergonomie auf das Zustandekommen des Problemlöseergebnisses kennt (ebd, S. 128).

Die Frage, ob ein bestimmter Varianzanteil der Problemlöseleistung durch die - *konstruktirrelevante* und diagnostisch nicht in jeder Situation relevante - Computererfahrung bedingt wird, wurde bislang vergleichsweise selten untersucht. U. Funke (1992a, S. III-5) berichtet über die geringe Anfälligkeit der Problemlöseleistung im Szenario „DISKo" gegenüber verschiedenen Formen der Vorerfahrungen mit Computern. Diese Vorerfahrungen blieben bei einer Stichprobe von 123 Personen allesamt ohne Einfluß auf das „DISKo-Gesamtergebnis". Hinsichtlich der Verhaltensindizes galt, daß sich *„die verschiedenen Vorerfahrungen im Umgang mit Computern"* nur *„vergleichsweise wenig auf die Strategien"* auswirken (1991, S. 121). Laut Kreuzig (1995b, S. 397) gibt es auch für das Szenario „Manage" keinen Hinweis darauf, daß PC-Kenntnisse Einfluß auf das Ergebnis nehmen. Hasselmann (1993, S. 203) konnte ebenso keinen statistisch bedeutsamen Zusammenhang zwischen der Steuerungsleistung in der „Textilfabrik" und der Computererfahrung ermitteln. Schließlich ergab sich auch für die Teilgruppe der Studenten aus der Studie von Putz-Osterloh und Haupts (1990, S. 142) keine Wirkverbindung zwischen der Bearbeitung zweier Szenarien und dem Computer-Vorwissen.

Demgegenüber beobachtete Obermann (1995, S. 405), daß ältere Kandidaten ohne PC-Erfahrung Schwierigkeiten mit dem Szenario „Airport" haben. In der Berliner Untersuchung (siehe Süß, 1996, S. 182 f.) zeigten sich mittlere bis starke Zu-

sammenhänge der Computererfahrung sowohl mit den Steuerungsleistungen in der „Schneiderwerkstatt" als auch mit den Leistungen im Intelligenztest. Das Ausmaß der Computererfahrung zum ersten Meßzeitpunkt korrelierte zu r=.41 mit dem Aggregat der folgenden drei Steuerungsdurchgänge in der „Schneiderwerkstatt" und zu r=.40 mit den ein Jahr später erhobenen Steuerungsleistungen. Die Interpretation der Daten wird allerdings dadurch erschwert, daß die Computererfahrung eben auch mit den Intelligenz- und Wissensmaßen konfundiert war.

Das in der Berliner Untersuchung zur Erfassung der Computerkenntnisse via Selbsteinschätzung eingesetzte Instrument wurde - neben anderen - auch in der Mannheimer Studie zu Determinanten des komplexen Problemlösens eingesetzt (Wittmann et al., 1996, S.16 f). Erneut zeigten sich signifikante Korrelationen der so erhobenen Computererfahrung zur Steuerungsleistung bei den Systemen „Schneiderwerkstatt" und „PowerPlant". Bei der nur von einer Teilstichprobe gesteuerten Simulation „Learn" verfehlte die entsprechende Korrelation hingegen knapp die Signifikanzgrenze. Auch in dieser Studie war die Computererfahrung nicht nur mit der Steuerungsleistung, sondern auch mit der Intelligenztestleistung konfundiert.

Zusammenfassend kann der Verdacht, daß die Problemlöseleistung zu einem Teil durch das Ausmaß an Computererfahrung bedingt ist, gegenwärtig nicht ausgeräumt werden. Zwar liegen Studien vor, in denen sich keine entsprechenden Effekte zeigten, allerdings sind hier Zweifel an der Qualität der (teilweise Single-Item) Messung der Computererfahrung angebracht. Entsprechende Befunde lassen sich erst dann interpretieren, wenn die Güte der Messung gesichert ist - dieser psychometrische Aspekt wird aber meist erst gar nicht thematisiert. Neben psychometrischen Überlegungen sind aber auch inhaltliche Bedenken anzumelden. Der Befund einer vermeintlichen Unabhängigkeit von Problemlöseleistungen, die über computergestützte Szenarien erhoben werden und der Computererfahrung steht im Widerspruch zu entsprechenden Erkenntnissen aus anderen Bereichen. So berichten Kleinmann und Strauß (1995, S. 128), daß nach den Ergebnissen der Forschung zur Softwareergonomie je nach Programmgestaltung große Differenzen zwischen erfahrenen Computernutzern und Laien bei der Bearbeitung des eigentlichen Sachproblems zu erwarten sind. Hamborg (1996) kommt nach einer knappen Literaturübersicht und nach einem eigenen Experiment zu Fehlerhäufigkeiten bei der Nutzung von Anwender-Software durch Experten und Novizen ebenfalls zu dem Schluß, daß Anfängern bei der Arbeit mit Software mehr Ineffizienzen und mehr Fehler der intellektuellen Regulationsebene (Fehler, die mit Denk- und Planungsprozessen verbunden sind) als Experten unterlaufen. In der Studie des Autors mit unterschiedlich komplexen Textverarbeitungssystemen zeigte sich eine Interaktion von Expertise und Programmkomplexität. Computeranfänger begehen - in Relation zu Experten - bei komplexen Systemen mehr Fehler und arbeiten ineffizienter als mit einem Sy-

stem von geringer Komplexität. Gerade die zuletzt genannte möglicherweise moderierend wirkende Komplexität der Programme läßt über das allgemeine Problem der computergestützten Diagnostik hinaus ein möglicherweise spezifisches Problem der Diagnostik mit computergestützten Problemlöseszenarien aufscheinen, da die Steuerung eines Szenarios in der Regel komplexer ist als beispielsweise die Reaktion auf computergestützt dargebotene Intelligenztestitems.

Bislang wurden lediglich (teilweise untaugliche) Versuche unternommen, die Effekte der Computer*erfahrung* auf die Problemlöseleistung zu erfassen. Mögliche Effekte der *Einstellung* zur Arbeit mit Computern fanden hingegen bislang überhaupt keine Berücksichtigung. Eine solche Beschränkung auf den Aspekt der Computererfahrung wäre aber nur dann sinnvoll, wenn Computererfahrung und Einstellung hochgradig korreliert wären, so daß z.B.- wie häufig angenommen wird - Personen durch die Erfahrung mit Computern von ihren negativen Einstellungen gegenüber Computern „automatisch geheilt" würden. Dies ist nicht der Fall, vielmehr schlagen sich negative Einstellungen und Emotionen gegenüber Computern selbst bei gegebener Computererfahrung negativ auf die Lösung der jeweils computergestützt dargebotenen Aufgaben nieder (siehe z.B. Rosen & Maguire, 1990).

Diagnostisch bedeutet die Ungewissheit hinsichtlich der Effekte der Computererfahrung und -einstellung auf den Umgang mit Szenarien, daß das Problemlösegütemaß nur zu einem Teil als Indikator für die interessierende Fähigkeit gewertet werden kann. Zu einem weiteren - unbekannten - Teil ist das Problemlösegütemaß möglicherweise auch ein Indikator für ein anderes Personmerkmal. Selbst wenn die Computererfahrung als Qualifikationsmerkmal in einzelnen Fällen von zusätzlichem diagnostischen Interesse sein mag, empfiehlt es sich nicht, die Erhebung dieses Merkmals mit einem anderen Merkmal unauflöslich in einem Wert zu vermengen. Die sich in einigen Untersuchungen andeutenden Leistungsunterschiede zwischen Gruppen mit und ohne Computererfahrung können außerdem als Hinweis auf ein mögliches Fairneßproblem (siehe oben) gedeutet werden. Dieses potentielle Fairneßproblem trifft zunächst allgemein die Gruppe der computerunerfahrenen Diagnostikanden. Die Tatsache, daß sich die Computererfahrung und -einstellung aber geschlechtsspezifisch unterscheidet, läßt aus dem allgemeinen potentiellen Fairneßproblem aufgrund der gruppenspezifischen Computererfahrung ein potentielles geschlechtsspezifisches Fairneßproblem erwachsen. Es ist zur Zeit nicht auszuschließen, daß Frauen durch den diagnostischen Einsatz von Problemlöseszenarien diskriminiert werden. Diese Schlußfolgerung auf eine potentielle Benachteiligung der Frauen soll im folgenden etwas ausführlicher hergeleitet werden. Ebenso denkbar ist aber auch eine altersspezifische Ausprägung der Computerkenntnisse und -einstellungen, so daß sich auch bezüglich dieses demographischen Merkmals Gruppeneffekte zeigen könnten - auf diesen Aspekt wird im Theorieteil der vorliegenden

Arbeit nicht weiter eingegangen, weil diesbezüglich zur Zeit keine Studien aus dem Kontext der Problemlöseforschung vorliegen.

10.1.3 Geschlechtsspezifische Unterschiede in der Problemlöseleistung sowie in der der Computererfahrung und -einstellung

Über eine Problemlöseleistungsdisparität zuungunsten der Frauen berichteten Kreuzig und Schlotthauer (1991, S. 109), Locher (1997, S. 121f.) Süß et al. (1992) sowie Wittmann et al. (1996).

Diese geschlechtsspezifischen Leistungsunterschiede können im Kontext der im vorherigen Abschnitt thematisierten Computererfahrung und -einstellung diskutiert werden. In der Berliner Untersuchung (siehe Süß, 1996, S. 184) war die Computererfahrung zu r=.66 (Erstuntersuchung) bzw. zu r=.63 (Wiederholungsuntersuchung) mit dem demographischen Merkmal „Geschlecht" korreliert. Gerade drei von 97 Schülerinnen erreichten hinsichtlich ihrer Computererfahrung den Mittelwert der Schüler. Während sich die Geschlechtsunterschiede in der Problemlöseleistung in der Berliner Untersuchung zu einem großen Teil (aber nicht vollständig) auf die vergleichsweise geringer Computererfahrung der Frauen zurückführen ließ, blieb es in der Mannheimer Untersuchung (Wittmann et al., 1996) auch dann bei dem Befund signifikant geringerer Steuerungsleistungen der Frauen, wenn die Computererfahrung auspartialisiert wurde.

Die Frage, ob es bei der Bearbeitung von computergestützten Problemlöseszenarien geschlechtsspezifische Besonderheiten gibt, bedarf dringend der Klärung. Dabei empfiehlt es sich, diese Frage gemeinsam mit dem Effekt von Computererfahrungen und -einstellungen auf das Problemlöseverhalten zu untersuchen. Mädchen und Frauen erweisen sich im Durchschnitt nämlich als weniger computererfahren als Jungen und Männer und zeigen im Geschlechtervergleich eine vergleichsweise desinteressierte, skeptische bis ablehnende Haltung gegenüber Computern.

Krahn (1990, S. 179) fasst die Ergebnisse verschiedener Studien zum Thema „Mädchen und Computer" wie folgt zusammen: *„Zum gegenwärtigen Zeitpunkt haben Mädchen in der Regel weniger Kennntnisse über Computer als Jungen, seltener Zugang zu einem eigenen Gerät, sind seltener bei Computerspielen anzutreffen und wenden insgesamt weniger Zeit für den Computer auf als ihre Klassenkameraden."* Auch zu späteren Zeitpunkten erhobene Daten bestätigen im wesentlichen diese Aussage (z.B. Lang, 1992). Die Befunde sind nicht auf die Kindheit und Jugend beschränkt, sondern gelten auch für das Erwachsenenalter. Im universitären und im beruflichen Kontext erwerben Männer einer Übersicht von Rosen und Maguire (1990, S. 185) zufolge beispielsweise mehr Computererfahrung als

Frauen. Als Anhaltspunkt für das unterschiedliche Interesse von Frauen und Männer kann auch das Ergebnis einer Marktanalyse gelten, (berichtet nach Dworschak in der Wochenzeitschrift „Die Zeit", Nr. 47, 18) die besagt, daß 80% der Leserschaft von PC-Fachzeitschriften männlich sind.

Krahn (1990) sieht in der geringeren Computererfahrung der Mädchen keinen Ausdruck eines generell geringeren Interesses der Mädchen an den neuen Technologien. Vielmehr setzen die Mädchen ihrer Ansicht nach andere Interessensschwerpunkte, indem sie sich mehr für Anwendungsmöglichkeiten und gesellschaftliche Auswirkungen der Technik interessieren. Diese Einschätzung steht in Einklang mit der allgemeinen Feststellung von Todt (1992, S. 315), *„daß vor allem Mädchen bei ihrer Reaktion auf Technik davon abhängig sind, inwieweit die Technik so eingebettet bzw. funktionalisiert ist, daß sie für das Leben und die Lebensziele der Mädchen bedeutsam ist."*

Als Ursachen für die Technik-Distanz der Mehrheit der Mädchen und Frauen werden geschlechtsspezifische Rollenzuschreibungen und -erwartungen angesehen. Um den bildungspolitische unerwünschten Trend zu geschlechtsspezifisch unterschiedlich ausgeprägten Erfahrungen mit und Einstellungen gegenüber der „Zukunftstechnologie" Computer entgegenzusteuern, werden entsprechend Unterrichtsmethoden und -inhalte gefordert, die geeignet sind, die Rollenstereotypen aufzubrechen (siehe z.B. Niederdrenk-Felgner, 1993).

10.2 Zur Verfälschbarkeit der mit Hilfe von computergestützten Problemlöseszenarien gewonnenen diagnostischen Informationen

Wie bei anderen Leistungstests können auch bei Problemlöseszenarien Leistungs*defizite* simuliert werden. Im Unterschied zu anderen Prüfverfahren - insbesondere zu Intelligenztest - lassen sich aber die Ergebnisse der Bearbeitung von Problemlöseszenarien auch im Sinne einer *positiven* Leistung relativ leicht und weitgehend verfälschen. Damit sind nicht die üblichen Traininigseffekte gemeint, die für Problemlöseszenarien (siehe z.B. K.J. Klauer, 1996a) prinzipiell ebenso wirksam werden dürften wie für andere Leistungsprüfverfahren (siehe z.B. die Zusammenstellung der entsprechenden Effekt-Größen für Trainings aus sieben Metaanalysen bei Lipsey und Wilson, 1993). Hinsichtlich der Höhe dieses prinzipiellen Trainingseffekts muß in Übertragung einer von Schuler (1996, S. 131) im Assessment-Center Kontext angestellten Überlegung allerdings befürchtet werden, daß Trainingseffekte bei Problemlöseszenarien besonders hoch ausfallen, da diese Aufgabentypen stark

von der Art der Herangehensweise abhängig sind und einen hohen Neuigkeitsgrad aufweisen. Besonders thematisiert werden soll hier aber der bislang nicht beachtete Aspekt, daß (1) eventuell bereits durch „Zurückhaltung" in der Steuerung und/oder (2) auf jeden Fall durch die Anwendung einfachster und leicht kommunizierbarer Regeln, Leistungen simuliert werden können, die die Verdienste sich redlich bemühender Diagnostikanden möglicherweise weit übertreffen.

Die Ursache dieser Besonderheit von Problemlöseszenarien liegt in der bereits mehrfach angesprochenen (Eigen-)Dynamik der Systeme. Diese Dynamik kann - auch in Kombination mit der Überforderung der Steuerer, siehe Kapitel 7 - dazu führen, daß einerseits Eingriffe in das Szenario fatale Folgen zeitigen, während andererseits bei keinen oder sehr zurückhaltenden Eingriffen letztendlich die Voreinstellung des Szenarios über das Endergebnis entscheidet. Häufig liegen die Werte, die man mit einem „ungesteuerten Null-Lauf" erzielt, über den Werten zahlreicher „agierender" Problemlöser. So brachten es - siehe oben, Abschnitt 7.3 - beispielsweise mehr als die Hälfte aller Probanden der Berliner Erstuntersuchung im ersten Steuerungsdurchgang mit ihren Eingriffen bei der (sehr schwer handhabbaren) „Schneiderwerkstatt" zu Ergebnissen, die schlechter waren, als wenn sie die Voreinstellung unverändert übernommen hätten. Anders formuliert: wer sich in einer solchen kaum zu steuernden Situation der aktiven Aufgabenbearbeitung weitgehend enthält, kann eine Leistung vortäuschen, die besser ist als diejenige der Hälfte seiner aktiven Konkurrenten. Eine solche „Zurückhaltung" zu Zwecken der Leistungssimulation muß nicht unbedingt auffallen. In vielen Szenarien kann man minimale Veränderungen weniger zentraler Variablen „gefahrlos" vornehmen (z.B. Variablen in einem Takt um 5 % herauf und im nächsten wieder runter setzen), so daß sich eine solche „Täuschung" nicht ohne weiteres im Rechnerprotokoll nachweisen läßt. (Wer will nach welchen Kriterien entscheiden, wann es sich um eine „Verfälschung" und wann um einen „zurückhaltenden" Steuerer handelt? Auch muß es sich ja nicht um eine vorsätzliche Täuschung handeln, vielleicht ist der Steuerer wirklich zurückhaltend - ist er deshalb aber auch automatisch positiv zu bewerten?). Dieser Verfälschungsgefahr kann man zwar durch eine ungünstigere Voreinstellung (mit negativer Eigendynamik) entgegenwirken, damit erhöht man aber wieder die nicht zu unterschätzende Gefahr, eine zu schwere Aufgabe zu schaffen (siehe Kapitel 7).

Ein hinsichtlich der Aussagekraft der Instrumente bedrohliches Ausmaß an Verfälschbarkeit ergibt sich, wenn man in Rechnung stellt, daß einige Diagnostikanden erfolgreich bemüht sein werden, sich vorab zumindestens ein minimales Wissen über das diagnostische Verfahren zu verschaffen, mit dem sie konfrontiert werden. Hinsichtlich der Intelligenzprüfverfahren hat dieses Interesse der Diagnostikanden nach einer „Testvorbereitung" mittlerweile einen veritablen Markt an „Testknackerbüchern" geschaffen. Ein vergleichbarer Organisationsgrad der Informationsmög-

lichkeiten über Problemlöseszenarien steht zwar (noch) aus, es wäre aber naiv und fahrlässig in Rumpelstilzchen-Manier anzunehmen, daß es diesbezüglich überhaupt kein Informationsinteresse und keine Informationsquellen gibt. Gerade bei regelmäßig durchgeführten internen Auswahlverfahren einer Organisation mit ein und demselben Szenario ist mit den Effekten eines „Erfahrungsaustausches" unter Kollegen mit und ohne Szenarienerfahrung zu rechnen. Auch Erfahrungen aus szenarienähnlichen „Spielen" sowie die Ergebnisse systematischer Literaturrecherchen können in die „inoffiziellen" Vorabinformationen eingebracht werden. Das - wie auch immer erworbene - Vorabwissen kann dazu führen, daß die Diagnostikanden durch die unreflektierte Anwendung simpelster Regeln eine hervorragende Problemlöseleistung simulieren. So reicht es bei einigen Szenarien, die sich in ihrer semantischen Einkleidung an einen Produktionsbetrieb anlehnen, z.B. meistens aus, lediglich „ausreichend Rohmaterial einzukaufen", und schon läuft die „Fabrik" wie das Schiff der Phäaken, von dem Homer erzählt, daß es ohne Steuermann geradeaus in den Hafen fährt. Wendet man bei dem Szenario „DISKo" beispielsweise diese simple Regel an und kauft über 12 Takte hinweg lediglich stets im gleichen Umfang (z.B. 1000) „Rohmaterial" ein, so erzielt man hinsichtlich der Zielvariablen „Gesamtvermögen" ein Ergebnis, welches nur noch von neun Prozent der - dem Programm in Form von Normwerten beiliegenden - studentischen Vergleichsgruppe überboten wird. Tatsächlich ist davon auszugehen, daß sich Diagnostikanden auch mehr Informationen als nur den Hinweis auf eine Variable merken können. Mit relativ simplen Informationen über eine günstige Variablenkonstellation sowie über einige Faustregeln zur Steuerung lassen sich aber zahlreiche Szenarien in den Bereich von überdurchschnittlichen Ergebnissen steuern. Dies haben die Versuche gezeigt, in denen Szenarien nicht von Probanden, sondern von Programmen - teilweise mit nur wenigem bzw. ohne Strukturwissen - auf einem Leistungsniveau gesteuert wurden, welches durchaus dem Niveau „richtiger" Probanden entsprach (siehe z.B. Kluwe, 1991; Kluwe et al., 1989, 1991b; Ringelband et al., 1990; Schaub, 1993; Schoppek, 1996). Die Ergebnisse der Bearbeitung von Problemlöseszenarien müssen daher als verfälschbar eingestuft werden. Diese Verfälschbarkeit ist in einem weit größerem Ausmaß gegeben als bei Intelligenztests. Auch die Bearbeitung von Intelligenztests kann durch Vorabinformation und Trainings positiv beeinflußt werden. Kulik, Bangert-Drowns und Kulik (1984) berichten beispielsweise als Resultat einer Meta-Analyse über 38 Studien zu diesem Thema eine durchschnittliche Effektgröße des Trainings in Höhe von 0.33. Dies bedeutet bei Tests mit einem Mittelwert von 100 und einer Standardabweichung von 10, daß sich eine durchschnittliche Leistung durch ein Testtraining von 100 auf 103 verbessern ließe. Diesem moderaten Effekt durch ein Training stehen bei Problemlöseszenarien enorme Effekte durch die Befolgung von simplen Regeln gegenüber. Der oft und

gern als Nachteil gescholtene Aspekt der „Statik" von Intelligenztestaufgaben impliziert eben auch, daß sich die Aufgaben nicht „von allein" lösen. Weder „Zurückhaltung" noch ein auswendig gelerntes „Standardrepertoire" an Verhaltensweisen nehmen dem Diagnostikanden die Lösung von Intelligenztestaufgaben ab.

Man könnte nun geneigt sein, der Verfälschbarkeit der Ergebnisse der Szenarienbearbeitung durch permanente Variation der Voreinstellungen und/oder durch die Berücksichtigung von Verhaltensmaßen entgegenzuwirken. Die beständige Änderung der Voreinstellung geht allerdings zu Lasten der Vergleichbarkeit und vereitelt die notwendige Sammlung von Normdaten (siehe den nächsten Abschnitt). Der Rekurs auf Verhaltensmaße gleicht im Rahmen der Diskussion um die „Verfälschbarkeit" hingegen der Idee, den Teufel mit dem Beelzebub auszutreiben. Gerade erwünschte leistungsunabhängige „Verhaltensweisen" wie z.B. „Informationsfragen am Anfang stellen" lassen sich problemlos simulieren, eine auf die Bewertung von „Verhaltensmaßen" abgestelltes Problemlöseszenario ist hinsichtlich der „Verfälschbarkeit" mit Persönlichkeitstests zu vergleichen, die in dieser Hinsicht als besonders anfällig gelten. Lediglich die Willkür und Beliebigkeit der Verhaltensbewertung (siehe oben, Abschnitt 6.3.2.2) könnten hier als Argument gegen die These der Verfälschbarkeit angeführt werden. Eine willkürliche und intransparente Bewertung des Verhaltens vereitelt natürlich Effekte der „positiven Selbstdarstellung", da der Diagnostikand keine Chance hat zu erkennen, was positiv bewertet wird - dies wäre allerdings ein abgründiges Gegenargument.

10.3 Spezifische und grundsätzliche Probleme der Normierung der mit Hilfe von computergestützten Problemlöseszenarien gewonnenen diagnostischen Informationen

Zahlreiche Fragestellungen der angewandten Diagnostik erfordern empirisch gesichertes Vergleichswissen, um das Verhalten der zu diagnostizierenden Person im interindividuellen Vergleich mit anderen Personen aus einer relevanten Referenzpopulation beurteilen zu können. Die bei der Bearbeitung eines Problemlöseszenarios anfallenden Werte stellen zunächst - wie bei jedem Test - wahllose Werte mit willkürlichen Nullpunkten dar, die der *Eichung* bedürfen.

Das Ausmaß und die Qualität der vorliegenden Normen für Problemlöseleistungen ist - zum Zeitpunkt der Erstellung der vorliegenden Arbeit - unzureichend. Zunächst ist zu beklagen, daß häufig lediglich Anfallsstichproben geringen Umfangs zur Eichung herangezogen wurden. Obermann (1991, S. 13) beziffert den Umfang der Normierungsstichprobe für das Szenario „Airport" auf 72 Personen. Für das

Szenario „Heizölhandel" stehen - laut dem Handbuch von Hasselmann und Strauß (1993a, S. 37) - zwei Vergleichstichproben zur Verfügung: eine Gruppe „Berufsanfänger" (105 Studenten) sowie eine Vergleichsstichprobe „Führungsnachwuchskräfte", die auf den Daten von 17 (sic!) Bankmitarbeitern basiert. Die Anwender des Szenarios „DISKo" können ebenfalls auf zwei Vergleichsstichproben zurückgreifen, nämlich auf Normwerte für 150 Naturwissenschaftler und Ingenieure sowie auf Normwerte einer Gruppe von Studierenden. Zur Größe der Studentengruppe finden sich im Manual (U. Funke, 1992a, S. VI-12) keine Angaben.

Selbst wenn man in Rechnung stellt, daß mit der Zeit Daten einer größeren Anzahl von Personen zur Verfügung stehen dürften und man weiterhin die durchgängig gegebene Möglichkeit zur Neuanlage (spezifischer) Stichproben als ein grundsätzlich positives Merkmal einer computergestützten Diagnostik würdigt, so ändern diese Anmerkungen doch nichts an der Tatsache, daß die Eichung der hier beschriebenen Szenarien (bei anderen Anbietern sind zu diesem wichtigen Punkt teilweise gar keine Angaben verzeichnet) zunächst auf unzureichender Datenbasis erfolgte. Diese Klage gilt umso mehr, da - wie weiter oben gezeigt wurde - geschlechts- und altersspezifisch ausgeprägte Problemlöseleistungen bei einigen Anwendern möglicherweise auch den Wunsch nach Gruppennormen wecken.

Während die ungünstige Situation der unzureichenden Datenbasis prinzipiell überwunden werden kann, generiert die vergleichsweise ungünstige Reliabilität (siehe oben, Kapitel 8) der Problemlösegütemaße ein grundsätzliches Problem bei der Testeichung. Bei unzureichender Reliabilität kann der Standardmeßfehler die Messgenauigkeit der Eichskala übertreffen (siehe Lienert, 1967, S. 314 f.) Die vergleichsweise geringe Reliabilität der Problemlösegütemaße zwingt außerdem zur Wahl einer recht groben Norm. Ein weitere Einschränkung bei der Normierung ist durch die - für Problemlöseszenarien häufig berichtete - Verletzung der Normalverteilungsannahme bedingt. Dieser Umstand bedingt ebenfalls informationsarme verteilungsfreie Normen.

10.4 Zur Ökonomie und Praktikabilität diagnostisch genutzter computergestützter Problemlöseszenarien

Schuler (1996, S. 174) sieht in dem vom Testkuratorium genannten Aspekt der „Ökonomie" einen Teilaspekt der organisationalen *Effizienz* oder *Praktikabilität* eines Verfahrens. Neben der „Ökonomie" zählen hierzu laut Schuler (ebd.) auch die Aspekte „Aufwand", „Ziel", „Schwierigkeit (Mühe, Kompetenzerfordernis)" und „Verfügbarkeit". Vernachlässigt man die für alle Verfahren geltenden Effekte der

Selektions- und Basisquote auf den „Nutzen" des Verfahrens, so dürfte sich auch beim Einsatz computergestützter Problemlöseszenarien - setzt man einmal eine befriedigende Kriteriumsvalidität voraus - ein betriebswirtschaftlicher Nutzen ergeben. Ein direkter Nutzensvergleich mit Intelligenztestverfahren muß im konkreten Einzelfall berechnet werden, allgemein läßt sich unter Praktikabilitäts-Gesichtspunkten lediglich auf die verfahrensspezifischen Stärken und Schwächen verwiesen. So sind Problemlöseszenarien in der Regel teurer in der Anschaffung. Während paper-pencil-Intelligenztests in großen Gruppen appliziert werden können, wird die Gruppengröße bei der Anwendung von Szenarien durch die Anzahl der in einem Raum verfügbaren Personalcomputer begrenzt. Die Computer stehen am Tag der Testung nicht für ihre eigentliche Nutzung zur Verfügung. Die zuletzt genannten potentiellen Praktikabilitäts-Nachteile können allerdings vernachlässigt werden, falls ohnehin nur eine kleine Anzahl an Personen diagnostiziert werden soll. Neben den allgemeinen Fertigkeiten in der Testdurchführung sollte die Untersuchungsleitung im Fall der Problemlöseszenarien auch noch über ein gewisses Maß an Computererfahrung verfügen. Während der Testdurchführung selbst ist die Versuchsleitung bei den Problemlöseszenarien wenig gefordert, die vereinzelt vorgeschlagene „Selbstadministration" (U. Funke, 1992a, S. II-6) schießt allerdings über das Ziel der Praktikabilität hinaus. Als praktisch anzusehen ist die bei Problemlöseszenarien rasch und automatisch erstellte Auswertung und Befunderstellung, ein Vorteil gegenüber der paper-pencil-Variante von Intelligenztests. Viele der hier genannten Praktikabilitäts Vor- und Nachteile der Problemlöseszenarien spiegeln ohnehin keine verfahrensspezifischen, sondern *mediumsspezifischen* Gesichtspunkte und würden entsprechend genauso für computergestützt dargebotene Intelligenztestverfahren gelten. Allein unter dem Aspekt der Ökonomie bzw. allgemeiner unter Praktikabilitätsgesichtspunkten spricht sicher nichts gegen den Einsatz von Problemlöseszenarien, sofern die Berücksichtigung der Erkenntnisse dieses Instruments die Validität der diagnostischen Entscheidung gegenüber einer isolierten Betrachtung der jeweiligen Verfahrensalternativen nachweislich erhöht. Gerade bei Arbeitsplätzen mit hoher Wertschöpfung und/oder großen Leistungsdifferenzen lohnen sich in der Regel eignungsdiagnostische Verfahren (siehe z.B. Barthel, 1988; Funke & Barthel, 1990) und mit hoher Wahrscheinlichkeit lohnen - unter der Voraussetzung der Kriteriumsvalidität - auch die vergleichsweise hohen Anschaffungskosten von Problemlöseszenarien.

10.5 Zusammenfassung, Schlußfolgerungen und Ausblick

Eine Prüfung der Fairneß von Entscheidungen, die aufgrund der Ergebnisse computergestützter Problemlöseszenarien getroffen werden, steht noch aus. Bedenklich sind in diesem Zusammenhang die bislang unzureichend geklärten Effekte der Computererfahrung und -einstellung auf die Problemlöseleistung. Da die Computererfahrung und -einstellung geschlechtsspezifisch (und möglicherweise auch altersspezifisch) variiert, werden u.U. mit dem diagnostischen Einsatz von computergestützten Problemlöseszenarien Frauen und Alte unter bestimmten Fairneßgesichtspunkte diskriminiert. Auch wenn eine solche Diskriminierung zur Zeit nicht empirisch nachgewiesen werden kann, so gilt es doch, den begründeten Verdacht zu prüfen. Die für Analysen auf Subgruppenebene benötigte Stichprobengröße konnte in der vorliegenden Studie aber nicht erzielt werden, so daß der empirische Teil der Arbeit keinen Beitrag zur Frage der geschlechtsspezifischen Leistungsdifferenzen und Kriteriumsvalidität leistet. Der generelle Effekt der Computererfahrung und -einstellung auf die Problemlöseleistung soll im Empirie-Teil allerdings Berücksichtigung finden.

Steuerungsleistungen können – je nach Szenario – möglicherweise verfälscht werden, indem die Diagnostikanden entweder eine positive Voreinstellung mehr oder minder unverändert übernehmen oder sich Vorabinformationen über günstige Parameter sowie über „Faustregeln" zur Steuerung verschaffen. Aus der Szenarienbearbeitung abgeleitete Verhaltensmaße sind in einem ungleich stärkeren Ausmaß von der Verfälschbarkeit betroffen.

Hinsichtlich der Normierung sind im Einzelfall die unzureichende Quantität der Eichstichproben, grundsätzlich aber die reliabilitätsbedingten Genauigkeitseinbußen der Normen zu beklagen. Auch die Tatsache, daß die Problemlösegütemaße häufig nicht normalverteilt sind, wirkt sich in Form von verteilungsfreien Normen ungünstig auf die Informationsausschöpfung aus.

Unter Praktikabilitätsgesichtspunkten unterscheiden sich computergestützte Problemlöseszenarien nicht wesentlich von anderen computergestützt dargebotenen diagnostischen Verfahren, mediumsspezifische Vor- und Nachteile sind hier eher im Kontrast zu paper-pencil Anwendungen auszumachen. Vernachlässigt man die für alle Verfahren geltenden Effekte der Selektions- und Basisquoten sowie weitere Spezifika des diagnostischen Entscheidungsprozeßes, so kann der ökonomische Nutzen von Problemlöseszenarien dann als gegeben angesehen werden, wenn der Nachweis der Kriteriumsvalidität erbracht ist.

11. Fragestellungen

Die im folgenden dargestellte Studie verfolgte allgemein die Frage, inwieweit computergestützte Problemlöseszenarien diagnostischen Qualitätsstandards gerecht werden und die Frage, ob bzw. unter welchen Umständen die beim Umgang mit Problemlöseszenarien erzielten Steuerungsleistungen als Ersatz oder Ergänzung der herkömmlichen Fähigkeitsdiagnostik mit Intelligenztests genutzt werden können.

Zunächst sollten einige allgemeine Voraussetzungen des diagnostischen Einsatzes von Problemlöseszenarien geprüft werden. Dabei wurde hinterfragt, ob die Indikatoren der Problemlöseleistung - die Gütemaße - tatsächlich das interessierende leistungsrelevante Verhalten der Testanden intern valide abbilden und ob der Einfluß fördernder oder hemmender Rahmenbedingungen der Szenarien weitgehend ausgeschlossen werden kann. Über die Analyse irrelevanter äußerer Bedingungen hinaus sollte auch der Einfluß von Personmerkmalen auf die Steuerungsleistung untersucht werden, nämlich der Einfluß der Computererfahrung und der Einstellung gegenüber Computern sowie die Effekte des allgemeinen szenarienspezifischen Vorwissens und des Alters.

Durch den Einsatz von zwei Problemlöseszenarien wurde ein empirischer Zugang zur Frage nach der Generalisierbarkeit von Steuerungsleistungen und somit zu einem Aspekt der Frage nach der Konstruktvalidität eröffnet. In den diesbezüglichen Analysen ging es darum zu klären, in wie weit der Erfolg beim Problemlösen von dem spezifischen System abhing, welches jeweils zur Diagnose verwandt wurde.

Die Konstruktvalidität war auch Gegenstand der Analysen zu den kognitiven Voraussetzungen der Steuerungsleistungen. Konkret sollte die Analyse des Zusammenhangs von Steuerungsleistungen mit Intelligenz- und Wissensleistungen Aufschluß darüber geben, ob es sich bei der Problemlösefähigkeit um eine gegenüber Intelligenz und Wissen gesonderte und eigenständige Fähigkeit handelt. Dieser Aspekt ist für den möglichen Einsatz von computergestützten Problemlöseszenarien zur Fähigkeitsdiagnostik von zentraler Bedeutung. Sowohl die Auswahl von geeigneten Meßinstrumenten bei einer gegebenen diagnostischen Fragestellung als auch die Interpretation der diagnostischen Daten setzt Annahmen über das durch die verfahrensspezifischen Daten indizierte Konstrukt voraus.

Im Mittelpunkt der Studie stand die Frage nach der Bedeutung der computergestützten Problemlöseszenarien für die eignungsdiagnostische Praxis. Zunächst sollte geklärt werden, ob die bei der Bewältigung von Problemlöseszenarien bzw. die bei der Lösung von Intelligenztestaufgaben mutmaßlich unter Beweis gestellten Lei-

stungen und Fähigkeiten nach Einschätzung von berufserfahrenen Personen für den Berufserfolg relevant sind. Primäres Ziel der Untersuchung war die Ermittlung empirischer Anhaltspunkte für die Kriteriumsvalidität computergestützter Problemlöseszenarien, wobei die bislang empirisch kaum untersuchte prädiktive Validität der Szenarien im Vordergrund stand. Wie treffsicher - so lautete die Leitfrage - sind die aufgrund computergestützter Problemlöseszenarien getroffenen Diagnosen in bezug auf die im Berufsalltag benötigten und gezeigten kognitiven Leistungen und Fähigkeiten? Weiterhin sollte geklärt werden, welchen zusätzlichen Nutzen der Einsatz von Problemlöseszenarien gegenüber den vorhandenen diagnostischen Verfahren der intellektuellen Leistungstests bringt. Intelligenz und Problemlösen stellen auf der Ebene der theoretischen Sprache zumindest deutlich überlappende Bereiche dar, so daß der gemeinsame Einsatz der jeweils zugeordneten Meßinstrumente die theoretische Auseinandersetzung um empirische Argumente ergänzen sollte.

Die Darstellung der zur Klärung der Fragen durchgeführten Studie und ihrer Ergebnisse gliedert sich in zwei Teile. Zunächst (Kapitel 12 bis 15) werden die Methoden und die Befunde der Prädiktorerhebung beschrieben und diskutiert. Dieser Teil umfaßt die Prüfung der Voraussetzungsfreiheit der Steuerungsleistung (Kapitel 14) und die empirischen Befunde zur Konstruktvalidität der Steuerungsleistung (Kapitel 15). Der zweite Teil ist dann der Darstellung und Interpretation der Methoden und Befunde zur retrograden und konkurrenten (Kapitel 16) sowie zur prädiktiven Kriterienerhebung (Kapitel 17) gewidmet. Den Abschluß bildet die Diskussion der Befunde in Kapitel 18.

12. Untersuchungsmethodik

12.1 Untersuchungsteilnehmer der Prädiktorenerhebung

Als Zielgruppe der Untersuchung wurden Vertreter einer Berufsgruppe ausgewählt, die mutmaßlich beruflich häufig mit - unter Zeitdruck zu bewältigenden - dynamischen, komplexen, vernetzten und intransparenten Problemen konfrontiert sind, nämlich Führungskräfte (gehobener und höherer Dienst) der Polizei. Die Auswahl der Untersuchungsgruppe entsprach also der in Kapitel 3.1.1 beschriebenen „plausibilitätsbedingten Anforderungskorrespondenz", die auf der Ebene der Theoriesprache erzielte tatsächliche Übereinstimmung zwischen beruflichen und szenarienspezifischen Anforderungen sollte im Rahmen der Arbeit empirisch überprüft werden (siehe Abschnitt 17.2.2). Die Untersuchungsgruppe setzte sich aus 104 überwiegend (92 %) männlichen Angehörigen des Polizeivollzugsdienstes im Alter zwischen 28 und 57 Jahren (Median =35; SD=6,26) zusammen. 55 Beamte (53%) waren der Kriminalpolizei, 47 Beamte (45 %) der Schutzpolizei und 2 Beamte der Wasserschutzpolizei in Niedersachsen (65), Schleswig-Holstein (20), Nordrhein-Westfalen (17) und Bremen (2) zugeordnet. Es handelte sich um eine berufserfahrene Gruppe. Von den 83 Personen, die nähere Angaben zu ihrer Berufserfahrung machten, verfügte jede Person über eine mindestens zehnjährige (Median= 17; SD =6,63) polizeispezifische Berufserfahrung. Als Angehörige des gehobenen Dienstes verfügten die Untersuchungsteilnehmer über die (interne) Fachhochschulreife oder über das Abitur. (Siehe die Übersicht zu den genannten Merkmalen in Tabelle 4.) Um zu gewährleisten, daß die untersuchte Gruppe eine möglichst breite Streuung im Kriteriumsverhalten - dem Berufserfolg - aufweist, wurden Personen aus unterschiedlichen Ebenen der hierarchisch strukturierten Organisation „Polizei" in der Studie berücksichtigt. Ein Kriterium für den Berufserfolg im Polizeibereich ist der - theoretisch für alle Beamten des gehobenen Dienstes mögliche, de facto aber seltene - Aufstieg in den höheren Polizeivollzugsdienst. Als mutmaßlich weniger erfolgreiche Gruppe wurden bei der Prädiktorerhebung 44 Polizeibeamte berücksichtigt, die sich zum Zeitpunkt der Prädiktorerhebung nicht in einem aktuellen Aufstiegsbewerbungsverfahren befanden. Dabei handelte es sich um Personen, die sich entweder aus eigener Entscheidung nicht für den Aufstieg beworben hatten oder aber bereits in der Vergangenheit im Aufstiegsverfahren gescheitert waren bzw. die polizeiintern gesetzten Voraussetzungen für eine Teilnahme am Aufstiegsverfahren (noch) nicht erfüllten. Demgegenüber wurden als erfolgreiche Gruppe 20 Personen in die Studie aufgenommen, die die Aufstiegsbewerbung in jüngster Vergangenheit mit Erfolg absolviert hatten und somit bereits zum Zeitpunkt der Prädiktorerhebung

sogenannte „Ratsanwärter" waren. Als mittlere bzw. noch unentschiedene Gruppe konnten schließlich 40 Polizisten gewonnen werden, die sich aktuell in der Aufstiegsbewerbung befanden. Die Berücksichtigung dieser unterschiedlichen Gruppen bedingte eine notwendige Variation bei einem der eingesetzten Meßinstrumente (Intelligenztest, siehe unten Abschnitt 12.2.3). Außerdem mußte ein systematischer Drop-out bei der Erhebung eines Kriterienmaßes in Kauf genommen werden: Die ein bis zwei Jahre später eingeholte Beurteilung der aktuellen beruflichen Leistung (siehe unten, Abschnitt 17) konnte für die erfolgreichen Aufstiegsbewerber nicht eingeholt werden, da diese ihre Berufstätigkeit für eine Ausbildung zum höheren Dienst an der Polizeiführungsakademie unterbrechen.

Ein für die Interpretation der Daten wichtiges Merkmal der Studie betrifft die Vorausgelesenheit der Gruppe nach Intelligenztests. Intelligenztests - und insbesondere Tests zur Verarbeitungskapazität - gehören in einigen Bundesländern zum Standard-Repertoire der Personalauswahlpraxis der Polizei. 85,6 % der hier analysierten Untersuchungsgruppe gaben an, bereits zu einem früheren Zeitpunkt mindestens einmal an einem Intelligenztest teilgenommen zu haben. Dabei dürfte es sich überwiegend um Intelligenzprüfungen im Rahmen von Auswahlverfahren gehandelt haben; für 80% der Gesamtgruppe der Teilnehmer war dies entsprechenden Angaben zufolge definitiv der Fall. Da die Daten früherer Untersuchungen nicht zugänglich waren, konnte der Effekt der Vorauswahl nicht kontrolliert werden. Die Tatsache, daß es sich um eine (u.a.) nach Intelligenztests vorausgewählte Gruppe handelte, legt nahe, daß die Daten der vorliegenden Studie die Vorhersageleistungen der Intelligenztests unterschätzen, und den Vergleich der Vorhersageleistungen von Intelligenz und Problemlöseszenarien zuungunsten der varianzeingeschränkten ersten Verfahrensgruppe verzerren. Dieses Risiko wurde bei der Planung der Studie in Kauf genommen, da gerade die mögliche Varianzeinschränkung auf Seiten der Intelligenzdiagnostik die Suche nach alternativen Instrumenten zur Fähigkeitsdiagnostik motiviert. Nach Schorr (1995, S. 9) sind Intelligenztests diejenigen Verfahren, die in Unternehmen am häufigsten eingesetzt werden werden. Wenn aber der Zugang von neuen Mitarbeitern aufgrund von Intelligenztestverfahren erfolgt und somit innerhalb einer Organisation eine hinsichtlich der Testintelligenz homogen leistungsstarke Gruppe erzeugt wird, ist die Verwendbarkeit der Intelligenztestverfahren für weitere Eignungsentscheidungen aufgrund der eingeschränkten Varianz in Frage gestellt. Gerade für diesen eignungsdiagnostischen „Spezialfall" könnten Problemlöseszenarien rein theoretisch eine Alternative zu Intelligenztests darstellen. Um diese Möglichkeit prüfen zu können, wurde die Gefahr einer Varianzeinschränkung auf Seiten der Intelligenz in der vorliegenden Studie billigend in Kauf genommen. Der Umstand einer möglichen Varianzeinschränkung der Intelligenztestverfahren ist bei der Interpretation diesbezüglicher Ergebnisse gedanklich zu berücksichtigen.

Tab. 4: Demographische und berufliche Merkmale der Untersuchungsgruppe

N		104
Geschlecht	Frauen	8 %
	Männer	92 %
Alter	28 / 29 Jahre	3 %
	30 - 39 Jahre	72 %
	40 - 49 Jahre	15 %
	50 - 57 Jahre	10 %
Bildung	Polizeiinterne Fachhochschulreife	46 %
	Fachhochschulreife	13 %
	Hochschulreife	41 %
Bundesländer	Niedersachsen	63 %
	Schleswig-Holstein	19 %
	Nordrhein-Westfalen	16 %
	Bremen	2 %
Aufgabenbereich	Schutzpolizeidienst (inkl. Wasserschutz)	47 %
	Kriminalpolizeidienst	53 %
Aufstiegspotential	gering (aktuell[1] keine Aufstiegsbewerbung)	42 %
	mittel (aktuelle[1] Aufstiegsbewerbung)	39 %
	hoch (Aufstiegsbewerbung erfolgreich)	19 %

[1)] „aktuell": zum Zeitpunkt der Prädiktorerhebung

12.2 Meßinstrumente und ihre psychometrische Qualität

12.2.1 Problemlöseszenarien

Mit den Szenarien „Schneiderwerkstatt" und „DISKo" wurden zwei Problemstellungen mit einer Rahmengeschichte aus dem Bereich der Wirtschaft eingesetzt. Als allgemeine Beschreibung kann für beide Szenarien gelten, daß die Testanden die

Verantwortung für alle bedeutenden Entscheidungsbereiche einer ihnen anvertrauten „Fabrik" übernehmen. Die „Fabrik" ist über einen bestimmten, in Takte oder „Monate" gegliederten Zeitraum zu steuern. Ziel der Steuerung ist die Maximierung des „Gesamtvermögens" am Ende der insgesamt zur Verfügung stehenden Bearbeitungszeit. Die Testanden können pro Takt „Entscheidungen" treffen und z.B. „Rohmaterial einkaufen", „Personal einstellen oder entlassen", „Maschinen kaufen oder verkaufen" oder die „Werbeausgaben verändern". Die Auswirkungen dieser Maßnahmen sowie ggf. die Auswirkungen von Maßnahmen vorheriger Takte werden nach Beendigung eines Taktes berechnet und den Testanden rückgemeldet. Die Information über den aktuellen Zustand ihrer „Fabrik" erfolgt über die Angabe bestimmter Kennwerte, z.B. Stand des „Gesamtvermögens". Diese Informationen können die Testanden als Ausgangsbasis für die Entscheidungen im folgenden Takt nutzen. Die beiden folgenden Abschnitte geben einige spezifische Informationen zu den Szenarien „Schneiderwerkstatt" und „DISKo", vorab soll aber noch eine Begründung für die Auswahl von Szenarien mit einer semantische Einkleidung aus dem Bereich der Wirtschaft gegeben werden.

Die Untersuchungsteilnehmer arbeiteten in unterschiedlichen Aufgabenbereichen der Polizei. Man kann annehmen, daß jeder Polizist für jeweils seinen Bereich Experte ist. Ein System mit einer Rahmengeschichte aus dem Alltag *eines* Polizeibereiches würde somit einige Teilnehmer von der wissensmäßigen Seite her privilegieren oder diskriminieren. Die semantische Einbettung aus dem Bereich Wirtschaft setzte hingegen vermutlich niemanden aufgrund seiner beruflichen Erfahrung in einen Wissensvorsprung. Darüber hinaus wurde angenommen, daß ein Minimum an Kenntnissen über Wirtschaft zum Allgemeingut gehört und somit jeder Teilnehmer einen Zugang zu der Rahmengeschichte der Systeme finden konnte. Die Annahme eines Minimums an Wirtschaftskenntnissen wurde durch den Einsatz eines entsprechenden Wissenstests überprüft (siehe Abschnitte 12.2.4 und 14.2).

12.2.1.1 Problemlöseszenario „Schneiderwerkstatt" (SWS)

Als Problemlöseszenario wurde mit der „Schneiderwerkstatt" (in der im Rahmen der Berliner Untersuchung von Süß et al. (1991) überarbeiteten Version der Fassung (2.3) von J. Funke) ein System mit einer betriebswirtschaftlichen semantischen Einbettung („Hemdenfabrik") eingesetzt. Den Testanden stellt sich die „Schneiderwerkstatt" als System mit 24 Variablen dar. Die Hälfte dieser Variablen ist zu Steuerungszwecken direkt zugänglich. Eine Vernetzungsgraphik der „Schneiderwerkstatt" findet sich bei Kersting und Süß (1995, S. 85). Nach der Erläuterung und der Übungsphase steuerten die Teilnehmer in direkter Interaktion mit dem Computer das Szenario für 40 Minuten (Zeitvorgabe) über 12 Bearbeitungstakte (sogenannte „Si-

mulationsmonate") mit der eindeutigen Zielvorgabe, das Gesamtvermögen des „Unternehmens" zu maximieren.

Die „Schneiderwerkstatt" wurde aus verschiedenen Gründen als Meßinstrument gewählt. Zunächst treffen die in Abschnitt 2.2 erläuterten Attribute komplexer Probleme auf die „Schneiderwerkstatt" zu. Ausschlaggebend war außerdem, daß mit diesem Meßinstrument in eigenen früheren Untersuchungen bereits wichtige Erfahrungen gesammelt werden konnten (Süß et al., 1991; 1993a; 1993b). Die „Schneiderwerkstatt" erlaubt eine hinreichend reliable Leistungsmessung (siehe oben, Abschnitt 8.1.1). Für die „Schneiderwerksatt" liegt - wie in Kapitel 2.3.2.3 gefordert - eine gründliche Analyse der Aufgabenmerkmale vor (Funke, 1983, 1986; Kersting, 1991). Als Ergebnis der sorgfältigen und umfangreichen Aufgabenanalyse und als Resultat einiger Modifikationen konnte für die hier verwendete Version der „Schneiderwerkstatt" mit einiger Wahrscheinlichkeit davon ausgegangen werden, daß das Problemlösegütemaß bei einer Untersuchungsgruppe mit Abitur bzw. Fachhochschulreife intern valide ist (siehe Kapitel 7). Die Möglichkeit, einen bewährten systemspezifischen und einen allgmeinen Wissenstest zur „Schneiderwerkstatt" einzusetzen (siehe unten, Abschnitte 12.2.3 und 12.2.4), sprach ebenfalls für den Einsatz dieses Instruments. Schließlich war aber auch entscheidend, daß dieses Programm den *„Klassiker unter den Szenarien mit Bezug zur Personalarbeit"* darstellt und *„die verschiedenen Varianten dieser Verfahrensfamilie bisher am häufigsten in anwendungsbezogenen Studien eingesetzt"* wurden (U. Funke, 1995a, S. 150).

12.2.1.2 Problemlöseszenario „DISKo"

Um der Frage nachgehen zu können, inwieweit Steuerungsleistungen nicht nur systemspezifisch sondern generalisierbar sind, wurde neben der „Schneiderwerkstatt" noch ein weiteres Problemlöseszenario eingesetzt. Die Auswahl dieses weiteren Systems orientierte sich zum einen explizit am Anspruch des Meßinstruments und verlangte den Einsatz eines der bereits im diagnostischen Handel befindlichen computergestützten Problemlöseszenarien. Zum anderen sollten die inhaltlichen Aufgabenmerkmale, also die semantische Einkleidung, konstant gehalten werden. Die Wahl fiel deshalb auf das *d*iagnostische, *i*nteraktive *S*zenario zur *Ko*mplexitätssimulation „DISKo" (U. Funke, 1992a). „DISKo" ist von der semantischen Einkleidung her der „Schneiderwerkstatt" vergleichbar, der Testand übernimmt auch hier die Führung eines Wirtschaftsbetriebes („Chipfabrik"). Die Anlehnung des Systems an die „Schneiderwerkstatt" wurde vom Programmautor U. Funke bewußt gewählt, um *„einen eventuellen Vergleich von Ergebnissen mit anderen Systemen zu ermöglichen"* (Schuler, Funke, Moser und Donat, 1995, S. 110). Hinsichtlich der formalen Merkmale (Anzahl und Art der Vernetzung der Variablen) unterscheidet sich diese Auf-

gabe hingegen von der „Schneiderwerkstatt". „DISKo" verfügt über 17 direkt und 24 indirekt beeinflußbare Systemvariablen. Das Szenario bietet den Teilnehmern außerdem vergleichsweise mehr Verhaltensmöglichkeiten und erfaßt mehr Verhaltensdaten. Schließlich sprach die umfangreiche und detaillierte Dokumentation (U. Funke, 1992a) für den Einsatz dieses Programms. Angaben zur Reliabilität der mit „DISKo" erzielten Messungen wurden weiter oben bereits referiert (Abschnitt 8.1.1). Auch die Tatsache, daß bei „DISKo" programmintern berechnete Parameter Informationen zum systemspezifischen Sachwissen der Testanden erfassen, wurde bei der Auswahl des Meßinstruments positiv berücksichtigt. Grundlage des hier für die Wissensdiagnose berücksichtigten programminternen Parameters sind die bei der Steuerung von „DISKo" möglichen „Testdurchläufe". Im Rahmen dieser „Testdurchläufe" kann der Testand „Probemaßnahmen" eingeben, die keine ernsthafen Konsequenzen haben, d.h. der für die Bewertung maßgebliche Systemzustand wird nicht verändert. Dem Testand wird eine Rückmeldung darüber gegeben, welche Auswirkungen seine „Probemaßnahmen" hätten, wobei er den „Zeitraum", über den der Effekt abgeschätzt werden soll (z.B. über einen oder über zwölf Takte oder „Monate"), vorab wählen kann. Nachdem die „Probemaßnahmen" eingegeben wurden, öffnet sich ein Prognosefenster, in dem der Testand Vorhersagen über die vermuteten Effekte seiner Maßnahmen auf bestimmte (von ihm selbst wählbare) Systemvariablen tätigen kann. Dabei wird lediglich nach der vermuteten Richtung des Effekts (Erhöhung oder Verminderung der abhängigen Variablen) gefragt. Der Testand erhält dann eine Rückmeldung darüber, ob seine Prognose zutreffend war oder nicht. Die Analyse der Anzahl korrekter Prognosen kann als Indikator für das „semi-quantitative Wissen" (siehe unten, Abschnitt 9.1.3.1) über den Zusammenhang von Variablen interpretiert werden („Vorzeichenwissen"). Konkret wird in der vorliegenden Arbeit als Wissensindikator für das Szenario „DISKo" die Differenz der „richtigen" abzüglich der „falschen" Prognosen verwendet. Problematisch ist diese „online"- Wissensdiagnose besonders bei allen Personen, die keine oder wenige Testdurchläufe durchgeführt haben und/oder keine oder nur wenige Prognosen abgegeben haben. Diesen Personen wird dieser Auswertung zufolge ein „neutraler" Wert für das Systemwissen attestiert. Ein anderes Problem dieser Art der Wissensdiagnose besteht darin, daß die Schwierigkeit der Prognosen bei der Bewertung keine Berücksichtigung findet. Ein Testand, der drei offensichtliche Variablenrelationen (z.B. „eine Erhöhung des Werbeetats führt zu einer Erhöhung der Nachfrage") richtig erkennt, erhält den gleichen Wert wie ein Proband, der drei weniger saliente und somit „schwierige" Variablenrelationen richtig einschätzt.

Aufgrund der Aktivitäten, die der Testand während der „DISKo"-Bearbeitung tätigt, werden programminterne Indikatoren für verschiedene Verhaltensbereiche berechnet. Der in der vorliegenden Arbeit berücksichtigte Parameter zur Beurteilung

der Verhaltensweisen und Strategien der Testanden stellt laut Handbuch (U. Funke, 1992a) eine Aggregation der Leistungen in den sechs Bereichen „Informationsgewinnung", „Analyse/Feedbacksuche", „Probehandeln/Testläufe", „Systemwissen/Hypothesen", „Entscheidungen" und „Inhalte aller Aktivitäten" dar. Für weitere Informationen über „DISKo" sei auf die entsprechenden Darstellungen bei U. Funke (1992a, 1995a) verwiesen.

Nach einer Einführung steuerten die Teilnehmer das System für 50 Minuten mit der eindeutigen Zielvorgabe, das Gesamtvermögen zu maximieren sowie sich ein größtmögliches Verständnis für die Zusammenhänge der einzelnen Variablen zu erarbeiten. Die Teilnehmer wurden instruiert, das Szenario über 12 Bearbeitungstakte (sogenannte „Simulationsmonate" mit Entscheidungen) zu steuern.

12.2.2 Wissenstests

12.2.2.1 Systemspezifischer Wissenstest zur „Schneiderwerkstatt" (WIS)

Zur Erfassung des systemspezifischen Sachwissens der Probanden über Aspekte des Systems „Schneiderwerkstatt" wurden zwei Skalen aus dem sogenannten „*WIS-2*"-Test eingesetzt. Dieser Test wurde von Kersting nach den Methoden zur Konstruktion kontentvalider Meßverfahren entwickelt und evaluiert (Kersting, 1991; Kersting und Süß, 1995). Eingesetzt wurden die Skalen „Variablen-Relationen" und „Variablen-Eigenschaften". Beim ersten Aufgabentyp werden den Probanden in 20 Items jeweils sechs Aussagen über alle Möglichkeiten der Ausgestaltung einer direkten Relation zwischen zwei Variablen der „Schneiderwerkstatt" vorgegeben, gefragt ist nach der richtigen Relation. Cronbach's alpha betrug in der hier thematisierten Untersuchung .58. Bei der Interpretation dieses Wertes ist zu berücksichtigen, daß diese Methode der Bestimmung der internen Konsistenz von eindimensionalen Tests ausgeht. Der Konstruktion des Wissenstests lagen keine entsprechenden Dimensionalitätsannahmen zugrunde. Hinsichtlich der psychometrischen Güte des Tests ist daher die weiter unten berichtete Angabe zur Stabilität sinnvoller interpretierbar.
Der zweite Aufgabentyp umfasst fünf Items. In jedem Item werden sechs qualitative Aussagen zu einer Variablen und deren Relationen getroffen. Jede zutreffende Aussage ist anzustreichen (Cronbach's alpha =.60).

Die interne Konsistenz einer Skala mit allen 25 Items aus beiden Aufgabentypen betrug alpha=.68. Die Aggregation der beiden z-transformierten Einzelskalen zu einer Gesamtskala „Sachwissen" erschien demnach gerechtfertigt.

Als Indikator der Stabilität dieser Messung kann ein Wert herangezogen werden, der in der Berliner Erst- und Wiederholungsuntersuchung berechnet wurde. In die-

sem Datensatz betrug die Stabilität des entsprechenden Aggregats über ein Jahr r=.70 (N=113). Der Test (inkl. Beispielitems) und seine Entwicklung sind, ebenso wie Daten zu seiner Bewährung, bei Kersting (1991) sowie Kersting und Süß (1995) ausführlich dokumentiert.

Für die Bearbeitung der beiden Skalen standen (inkl. einer kurzen Einführung) knapp 20 Minuten zur Verfügung.

Eingesetzt wurde außerdem die Skala „Eingriffswissen" (siehe Kersting und Süß, 1995). Die interne Konsistenz dieser Skala war in der vorliegenden Studie derart unbefriedigend, daß bei der Analyse der Untersuchungsergebnisse auf eine weitere Berücksichtigung dieser Variablen verzichtet wurde.

12.2.2.2 Allgemeiner Kenntnistest Wirtschaft (DKT-W)

Deklarierbares Wirtschaftswissen mit relativ hohem Allgemeinheitsgrad wurde mit dem Subtest Wirtschaft des „Differentiellen Kenntnis-Tests" der *Deutschen Gesellschaft für Personalwesen e.V. (DGP)* erhoben. In dieser Skala wird mit 20 Items im multiple-choice Format Wissen über die Bedeutung einfacher Wirtschaftsfachbegriffe und über wirtschaftliche Zusammenhänge erfragt. Beispielsweise sollen die Testanden die richtige Erläuterung der Begriffe *„Embargo"*, *„Schuldverschreibung"*, *„Komplementäre"* und *„Dividende"* unter jeweils vier Distraktoren herausfinden. Für die Testbearbeitung standen fünf Minuten zur Verfügung. Bei einer Gruppe von 180 Bewerbern für den öffentlichen Dienst betrug Cronbach's alpha .71. (In der vorliegenden Untersuchung wurde dieser Test nicht auf Itemebene abgelocht, so daß Cronbach's alpha nicht bestimmt werden konnte). Hinweise auf die befriedigende psychometrische Qualität des Tests ergeben sich auch aus der Tatsache, daß für diesen Test bzw. für Vorläuferversionen in mehreren Bewährungskontrollen der Nachweis der Kriteriumsvalidität erbracht wurde (z.B. Graudenz, 1982; Kleinevoss, 1983; Seggebruch, 1982; Weber & Werner, 1983; Wolf, 1990).

12.2.3 Intelligenztests

Aus untersuchungstechnischen Gründen konnte nicht für alle Teilnehmer ein vollständig identisches Intelligenztestverfahren eingesetzt werden. Eine Erläuterung und Diskussion dieser Versuchsplanproblematik finden sich in Abschnitt 12.2.3.3, vorab sollen aber die beiden eingesetzten Verfahren dargestellt werden.

12.2.3.1 Berliner Intelligenzstruktur-Test

Zur Messung der Intelligenz wurde bei 61 % der Teilnehmer der Test zum Berliner Intelligenzstrukturmodell („BIS-4 Test", Jäger et al., 1997) eingesetzt. Der Test besteht aus 45 Aufgabentypen (siehe Tabelle 5), die mit dem Ziel einer möglichst guten Repräsentation aus einem ca. 2000 Typen umfassenden Inventar der bisher publizierten Intelligenz- und Kreativitätsaufgaben ausgewählt und adaptiert wurden. Der Test und seine psychometrische Güte ist bei Jäger et al. (1997) dokumentiert.

Tab. 5: Verteilung der 45 BIS-Aufgaben auf die Skalen und Zellen des BIS (nach Jäger et al.,1997)

	F	V	N
B	BD Buchstaben Durchstreichen OE Old English ZS Zahlen-Symbol-Test	KW Wörter Klassifizieren TG Teil-Ganzes UW Unvollständige Wörter	RZ Rechen-Zeichen SI Sieben-Teilbar XG X-Größer
M	WE Wege-Erinnern FP Figuren-Paare OG Orientierungs-Gedächtnis	PS Phantasiesprache ST Sinnvoller-Text WM Worte-Merken	ZP Zahlen-Paare ZW Zahlen Wiedererkennen ZZ Zweistellige Zahlen
E	OJ Objekt-Gestaltung ZF Zeichen-Fortsetzen LO Layout ZK Zeichen-Kombinieren	AM Anwendungs-Möglichkeiten EF Eigenschaften-Fähigkeiten MA Masselon IT Insight-Test	DR Divergentes Rechnen TN Telefon-Nummern ZG Zahlen-Gleichungen ZR Zahlenrätsel
K	AN Analogien (figural) AW Abwicklungen BG Bongard CH Charkow FA Figuren-Auswahl	WA Wortanalogien (verbal) TM Tatsache-Meinung SL Schlüsse SV Schlüsse-Vergleichen WS Wortschatz	BR Buchstabenreihen SC Schätzen TL Tabellen-Lesen RD Rechnerisches Denken ZN Zahlenreihen

Randspalten und -zeilen: Die BIS-Fähigkeiten:

„K" Verarbeitungskapazität
„E" Einfallsreichtum
„M" Merkfähigkeit
„B" Bearbeitungsgeschwindigkeit
„V" Sprachgebundenes Denken
„N" Zahlengebundenes Denken
„F" Anschauungsgebundenes, figural-bildhaftes Denken

(Erläuterungen zu den Fähigkeitskomponenten in Jäger, 1984)

12.2.3.2 Intelligenztest der DGP, ergänzt um Aufgaben aus dem BIS-Test

39% der Teilnehmer bearbeiteten einen unveröffentlichten Intelligenztest der *Deutschen Gesellschaft für Personalwesen e.V.* (*DGP*), der um 11 Aufgaben aus dem BIS-4 Test aufgestockt wurde. Der ursprüngliche *DGP*-Test umfasste Aufgaben, die in Termini des Berliner Intelligenzstrukturmodells („BIS", Jäger, 1982) die verbale (5), numerische (4) und figurale (1) Verarbeitungskapazität erfassen (in Klammern: Anzahl der entsprechenden Aufgaben) und die zum Teil typgleich mit den Aufgaben des BIS-4 Tests sind (siehe Tabelle 7). Tabelle 6 verzeichnet die Zuordnung der Aufgaben zu den BIS-Zellen, die Aufgabenbezeichnungen (samt) Abkürzungen sowie die Anzahl der Items und das jeweilige Cronbach's alpha. Die Berechnungen der internen Konsistenz wurden an einer Gruppe von Bewerbern für eine Ausbildung zum gehobenen Verwaltungsdienst vorgenommen, die bei Kersting (1996, S. 110) näher beschrieben ist.

Tab. 6: Verteilung der 10 *DGP*-Aufgaben zur Verarbeitungskapazität auf die BIS-Zellen; Erläuterung der Zellen: siehe Tabelle 5.
Die in der Tabelle berichteten Kennwerte beziehen sich auf eine bei Kersting (1996) dokumentierte Datenbasis

BIS-Zelle	Abkürzung und Bezeichnung		Items	N	Cronbach's α
KV	ÄW	Ähnliche Wortbedeutungen	20	353	.54
	AG	Analogien	23	353	.64
	SL	Schlüsse	20	353	.71
	TA	Textanalyse	18	353	.61
	WS	Wortschatz	25	260	.74
KN	ZZ	Zahlenmatrizen	15	353	.73
	TX	Textrechenaufgaben	17	353	.68
	ES	Ergebnisse Schätzen	18	353	.58
	TS	Tabellen und Statistiken	21	337	.58
KF	VB	Verschiedene Beziehungen	15	353	.55

Hinzu kamen zwei Aufgaben („Computerausdruck", abgekürzt „CA" und „Postaufgabe", abgekürzt „PA") zum Arbeitsverhalten mit numerischen Material, die der BIS-Zelle numerische Bearbeitungsgeschwindigkeit zugeordnet wurden.

Neben den hier für die Denkaufgaben im engeren Sinne berichteten Maßen der internen Konsistenz ergeben sich weitere Hinweise auf die psychometrische Qualität

des *DGP*-Tests aus der Tatsache, daß für die einzelnen Aufgaben in mehreren Bewährungskontrollen der Nachweis der Kriteriumsvalidität erbracht wurde (z.B. Bretz & Oldendörp, 1992; zu deutlichen Einschränkungen der Kriteriumsvalidität für die hier interessierende Berufsgruppe „Polizei" siehe aber: Wolf, 1990).

Um die Ähnlichkeit der beiden Testbatterien über die typgleichen Aufgaben hinaus noch weiter zu erhöhen, wurden zusätzlich zu diesen Aufgaben elf Subtests aus dem BIS-Test eingesetzt (siehe Tabelle 7). Mit acht der zusätzlich applizierten Aufgaben wurden die vom *DGP*-Test nicht abgedeckten BIS-Zellen mit jeweils einem Test repräsentiert. Außerdem wurde eine BIS-Aufgabe aufgenommen, die der BIS-Zelle BN zuzuordnen ist. Diese BIS-Zelle BN war im *DGP*-Test lediglich mit Aufgaben zum Arbeitsverhalten abgedeckt. Schließlich wurden zwei BIS-Aufgaben aufgenommen, die der Zelle KF zuzuordnen sind (siehe auch Tabelle 7). Die Gesamttestbatterie (*DGP*-Aufgaben zuzüglich den 11 Aufgaben aus dem BIS-4 Test) wird in der vorliegenden Arbeit mit der Bezeichung „*DGP* & Teil-BIS" abgekürzt.

12.2.3.3 Begründung und Diskussion des Einsatzes von zwei nicht vollständig identischen Intelligenztestverfahren

Daß nicht für alle Untersuchungsteilnehmer ein vollständig identisches Intelligenztestverfahren angewendet werden konnte, stellt ein untersuchungsplanerisches Problem der vorliegenden Studie dar. Dieses Problem ist dem „Feldcharakter" der Untersuchung geschuldet: Für diejenigen Probanden, die den Intelligenztest im Rahmen ihrer Aufstiegsbewerbung bearbeitet haben (siehe Abschnitt 12.3, dort als „Gruppe II" bezeichnet), mußte das Standardinstrument dieses Auswahlverfahrens - der *DGP*-Intelligenztest - verwendet werden. Lediglich eine Aufstockung der *DGP*-Testbatterie um 11 BIS-Aufgaben ließ sich erzielen. Weitere Gestaltungsmöglichkeiten hinsichtlich des bei dieser Gruppe eingesetzten Intelligenztests hatte der Autor der Studie aufgrund des vorgegebenen formalen/zeitlichen Rahmens nicht. Nur bei den übrigen Probanden („Gruppe I" in Tabelle 8) konnte das Intelligenz-Meßinstrument frei gewählt werden. Die naheliegende Planungssvariante, bei den übrigen Probanden nun ebenfalls den *DGP*-Test einzusetzen, ließ sich aus zeitlichen Gründen nicht realisieren. Die Gesamtuntersuchung mußte für die Gruppe I innerhalb eines Arbeitstages durchgeführt werden, da eine über einen Tag hinausgehende Freistellung der Beamten nicht erwirkt werden konnte. Dem Einsatz von zwei Problemlöseszenarien, dem Einsatz des Wissenstests sowie der Befragung nach Aspekten der Erfahrung im Umgang mit Computern sowie der Einstellung gegenüber der Arbeit mit Computern und der Akzeptanzbefragung wurde Priorität eingeräumt, da diese Gesichtspunkte in den bislang vorliegenden Studien vergleichsweise selten in dieser Systematik berücksichtigt wurden. Somit standen für die Messung

der Intelligenz nur noch ca. drei Stunden zur Verfügung. Aufgrund dieses Planungsansatzes war bei Gruppe I der Einsatz des *DGP*-Tests, dessen Bearbeitung mehr Zeit in Anspruch nimmt, nicht möglich. Der aus untersuchungsplanerischer Sicht nicht optimale Einsatz zweier nicht vollständig identischer Intelligenztestbatterien ist vor dem Hintergrund des teilweise vorliegenden Feldcharakters der hier thematisierten anwendungsbezogenen Forschungsstudie zu bewerten. Bei einer herkömmlichen Laboruntersuchung mit studentischen Versuchspersonen wäre diese Problematik nicht entstanden. Dieser Vorteil einer untersuchungsplanerisch weniger angreifbaren Laboruntersuchung wäre allerdings mit einer deutlichen Einschränkung der Relevanz der Studie für die diagnostische Praxis erkauft worden. Bei der Entscheidung für den Einsatz zweier nicht vollständig identischer Testverfahren wurde auch in Rechnung gestellt, daß bei Operationalisierungen von Konstrukten durch verschiedene, breit angelegte Intelligenzstrukturtests im Bereich der Intelligenzforschung eine höhere Übereinstimmung erzielt werden kann als bei den oft sehr speziellen Operationalisierungen von Konstrukten in anderen Forschungsbereichen. Jensen (1984, S. 570) berichtet eine durchschnittliche Interkorrelation von r=.77 (r=.86 nach Attenuationskorrektur) zwischen 30 verschiedenen Intelligenztests.

Dem Problem der beiden nicht vollständig identischen Testbatterien wurde in der Untersuchungsplanung vor allem durch die folgenden beiden Maßnahmen begegnet:

(1) Bei der Gruppe, bei der eine Wahlmöglichkeit hinsichtlich des Intelligenztests bestand, wurde ein Test ausgewählt, der a priori deutliche Gemeinsamkeiten mit dem für die übrige Gruppe „zwangsläufig" vorgegebenen *DGP*-Test aufweist.

(2) Um eine Angleichung der Messungen zu erzielen, wurde der „zwangsläufig" vorgegebene *DGP*-Test um elf Aufgaben aus dem BIS-Test ergänzt.

Diese beiden Maßnahmen führten dazu, daß ein Teil der Aufgaben (im Sinne einer Testhälfte) in beiden Gruppen gemeinsam eingesetzt wurde und somit die Parallelität der Messungen bestimmt werden konnte. Die erste Maßnahme, die Wahl des BIS-Tests für die Gruppe, bei der eine Wahlmöglichkeit hinsichtlich des eingesetzten Intelligenztests bestand, ist durch die deutlichen Gemeinsamkeiten zwischen dem BIS-Test und dem *DGP*-Test begründet[13]. Insbesondere sind hier die in beiden Tests enthaltenen fünf *typgleichen* Aufgaben (in Tabelle 7 unterstrichen) zu nennen. Hinsichtlich der Aufgaben der beiden Testbatterien, die keine direkte Typgleichheit

[13] Die Gemeinsamkeiten zwischen dem BIS-Test und dem *DGP* Test sind u.a. testentwicklungs-geschichtlich bedingt. A.O. Jäger war von 1955-1968 Leiter der *DGP* und hat dort sein Projekt „Dimensionen der Intelligenz" (Jäger, 1967) verwirklicht. Die in der DGP begonnenen und in Berlin weiterverfolgten Arbeiten haben zur Entwicklung des Berliner Intelligenzstrukturmodells und des dazugehörigen BIS-Tests geführt. Eine weitere Verbindung in der Testentwicklung der beiden Tests ist darin begründet, daß sowohl der BIS-4 Test als auch der hier thematisierte *DGP*-Test teilweise auf den WILDE-Test (Jäger & Althoff, 1994) aufbauen.

aufweisen, gilt es zu bedenken, daß Intelligenzaufgaben aus anderen Tests sich in das Berliner Intelligenzstrukturmodell (BIS) klassifizieren lassen. (Zur Invarianz des Modells über verschiedene Aufgabensätze hinweg siehe z.B. Jäger und Tesch-Römer, 1988; Schmidt, 1993). Der BIS-Test wurde u.a. eingesetzt, weil die Beziehungen zwischen den BIS-Testaufgaben und den *DGP*-Testaufgaben bereits mehrfach empirisch untersucht wurden. Schmidt (1986) hatte gezeigt, daß sich der *DGP*-Test in den Strukturrahmen des BIS integrieren läßt. Für Weiterentwicklungen von sieben der im hier analysierten *DGP*-Test enthaltenen Aufgaben (in Tabelle 7 mit einem Stern indiziert) konnten Kersting und Beauducel (1997) die hier vorgenommene Klassifikation in das BIS an einem Datensatz mit 3274 Personen replizieren.

Tab.7: Verteilung der Aufgaben aus den beiden Tests auf die Skalen und Zellen des BIS

AI	V		N		F	
	BIS	*DGP* & Teil-BIS	BIS	*DGP* & Teil-BIS	BIS	*DGP* & Teil-BIS
K	WA SL TM WS SV	AG* SL* ÄW TA* WS	RD SC TL ZN BR	TX* ES* TS* ZM*	***AW*** ***CH*** BG FA AN	***BIS_AW*** ***BIS_CH*** VB
E	***EF*** MA IT AM	***BIS_EF***	***TN*** DR ZG ZR	***BIS_TN***	***OJ*** ZF LO ZK	***BIS_OJ***
M	***PS*** WM ST	***BIS_PS***	***ZP*** ZZ ZW	***BIS_ZP***	***OG*** FM WE	***BIS_OG***
B	***KW*** TG UW	***BIS_KW***	***SI*** XG RZ	***BIS_SI*** CA PA	***BD*** OE ZS	***BIS_BD***

Kursiv- und Fettdruck: identische Aufgaben in beiden Tests;
Unterstreichung: typgleiche Aufgaben in beiden Tests;
Sternchen: Klassifikation empirisch bestätigt (Kersting & Beauducel, 1997)
Abkürzungen: siehe Tabellen 5 und 6 sowie Text

12.2.3.4 Skalenbildung und Parallelität der in beiden Teilgruppen eingesetzten Intelligenztests

Für beide Tests wurde eine Skala für die „Allgemeine Intelligenz (AI)" sowie für die „Verarbeitungskapazität (K)" gebildet. Die Beschränkung auf zwei Skalen entspricht dem Vorgehen der Auswertung der Kurzform im BIS-4 Test (siehe Jäger et al., 1997). Im *DGP*-Test stehen für diese beiden Skalen eine größere Zahl an unabhängigen Leistungsmessungen zur Verfügung als bei der Anwendung der Kurzform des BIS-4 Test. Analysen auf einem feineren Auflösungsgrad, d.h. auf Ebene der sieben BIS-Fähigkeiten wurden ausschließlich für die Teilnehmer durchgeführt, die den BIS-4 Test bearbeitet haben. Eine entsprechende Auswertung auf Skalenebene für die Gruppe, die den um 11 BIS-Aufgaben ergänzten *DGP*- Test bearbeitet hat, schien aufgrund der teilweise quantitativ zu schwachen Repräsentation der BIS-Zellen durch Aufgaben nicht vertretbar.

Um zu beurteilen, ob die beiden Testbatterien eine hinreichend große Ähnlichkeit bezüglich der Messung der „Allgemeinen Intelligenz" und der „Verarbeitungskapazität" gewährleisten, kann man sich zunächst den Umfang des in beiden Testbatterien gemeinsamen Anteils an Aufgaben vergegenwärtigen. Die elf BIS-4 Aufgaben, die bei beiden Gruppen eingesetzt wurden (Kursiv- und Fettdruck in Tabelle 7) und die fünf typgleichen Aufgaben (in Tabelle 7 unterstrichen) werden dabei als eine „Testhälfte" betrachtet, die *in beiden Gruppen* zur Anwendung gelangte. Lediglich hinsichtlich der anderen „Testhälfte" unterscheidet sich die für die beiden Gruppen vorgenommene Intelligenzmessung (weitere BIS-4 Aufgaben in der einen und *DGP*-Aufgaben in der anderen Gruppe). 16 der 45 Intelligenztestaufgaben der Gruppe I und 16 von 23 Intelligenztestaufgaben der Gruppe II (Gruppenbezeichnungen laut Tabelle 9) waren somit für beide Gruppen weitgehend identisch. Um einen Anhaltspunkt für die Parallelität der Fähigkeitsindikatoren zu erhalten, wurde nun dieser Anteil von 16 Aufgaben, der in beiden Gruppen weitgehend identisch war, mit denjenigen übrigen 29 Aufgaben der Gruppe I korreliert, die ausschließlich in Gruppe I eingesetzt wurden. Dazu wurden mit den 16 Aufgaben, die in beiden Gruppen eingesetzt wurden, Skalen für die „Allgemeine Intelligenz (AI)" und für die „Verarbeitungskapazität (K)" gebildet. Für die Gruppe I, die mit dem BIS-4 Test geprüft worden war, wurde außerdem aus dem verbleibenden Rest der Aufgaben (der übrigen „Testhälfte") ebenfalls eine Skala für die „Allgemeine Intelligenz" sowie für die „Verarbeitungskapazität" gebildet. (Beim *DGP*-Test war die verbleibende „Testhälfte" zu klein für eine solche Skalenbildung zum Zweck der Parallelitätsprüfung.) Die so gebildeten „Testhälften" wurden nun miteinander korreliert, über die Ergebnisse informiert Tabelle 8.

Die Korrelation der beiden „Testhälften" des BIS-4 Tests betrug nach Spearman-Brown Korrektur für die Skala „Allgemeine Intelligenz" $r_{korr}=.89$ und für die Skala „Verarbeitungskapazität" $r_{korr}=.90$. Der Wert für die Skala „Allgemeine Intelligenz" liegt nominell knapp unter, der Wert für die „Skala" Verarbeitungskapazität liegt nominell etwas über den Werten, die im Handbuch des BIS-4 Tests (Jäger et al., 1997) für Zufallsteilungen der Skalen berichtet werden. Somit liegt die Parallelität der beiden Testhälften für die Gruppe mit dem BIS-4 Test in einer Höhe, die angesichts der durch die Reliabilität gegebenen Begrenzungen als maximal mögliche Höhe gekennzeichnet werden kann. Die Ergebnisse dieser Analyse liefern einen deutlichen Hinweis darauf, daß die *in beiden Teilgruppen* gemeinsam eingesetzte „Testhälfte" eine hohe Parallelität mit dem BIS-4 Test aufweist und daher eine repräsentative Abbildung der „Allgemeinen Intelligenz" und der „Verarbeitungskapazität" im Sinne des BIS-Modells erlaubt. Aufgrund dieser Ergebnisse sowie aufgrund der hohen Ähnlichkeit zwischen den Aufgaben des BIS-4 Tests und des *DGP*-Tests erscheint es zulässig, die Ergebnisse der Studie für die Skalen „Allgemeine Intelligenz" und „Verarbeitungskapazität" auf Basis der Gesamtgruppe zu berechnen.

Tab. 8: Zur Parallelität der in beiden Gruppen gemeinsam eingesetzten „Testhälfte" mit der übrigen „Testhälfte" bei der mit dem BIS-4 Test geprüften Gruppe (N=63)

	AI	K
Einfache Korrelation	.80**	.82**
nach Spearman-Brown Korrektur	.89**	.90**

„AI" = Allgemeine Intelligenz,
„K" = Verarbeitungskapazität;
** $p<.01$; * $p<.05$;

Durch den teilweise gegebenen Feldcharakter der Untersuchung entstand das Problem, daß in den beiden Untersuchungsgruppen neben einem Anteil gemeinsamer Intelligenztestaufgaben auch ein Anteil unterschiedlicher Intelligenztestaufgaben vorgegeben wurde. Wie berichtet, wurde diesem Problem im wesentlichen durch zwei Maßnahmen begegnet. Einerseits wurden bei der Gruppe, bei der eine Wahlmöglichkeit hinsichtlich des verwendeten Intelligenztests bestand, ein Test ausgewählt, der deutliche Gemeinsamkeiten mit dem für die übrige Gruppe „zwangsläufig" vorgegebenen *DGP*-Test aufweist, und andererseits wurde der „zwangsläufig" vorgegebene *DGP*-Test zur Angleichung der Messungen um elf Aufgaben aus dem BIS-Test ergänzt. Diese Maßnahmen erlaubten die im vorherigen Abschnitt dargestellte Prüfung der Parallelität der Messungen über den Weg der „Testhalbierung". Als dritte Maßnahme wurde außerdem versucht, die Parallelität der Messungen direkt zu bestimmen, indem eine Subgruppe berücksichtigt wurde, die alle Intelligenzaufgaben

bearbeitet hatten. Leider konnten diese dritte Maßnahme nur für 18 der insgesamt 104 Teilnehmer realisiert werden. Die Zusammenhangsanalysen sprechen ebenfalls für eine ausreichende Parallelität der Messungen, die Aussagekraft der Befunde ist aber aufgrund der geringen Gruppengröße eingeschränkt[14].

Zur Absicherung gegen Methodenartefakte wurden die in der vorliegenden Arbeit berichteten relevanten Aussagen zu Außenbeziehungen der Intelligenzmaße zusätzlich auch für die „Testhälfte" berechnet, die in beiden Gruppen gemeinsam eingesetzt wurde. Wie zu erwarten, fielen diese Zusammenhangsmaße reliabilitätsbedingt geringer aus als die entsprechenden Maße für den Gesamttest, der relevante Aussagenbereich (Richtung und Signifikanz des Zusammenhangs) blieb aber zumeist unbeeinflußt. Beispielsweise verringerte sich der in Kapitel 15, Tabelle 14 berichtete Zusammenhang zwischen der Skala „Allgemeine Intelligenz" und der Steuerungsleistung in der „Schneiderwerkstatt" (neues Problemlösegütemaß) von r=.23 auf r=.20, wenn anstelle des Gesamttests nur diejenige Testhälfte berücksichtigt wurde, die in beiden Gruppen zum Einsatz kam. Die Treffsicherheit, mit der das Vorgesetztenurteil über die im Beruf gezeigten intellektuellen Leistungen durch die Skala „Allgemeine Intelligenz" vorhergesagt werden konnte (Abschnitt 17.3.1, Tabelle 25), verminderte sich von r=.43 auf r=.33, wenn die in beiden Gruppen unterschiedliche zweite Testhälfte bei der Auswertung unberücksichtigt blieb. Es ergaben sich aber keine inhaltlich vom Gesamtmaß abweichenden Befunde, so daß die in der Arbeit vorgenommene inhaltliche Interpretation der Befunde nicht durch den Hin-

---

[14] Um einen Anhaltspunkt für die Parallelität der mit den beiden unterschiedlichen Tests erhobenen Messungen zu gewinnen, wurde in der Studie eine Gruppe von 18 Personen berücksichtigt, die den *DGP*-Test bereits zu einem früheren Zeitpunkt („t1", im Durchschnitt 18 Monate vor der Untersuchung) bearbeitet hatten. Diese Gruppe bearbeitete in der vorliegenden Studie (Zeitpunkt „t2") den BIS-Test. Durch die zusätzliche wiederholte Vorgabe zweier *DGP*-Aufgaben („AG" und „VB", siehe Tabelle 6) konnte ein Indikator für die Stabilität des *DGP*-Tests ermittelt werden. Die Test-Retest Stabilität (berechnet als Korrelation der Aggregate der beiden Aufgaben zum Meßzeitpunkt „t1" und „t2") betrug r=.87. Die AI-Skalen der beiden Tests („BIS" und „*DGP ohne BIS-Aufgaben*") korrelierten zu r=.73 miteinander. Dieser Wert unterschätzt aus zwei Gründen die tatsächliche Parallelität der Messungen. Zum einen ist der Zusammenhang der Messungen in dem Maße eingeschränkt, in dem die beiden Tests von einer perfekten Retest-Reliabilität abweichen bzw. indem sich tatsächliche Merkmalsänderungen über die Zeit ergeben. Zum anderen ist der berichtete Wert ein Maß für den Zusammenhang zwischen den „ursprünglichen" *DGP*-Aufgaben und dem BIS-4 Test. In der vorliegenden Untersuchung wurden bei der Gruppe, die mit dem *DGP*-Test geprüft wurde, aber zusätzlich zu den „ursprünglichen" *DGP*-Aufgaben auch noch 11 Aufgaben aus dem BIS-4 Test appliziert. Dieser Umstand dürfte die Parallelität der Messungen vermutlich deutlich über das berichtete Maß hinaus erhöhen. Für ein Gedankenexperiment wurden einmal die 11 entsprechenden BIS-Aufgaben aus der Messung „t2" auch der Messung „t1" zugeschlagen. Dadurch erhöhte sich die Korrelation der beiden AI-Skalen auf r=.93. Dieser Wert ist aufgrund der doppelten Verrechnung der Aufgaben natürlich überschätzt. Die Parallelität der Messungen dürfte somit zwischen diesem unterschätzten Wert von .73 und dem überschätzten Wert von .93 liegen.

weis auf den Einsatz einer für zwei Gruppen unterschiedlichen zweiten Intelligenztesthälfte grundsätzlich in Frage gestellt werden kann.

12.2.4 Computererfahrung und Einstellung zur Arbeit mit Computern

12.2.4.1 Computererfahrung („CErfahr")

Mit einem neu konstruierten Fragebogen wurde die Qualität und Quantität des Ausmaßes der Computererfahrung erfaßt. Der Test umfaßte 35 Items in acht Frageblöcken (Cronbach's α der Gesamt-Skala = .91). Gefragt wurde beispielsweise nach der Vorerfahrung mit Computern und Softwaresystemen (z.B. Anzahl genutzter Programme, Programmiererfahrung), nach anwendungsbezogenem Wissen sowie nach der Häufigkeit, mit der Computer zu Arbeits- und Spielzwecken genutzt wurden. Für die Bearbeitung dieses Fragebogens - inklusive der im folgenden Abschnitt beschriebenen 10 Items zur Einstellung zur Arbeit mit Computern - standen 10 Minuten Zeit zur Verfügung.

12.2.4.2 Einstellung zur Arbeit mit Computern („CEin")

Im neu konstruierten Fragebogen „Einstellung zur Arbeit mit Computern" (10 Items, Cronbach's α =.61) ging es darum, ob ein Individuum eine eher positive oder eine eher zurückhaltende bis ablehnende Haltung zur der Arbeit mit Personalcomputern und gegenüber einer postulierten zunehmenden „Computerisierung" des Arbeitsplatzes einnimmt. Das Inventar bietet den Probanden pro Item vier Antwortmöglichkeiten, die von *„trifft gar nicht zu"* bis *„trifft genau zu"* reichen. Items, die eine tendenziell ängstliche, ablehnende Position zur Arbeit mit Computern schildern, wurden für die Datenanalyse umgepolt. (Beispielitem: *„Ohne Computer war vieles einfacher und funktionierte letztendlich genausogut oder sogar besser."*) Die so gemessene Einstellung zur Arbeit mit Computern war zu r=.45 mit der Computererfahrung (siehe Abschnitt 12.2.4.1) korreliert.

12.2.5 Weitere Instrumente

Mit einem kurzen Fragebogen („Fragen zur Steuerung") sollte erfaßt werden, wie gut die Teilnehmer mit der Aufgabe der Steuerung der beiden unterschiedlichen Problemlöseszenarien zurechtgekommen sind. Vorgegeben wurden neun Aussagen (z.B. *„Das Programm ist einfach zu bedienen"*). Die Teilnehmer sollten dann auf

einer sechsstufigen Skala ankreuzen, inwieweit sie den vorgegebenen Aussagen einmal in bezug auf die „Schneiderwerkstatt" und einmal in bezug auf „DISKo" zustimmen. Die interne Konsistenz, Cronbach's α, betrug für das Gesamturteil über die „Schneiderwerksatt" .81, über „DISKo" .88 (jeweils neun Items). Mit je einem Item (sechsstufige Likert-Skala) wurden die Teilnehmer darüber hinaus gefragt, ob die Unterschiedlichkeit der beiden Szenarien hinsichtlich (1.) der Bedienung und (2.) hinsichtlich der Inhalte (Größenordnungen, Variablenverknüpfungen) sich insofern ungünstig auswirkt, daß man sich bei der Steuerung des zweiten Programms zunächst umgewöhnen muß. Die abschließenden zwei Fragen „*Würden Sie bei so einer Computer-Problemlöseaufgabe gerne noch einmal mitmachen?*" und „*Haben Sie schon einmal eine ähnliche Aufgabe am Computer bearbeitet?*" konnten mit „ja" oder „nein" beantwortet werden.

Eingesetzt wurde außerdem ein Akzeptanzfragebogen mit dem erfaßt wurde, wie die Teilnehmer die beiden unterschiedlichen Meßinstrumente – Intelligenztests und Problemlöseszenarien – unter verschiedenen Aspekten (z.B. face-validity, Meßqualität, positives Erleben) beurteilen. Dieser Teil der Untersuchung ist an anderer Stelle dokumentiert (Kersting, 1998) und bleibt in der vorliegenden Arbeit ausgespart.

12.3 Untersuchungsdurchführung und -ablauf; Kontrolle der potentiellen Effekte unterschiedlicher Untersuchungsbedingungen sowie der Darbietungsabfolge

Tabelle 9 gibt den Ablauf der Untersuchung wieder. Die Variationen ergeben sich durch die Kontrolle der Darbietungsabfolge der Instrumente sowie durch die weiter oben (Abschnitt 12.1) bereits angesprochene Berücksichtigung einer Gruppe, die den Intelligenztest im Rahmen ihres Auswahlverfahrens für den Aufstieg zum höheren Polizeivollzugsdienst absolvierten. Während für diese Gruppe (Gruppe II in Tabelle 9) im Umfang von 40 Personen die einzelnen Untersuchungskomponenten an zwei aufeinanderfolgenden Tagen dargeboten wurden, fanden für die Gruppe I alle Erhebungen an einem Tag statt. Die Gruppe II bearbeitete im Anschluß an die Intelligenztests noch weitere Kenntnistests und nahm an Assessment-Center Übungen teil. Für die Gruppe II stellte sich lediglich der zweite Untersuchungsabschnitt (u.a. Problemlöseszenarien, systemspezifischer Wisssenstest und Akzeptanzbefragung) als Forschungsuntersuchung dar, der erste Teil der Untersuchung, der Intelligenztest, wurde unter „Ernstfallbedingungen" absolviert.

Daß die eine Gruppe der Teilnehmer den Intelligenztest in einer belastenden Personalauslesesituation absolviert hat, die andere Gruppe aber unter reinen Gefällig-

Tab. 9: Untersuchungsablauf

Variation der Darbietungsabfolge in einzelnen Gruppen		
Gruppe Ia (N=28)	Gruppe Ib (N=36)	Gruppe II (N=40)
Erfassung personenbezogener Daten		Erfassung personenbezogener Daten
Fragebogen zur PC-Erfahrung und PC-Einstellung	Intelligenztest (BIS)	Intelligenztest (*DGP* + 11 BIS-Aufgaben)
Zwei Problemlöseszenarien[1] (im Anschluß an die „SWS“: WIS-Test)	Kenntnistest Wirtschaft (DKT-W)	Kenntnistest Wirtschaft (DKT-W)
	Fragebogen zur PC-Erfahrung und PC-Einstellung	Weitere Personalauswahlverfahren (Kenntnistests, AC-Komponenten)
Fragen zur Programm-Steuerung	Zwei Problem löseszenarien[1] (im Anschluß an die „SWS“: WIS-Test)	Fragebogen zur PC-Erfahrung und PC-Einstellung
Intelligenztest (BIS)	Fragen zur Programm-Steuerung	Zwei Problemlöseszenarien[1] (im Anschluß an die „SWS“: WIS-Test)
Kenntnistest Wirtschaft (DKT-W)		Fragen zur Programm-Steuerung
Fragebogen zur Akzeptanz		Fragebogen zur Akzeptanz

SWS: Problemlöseszenario „Schneiderwerkstatt“
BIS: Test zum Berliner Intelligenzstrukturmodell; Version 4
DGP: Intelligenztest der *Deutschen Gesellschaft für Personalwesen e.V.*
1) Problemlöseszenarien: zunächst Einführung und Übungsphase
Darbietungsabfolge der Problemlöseszenarien variiert:
a) SWS, Wissenstest zur SWS (WIS), DISKo (bei 54 Teilnehmern)
b) DISKo, SWS, Wissenstest zur SWS (WIS) (bei 50 Teilnehmern)
Grau hinterlegte Felder: in vorliegender Arbeit nicht berücksichtigt

keitsbedingungen könnte - ebenso wie die Variation der zeitlichen Abfolge (Untersuchung an zwei aufeinanderfolgenden Tagen versus Untersuchung an einem Tag) und des Umfangs der Aufgaben (zusätzliche Kenntnistests und AC-Aufgaben bei Gruppe II) - Effekte auf die Leistung zeitigen. Um dies zu prüfen, wurden die z-transformierten Leistungen in den 11 Intelligenztestaufgaben aus dem BIS, die von beiden Gruppen bearbeitet wurden, zu einem Aggregat zusammengefaßt. Eine Variation der Leistung in dieser Skala in Abhängigkeit von den Untersuchungsbedin-

gungen konnte nicht festgestellt werden. Skeptiker müßten erwarten, daß die unter „Ernstfallbedingungen" getesteten Personen entweder aufgrund der hohen Motivation besonders gut oder aber aufgrund der hohen Belastung besonders schlecht abgeschnitten hätten. Dies war nicht der Fall. Diejenigen 40 Teilnehmer, die den Intelligenztest unter „Ernstfallbedingungen" absolvierten (Aufstiegsbewerber), unterschieden sich in ihren Leistungen ($\underline{M}=1.49$, $\underline{s}=.4.29$) nicht von denjenigen 20 Teilnehmern, die zum Zeitpunkt der Untersuchung bereits die Zulassung zum höheren Dienst erzielt hatten (Ratsanwärter) und den Test unter Gefälligkeitsbedingungen bearbeiteten ($\underline{M}=2.14$, $\underline{s}=4.75$) ($\underline{t}=-.52$; $\underline{p}=$n.s.).

In einem multiplen Mittelwertsvergleich mit Hilfe der Prozedur Oneway zeigte sich mittels Tukey-Test, daß die Intelligenzleistung erwartungsgemäß in Abhängigkeit von dem unterschiedlichen Berufserfolgspotential der Gruppen („Aufstiegspotential" in Tabelle 4) variierte. Sowohl die Gruppe der Ratsanwärter, die den Test unter Gefälligkeitsbedingungen bearbeitete als auch die Gruppe der Aufstiegsbewerber, die unter Ernstfallbedingungen mit dem Test konfrontiert war, übertraf leistungsmäßig die Gruppe der 43[15] Personen, die sich zum Zeitpunkt der Erhebung nicht in einem Aufstiegsverfahren befanden ($\underline{M}=-2.38$, $\underline{s}=4.89$) und den Test ebenfalls rein freiwillig bearbeiteten ($\underline{F}$ (2,100) = 9.85; $\underline{p}<.01$). Auch hinsichtlich der Indikatoren der Steuerungsleistung (siehe Tabelle 13) in den beiden Problemlöseszenarien ließen sich keine Effekte der pro Gruppe unterschiedlichen Rahmenbedingungen der Untersuchung nachweisen.

Durch die Variation der Szenarienabfolge wurden mögliche Effekte der Darbietungsabfolge experimentell kontrolliert. Die Steuerungsleistung in beiden Szenarien (Problemlösegütemaße laut Tabelle 13, siehe unten) blieb statistisch unbeeinflußt von der Darbietungsabfolge.

Alle Probanden nahmen freiwillig an der Untersuchung teil, für die Zeit der Teilnahme an der Untersuchung gewährte die zuständige Behörde dienstfrei. Eine materielle Vergütung der Untersuchungsteilnahme bzw. ein materieller Leistungsanreiz bestand nicht. Den Teilnehmern wurde eine schriftliche Rückmeldung mit einem ipsativen Profil ihrer Leistungen im Intelligenztest sowie einer Rückmeldung über ihr Abschneiden in den beiden Problemlöseszenarien versprochen und im Durchschnitt ein Jahr nach der Untersuchung zugesandt. Mit dieser Rückmeldung erhielten die Teilnehmer einen Fragebogen, der überwiegend der Erhebung von Kennwerten für eine prädiktive Kriteriumsvalidierung aus der Sicht der Selbstbeurteilung diente, in dem aber auch einige Items der ursprünglichen Akzeptanzbefragung zum Akzeptanzaspekt „face validity" erneut vorgegeben wurden (siehe

[15] Für eine Person dieser Gruppe lagen keine Intelligenztestdaten vor, siehe Abschnitt 12.4.

Kersting, 1998). Die prädiktive Kriteriumsvalidierung aus Sicht der Teilnehmer ist nicht Gegenstand der vorliegenden Arbeit.

Die Erhebung wurde zu sieben Terminen in Gruppen von sieben bis maximal 20 Probanden durchgeführt. Bei größeren Gruppen wurde die Bearbeitung der Problemlöseszenarien parallel in zwei Räumen durchgeführt. Im März 1993 wurden die ersten Probanden, im Juni 1994 die letzten Probanden untersucht. Die Datenerhebung fand in Unterrichts- und Computerräumen der Polizei in Düsseldorf, Eutin, Hannover, Hannoversch-Münden und Neuss statt und wurden von erfahrenen Diplom-Psychologen oder psychologisch-technischen Assistenten durchgeführt. Bei der Durchführung der Problemlöseszenarien wurde aus Gründen höherer Standardisierung lediglich zwei Versuchsleiter berücksichtigt.

12.4 Datenausfälle

Für vier Probanden konnten die Ergebnisse der Bearbeitung des Szenarios „DISKo" nicht berücksichtigt werden, da diese Personen in den 50 Bearbeitungsminuten keinen bzw. nur einen Entscheidungstakt absolvierten (siehe Abschnitt 13.2.2).

Ein Proband aus der Gruppe Ia (laut Tabelle 9) versäumte den zweiten Teil der Untersuchung, der die Bearbeitung des Intelligenztests und des allgemeinem Wirtschaftswissenstests sowie die Akzeptanzbefragung vorsah, so daß für diese Variablen lediglich Daten von 103 Personen vorlagen.

12.5 Auswertungsmethoden

Die Datenanalyse folgte dem statistischen Entscheidungsmodell, wobei ein Signifikanzniveau von $\alpha = .05$ festgelegt wurde. Sofern keine weiteren Angaben getroffen werden, bezieht sich die Angabe der Irrtumswahrscheinlichkeit auf zweiseitige Tests, Irrtumswahrscheinlichkeiten für gerichtete Hypothesen werden als solche expliziert. Alle verwendeten Variablen wurden zunächst mit Hilfe des Kolmogorov-Smirnov-Tests auf Normalverteilung geprüft. Entsprechend der Empfehlung von Bortz (1989, S. 198 f.) wurde der α-Fehler beim Kolmogorov-Smirnov-Test auf das 25%-Niveau gesetzt. Die Variablen für die Steuerungsleistungen in den beiden Problemlöseszenarien und die bei dem Szenario „DISKo" programmintern berechneten Verhaltensmaße für die Subbereiche „Wissen", „Entscheidungen" und für das Gesamtmaß „Verhalten" sowie der Index des allgemeinen Wirtschaftswissens und die

Variable für die Computererfahrung und das Alter erfüllten diese Voraussetzung nicht und wurden daher normalisiert. Die übrigen Variablen waren gemäß des vorgenommenen Kolmogorov-Smirnov-Tests normalverteilt. Die Berechnungen wurden überwiegend mit parametrischen Verfahren durchgeführt, Abweichungen von dieser Regel werden benannt.

Bei allen im folgenden dargestellten Befunden gilt, daß die Voraussetzungen der jeweiligen Analysen wie Normalität, Linearität, univariate und/oder multivariate Varianzhomogenität, Abwesenheit von Multikollinearität etc. grundsätzlich geprüft wurden. Die entsprechenden Prüfungsergebnisse werden aber nur dann thematisiert, wenn die Voraussetzungen *nicht* erfüllt waren, d.h. falls die entsprechenden Prüftests signifikante Ergebnisse zeigten.

13. Problemlösegütemaße

13.1 Analysen zur Schwierigkeit der Szenarien

Beide Szenarien wurden von den Testanden mit dem Ziel gesteuert, das Gesamtvermögen am Ende der Bearbeitungszeit auf das höchstmögliche Niveau zu maximieren. Die Steuerungsleistung ist somit am Gesamtvermögen am Ende der Bearbeitungszeit zu messen. In Kapitel 7 der vorliegenden Arbeit wurde aber hervorgehoben, daß die interne Validität einzelner Problemlösegütemaße und die Verwendbarkeit der Steuerungsleistungen als diagnostische Information u.a. von der Schwierigkeit der zu steuernden Szenarien beeinflußt werden kann. Dem Vorgehen der Aufgabenanalyse bei der Berliner Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen entsprechend wurde daher zunächst geprüft, ob die Voraussetzungen für die Umsetzung des vorgegebenen Ziels überhaupt gegeben waren, d.h. ob die beiden Szenarien einen der Untersuchungsgruppe angemessenen Schwierigkeitsgrad aufwiesen.

Tabelle 10 gibt einige deskriptive Kennwerte zur Zielerreichung in den beiden Szenarien wieder, Abbildung 5 zeigt für beide Szenarien die Entwicklung des Gesamtvermögens über die Bearbeitungstakte hinweg. Während in der „Schneiderwerkstatt" ein akzeptabler Anteil der Testanden das Ziel der Steuerung erreichte, gelang es nur 17% der Teilnehmer im Szenario „DISKo" das Gesamtvermögen zu steigern. 38% der Teilnehmer erwirtschafteten in keinem

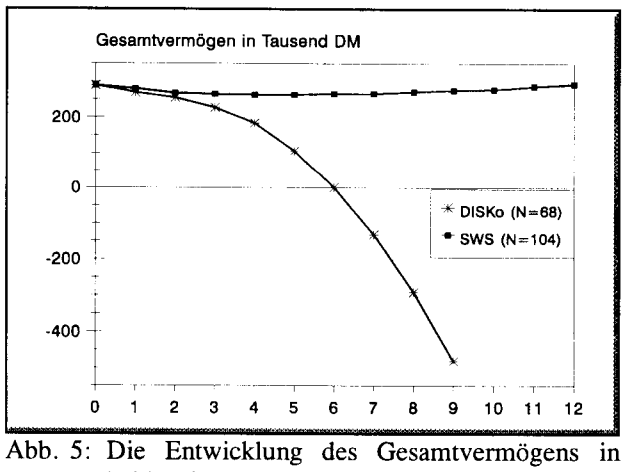


Abb. 5: Die Entwicklung des Gesamtvermögens in beiden Szenarien
(Zu Zwecken der Vereinheitlichung der Darstellung wurde zu allen „DISKo"-Werten der Betrag von 109925 hinzuaddiert. Die Gründe für die geringere Gruppengröße und geringere Anzahl an Bearbeitungstakten bei „DISKo" sind in Abschnitt 13.2.2 erläutert.)

einzigen „DISKo"-Bearbeitungstakt Gewinn, nur 19% gelang dies für mehr als der Hälfte der Bearbeitungstakte. Aus diagnostischer Perspektive besonders hervorzuheben ist der Vergleich der von den Teilnehmern erzielten Ergebnisse im Gesamtvermögen mit dem sogenannten „Null-Lauf": 83 % der Testanden haben durch ihre Eingriffe im Szenario „DISKo" ein schlechteres Ergebnis erzielt als wenn sie für 12 Takte die Programmvoreinstellung ohne irgendeine eigene Entscheidung bestätigt hätten. (Zur damit gegebenen Möglichkeit der Ergebnisverfälschung: Kapitel 10.2.)

Tab. 10: Deskriptive Kennwerte zur Zielerreichung in den beiden Szenarien

Prozentualer Anteil der Teilnehmer, ...	„Schneiderwerkstatt"	„DISKo"
die insgesamt Gewinn erwirtschaften konnten	61,5 %	17 %
die keinen Bearbeitungstakt mit Gewinn verzeichneten	6,7 %	38 %
die ein besseres Ergebnis erzielten als der Null-Lauf	96,2 %	17 %
mit Gewinn in mehr als d. Hälfte d. Bearbeitungstakte	63,5 %	19 %

Die deskriptiven Kennwerte weisen nach, daß die Steuerung des Szenarios „DISKo" für die Untersuchungsgruppe sehr schwer war. Die in Kapitel 7 dargestellte Aufgabenanalyse für eine frühere Version der „Schneiderwerkstatt" hatte ergeben, daß unter den Umständen eines zu schwer steuerbaren Programms das „Gesamtvermögen" kein intern valider Indikator der Steuerungsleistung war. Diese Aufgabenanalyse wurde nun auf das Szenario „DISKo" angewandt.

13.2 Aufgabenanalyse Szenario „DISKo"

13.2.1 Analyse der Gewinnspanne; Definition eines neuen Problemlösegütemaßes

Die in Kapitel 7 beschriebenen Annahmen der Aufgabenanalyse für die „Schneiderwerkstatt" gelten insofern für das Szenario „DISKo" als auch hier die Hauptgewinnmöglichkeit im Verkauf der Produkte (bei „DISKo": Computerchips) liegt. Die in diesem Programm zusätzlich möglichen Erlöse durch Patente, Forschungsaufträge und Recyling können demgegenüber vernachlässigt werden. Wie in Kapitel 7 erläutert, ist der Gewinn das *Produkt* aus der Anzahl verkaufter Chips und der Gewinnspanne pro verkauftem Chip. Es galt also zunächst zu prüfen, ob die Teilnehmer das Szenario „DISKo" in den Bereich der positiven Gewinnspanne steuern konnten. Dies war überwiegend nicht der Fall.

Abbildung 6 zeigt, daß es selbst im diesbezüglich besten Bearbeitungstakt weniger als 40 % der Testanden gelungen ist, das System zumindest kurzfristig in die Zone mit positiver Gewinnspanne zu steuern. (Zum Vergleich sind auch die entsprechenden Werte für die „Schneiderwerkstatt" in der Graphik abgetragen.) Durch die Beschränkung der Graphik auf die Daten der 68 Personen, die mindestens neun Bearbeitungstakte vorzuweisen hatten (siehe Abschnitt 13.2.2), wird die Bedeutung der negativen Gewinnspanne für die Testanden sogar noch unterschätzt. Bezogen auf den jeweils letzten Entscheidungsdurchgang und die Gesamtgruppe konnten sogar nur 19% eine positive Gewinnspanne aufweisen. Die Steuerung des Szenarios war offensichtlich zu schwer. Dieser Befund ist in seiner Geltung natürlich zunächst auf die analysierte Untersuchungsgruppe beschränkt. Allerdings erzielte der Proband, der dem Median der Gruppe am nächsten lag, mit seinem Endergebnis im Gesamtvermögen in Höhe von -549.734 DM nach der dem Programm beiliegenden Norm für berufstätige Ingenieure und Naturwissenschaftler noch einen Prozentrang von 33. Demzufolge war die Untersuchungsgruppe etwas, aber nicht deutlich leistungsschwächer als andere Gruppen. Es wäre interessant zu prüfen, ob die geschilderte Überforderung und die daran gebundene Konsequenz der mangelnden internen Validität (siehe unten sowie Kapitel 7) der am „Gesamtvermögen" orientierten Indikatoren der Steuerungsleistung sich auch bei anderer Untersuchungen zeigt.

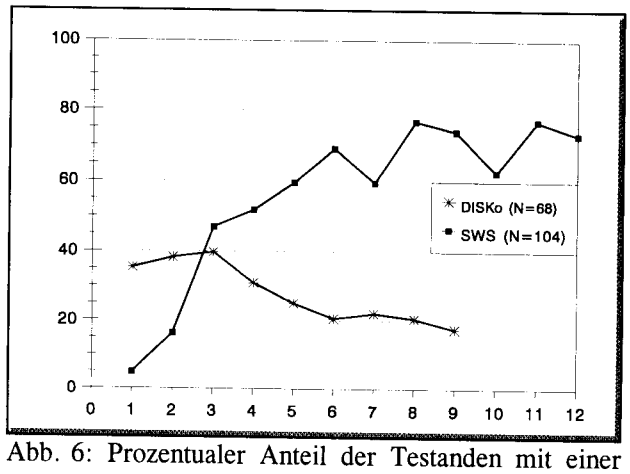


Abb. 6: Prozentualer Anteil der Testanden mit einer positiven Gewinnspanne in beiden Szenarien (Die Gründe für die geringere Gruppengröße und geringere Anzahl an Bearbeitungstakten bei „DISKo" sind in Abschnitt 13.2.2 erläutert.)

Solange die Gewinnspanne aber - wie für die überwiegende Mehrheit der Untersuchungsgruppe - negativ ist, erwirtschaften die Probanden mit jedem verkauften Chip Verluste. Wie im Theorieteil in Abbildung 4 veranschaulicht, unterscheidet das Problemlösegütemaß „Gesamtvermögen" unter diesen Umständen nicht mehr zwischen einem relativ guten Problemlöser (nur geringfügig negative Gewinnspanne, große Anzahl an verkauften Chips) und einem schlechten Problemlöser (deutlich negative Gewinnspanne, geringe Anzahl an verkauften Chips), weil die beiden Subziele „Gewinnspanne pro Chip" und „Anzahl verkaufter Chips" zur Be-

stimmung des Gesamtvermögens *multipliziert* werden und dabei das negative Vorzeichen der Gewinnspanne den ursprünglich Qualitätsnachweis eines guten Problemlösers - nämlich seine guten Verkaufszahlen - konterkariert. Die unter den spezifischen Bedingungen gegebene Inkompatibilität der Subziele konnten die Teilnehmer kaum erschließen, da der Wert für die Gewinnspanne pro Chip weder angezeigt wurde noch ohne weiteres selbst errechnet werden konnte.

Analog zu dem Vorgehen in der Berliner Untersuchung mit einer früheren Version der „Schneiderwerkstatt" (siehe oben, Abschnitt 7.3.2) wurde nun für „DISKo" ein neues Problemlösegütemaß definiert. Zunächst wurde die Kennwerte für die beiden Subziele „Anzahl an verkauften Chips" und „Gewinnspanne pro Chip" gebildet, wobei pro Teilnehmer die jeweils maximale Anzahl an Entscheidungsdurchgängen berücksichtigt wurde. Aus den beiden Teilgütemaßen wurde dann durch einfache *additive* Aggregation das neue Problemlösegütemaß gebildet, die Leistungen in den beiden Subgütekriterien wurden vorab z-transformiert. Aufgrund der schiefen Verteilung wurde das neue Problemlösegütemaß normalisiert. Das neue Problemlösegütemaß korrelierte zu $r=.40$ mit dem Kapitalendwert ($N=100$). Ein erster Hinweis auf die fehlende interne Validität des Indikators „Gesamtvermögen" und auf die interne Validität des neuen Problemlösegütemaßes ergibt sich aus den Interkorrelationen der Maße mit dem bei „DISKo" programmintern berechneten Index zur Beurteilung der von den Testanden gezeigten Verhaltensweisen und Strategien. Dieses Maß korrelierte zu .19 (n.s) mit der Variable „Kapitalendwert", aber zu .34 ($p<.01$) mit dem neuen Problemlösegütemaß ($N=100$).

Zu Vergleichszwecken wurde das neue Problemlösegütemaß auch für die „Schneiderwerkstatt" gebildet.

13.2.2 Weitere Probleme der Systemsteuerung; zusätzliche Variante des neu definierten Problemlösegütemaßes

Im Rahmen der Analyse der „DISKo"-Daten fiel auf, daß die Anzahl der „DISKo"-Bearbeitungstakte mit Entscheidungen über die Teilnehmer hinweg variierte (siehe Tabelle 11). Während einige Teilnehmer in den 50 Minuten keinen einzigen oder nur eine Entscheidungsdurchgang „spielten" (und von der Analyse ausgeschlossen werden mußten, siehe Abschnitt 12.4) brachten es andere in der gleichen Zeit auf bis zu 25 Bearbeitungstakte. Bei „DISKo" wird dem Probanden kein festes Ablaufschema vorgegeben, dies wird im Handbuch explizit als Vorzug des Szenarios herausgestellt (U. Funke, 1992a, S. 2-4). Diese Erweiterung der Verhaltensmöglichkeit wird mit Einbußen in der Standardisierung erkauft. Für die einzelnen Teilnehmer liegen in einem größeren Ausmaß quantitativ und qualitativ unterschiedliche Daten

vor als dies bei Problemlöseszenarien mit vorgegebenem Ablaufschema der Fall ist. Um hinsichtlich der Steuerungsleistung eine zwischen den Individuen und den beiden Szenarien vergleichbarere Datenbasis zu erhalten, wurden die Teilnehmer der vorliegenden Untersuchung explizit instruiert, 12 Entscheidungsdurchgänge zu absolvieren. Daß diese Instruktion nur von 18 Teilnehmern befolgt wurde, muß nicht als Indiz mangelnder compliance gewertet werden. Im Szenario „DISKo" werden die Teilnehmer standardmäßig nicht darüber informiert, in welchem „Entscheidungsmonat" ihre „Simulation" sich befindet. Lediglich ein spezielles Untermenü (graphische Verlaufsanalyse der Entscheidungen) erlaubt es, diesbezügliche Informationen zu generieren. Um diese Information aufzurufen, muß man sich durch verschiedene Pull-down Menüs wählen und insgesamt sechsmal die „Enter" Taste drücken. Kurzum, es liegt nahe zu vermuten, daß einige Teilnehmer nicht wußten, in welchem Entscheidungsmonat sie sich befanden. Das schlichte „Mitzählen" der einzelnen Entscheidungen wird dadurch erschwert, daß das Abschließen eines Entscheidungsdurchgangs große Ähnlichkeit mit dem Abschließen eines Testlaufs hat. In dem „Fragebogen zur Steuerung" (siehe oben, 12.2.5) äußerten insgesamt 86.4% der Teilnehmer, daß die Aussage *„Es fällt schwer, den Überblick zu bewahren"* etwas (18.4%), überwiegend (39.8%) oder genau (28.2%) auf die Bedienung des Szenarios „DISKo" zutrifft. (Zum Vergleich: hinsichtlich der „Schneiderwerkstatt" äußerten nur 17.5% der Teilnehmer entsprechende Bedenken ($t=11.58$; $p<.01$)) Das Befragunsergebnis mag zum Teil die Frustration der Teilnehmer über das eigene schlechte Abschneiden widerspiegeln, zum anderen Teil dürfte es aber auch auf wahrgenommene Bedienungsprobleme zurückzuführen sein.

Tab. 11: Bearbeitungstakte mit Entscheidungen in „DISKo"

↘ *Anzahl der „DISKo" Bearbeitungstakte mit „Entscheidungen"*												
0[1]	*1*[1]	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*
2	2	1	9	1	3	5	5	8	9	3	11	18
Häufigkeit ↗												
↘ *Anzahl der „DISKo" Bearbeitungstakte mit „Entscheidungen"*												
13	*14*	*15*	*16*	*17*	*18*	*19*	*20*	*21*	*22*	*23*	*24*	*25*
4	5	6	4	4	---	1	1	---	---	---	1	1
Häufigkeit ↗												

1) Teilnehmer von der weiteren Analyse ausgeschlossen

Um eine konstante Größe der Datenbasis zu sichern, wurden in den Abbildungen 5 und 6 nur die ersten neun Entscheidungsdurchgänge derjenigen 68 Probanden berücksichtigt, die mindestens neun Bearbeitungstakte vorzuweisen hatten.

Die Aufgabenanalyse förderte noch ein weiteres Problem der Steuerung des Szenarios „DISKo“ zu Tage: Die Anzahl der Entscheidungstakte war an einen negativen Steuerungserfolg in der Variable „Gesamtvermögen“ geknüpft, die beiden entsprechenden normalisierten Variablen korrelierten zu -.37 miteinander ($N=100$; $p<.01$). Verantwortlich für diesen Zusammenhang könnte das Simulations-Ereignis „Maschinenschäden“ gewesen sein. Bei fast der Hälfte der Probanden (48%) nahmen die Maschinenschäden im Laufe der Zeit kurz- oder längerfristig ein Ausmaß von über 95% an. Rund ⅓ der Teilnehmer (34%) hatten genau oder mehr als ⅓ ihrer Entscheidungsdurchgänge mit mindestens 95 % Maschinenschäden verbracht. Nur vier Personen ist es überhaupt gelungen, einen Maschinenschaden in dieser Größenordnung wieder zu „beheben“. Für die übrigen Teilnehmer mit diesem Ereignis galt: einmal Maschinenschaden, immer Maschinenschaden. Mit über 95% Maschinenschäden ist es kaum möglich zu produzieren und Gewinn zu erzielen, entsprechende Steuerungsindikatoren bewerten das Verhalten der Teilnehmer unter diesen Umständen nicht zutreffend. Hinzu kommt, daß das System „DISKo“ in der vorliegenden Version nicht immer sonderlich plausibel auf die Maßnahmen reagierte, die die Teilnehmer ergriffen, um dem durchaus registrierten Umstand der Maschinenschäden entgegenzusteuern. So blieb der Parameter „Maschinenschäden“ z.B. teilweise unbeeinflußt von den Gegenmaßnahmen „Maschinenverkauf“, „Einkauf neuer Maschinen“ und „Erhöhung der Instandhaltungskosten“ (auch bei einer Erhöhung der Instandhaltungskosten auf 100.000 DM blieb es z.B. noch über „Monate“ beim vollständigen Maschinenschaden). Ein Einzelfall soll dies illustrieren. Teilnehmer 5 reagierte auf die Variablenausprägung „99% Maschinenschäden“ mit dem Verkauf sämtlicher „Vollautomaten“ und kaufte - nach seiner Lesart - „neue“ und somit kaum schadensbehaftete „Halbautomaten“. Als Reaktion *stieg* die Variable „Maschinenschäden“ der „nagelneuen“ Halbautomaten von „99“ auf „100 %“. Das „vernünftige“ Verhalten des Teilnehmers lohnte sich nicht, er kam aus der einmal verfahrenen Situation nicht mehr heraus. Am Parameter „Gewinn“ orientierte Indikatoren der Steuerungsleistungen würdigten die vernünftige Handlung des Problemlösers nicht. Überspitzt kann man formulieren, daß sich das System „DISKo“ nach dem Ereignis „Maschinenschäden über 95%“ in den meisten Fällen der Steuerung entzog und somit die Steuerungsleistung für die von diesem Ereignis betroffenen Probanden nicht mehr beurteilt werden konnte. Die Kovariation zwischen der Anzahl der Entscheidungstakte und dem Ausmaß der Verluste mag sich dadurch erklären, daß die Wahrscheinlichkeit von Maschinenschäden im Umfang von über 95% mit zunehmender Anzahl an Entscheidungstakten wächst. Nur bei einem Probanden trat ein solches Ereignis bereits im vierten Bearbeitungstakt auf, 99 Probanden blieben in den ersten vier Takten von Maschinenschäden dieses Ausmaßes verschont.

Tabelle 12 zeigt die Verteilung des ersten Auftretens des Ereignisses „Maschinenschäden von mindestens 95%“ über die einzelnen Bearbeitungstakte.

Tab. 12: Verteilung des ersten Auftretens des Ereignisses „Maschinenschäden von mindestens 95%“ über die einzelnen „DISKo“ Bearbeitungstakte

↘ *erster „DISKo“ Bearbeitungstakt mit „Maschinenschäden in Höhe von min. 95%“*															
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*
---	---	---	1	5	14	12	6	4	2	1	1	---	1	---	1
Häufigkeit ↗															

Aus der Analyse ergibt sich die These, daß an der Zielvorgabe Gewinnmaximierung und somit auch an den Subzielen „Verkauf“ und „Gewinnspanne“ orientierte Indikatoren der Steuerungsleistung nur solange die Leistungen der Teilnehmer in dem Szenario richtig wiederspiegeln, solange das System aus der subjektiven Sicht der Teilnehmer halbwegs „steuerbar“ ist. Mit zunehmender Anzahl an Bearbeitungstakten entzog sich das System „DISKo“ (angesichts der zunehmenden Wahrscheinlichkeit eines fatalen Ausmaßes an Maschinenschäden) für viele Teilnehmer der Steuerung. Zusätzlich zu den Indikatoren der Steuerungsleistung über alle Bearbeitungstakte wurde daher das neue Problemlösegütemaß (siehe oben Abschnitt 13.2.1) *über die ersten acht Bearbeitungstakte* berechnet. Die Wahl fiel auf den achten Bearbeitungstakt, da zu diesem Simulationszeitpunkt einerseits genügend Entscheidungen für eine Beurteilung der Steuerungsleistung vorliegen, andererseits aber die Anzahl der Entscheidungstakte unter desaströsen Umständen (mindestens 95% Maschinenschäden) zumeist nicht größer ist als die Anzahl der „steuerbaren“ Takte. Die Daten aus den Entscheidungen 9-25 wurden bei diesem Maß ignoriert. Bei den 24 Teilnehmern, die weniger als acht Entscheidungen getroffen haben, wurde zur Bildung des Indikators die maximal mögliche Anzahl an Bearbeitungstakten einbezogen. Das auf die ersten acht Bearbeitungstakte beschränkte neue Problemlösegütemaß war zu $r=.09$ mit dem Kapitalendwert und zu $r=.80$ mit dem über alle Bearbeitungstakte berechneten neuen Problemlösegüemaß korreliert.

13.3 Überblick über die Indikatoren der Steuerungsleistung

Tabelle 13 gibt einen Überblick über die Indikatoren der Steuerungsleistungen und die verwendeten Abkürzungen, die in der Arbeit dargestellten Analysen beziehen sich auf die normalisierten Variablen.

Tab. 13: Indikatoren der Steuerungsleistung in beiden Szenarien

	alle Bearbeitungstakte		nur die ersten 8 Bearbeitungstakte
	„Schneiderwerkstatt"	„DISKo"	„DISKo"
Gesamtvermögen (Endwert)	***SWS-Gekap***	***DISKo-Gekap***	
neues Problemlösegütemaß	SWS-PLG	DISKo-PLG	DISKo-PLG8

Es gilt explizit festzuhalten, daß zur Beurteilung der diagnostischen Alltagstauglichkeit der Szenarien lediglich die Indikatoren der Steuerungsleistung herangezogen werden können, die mit der Zielvorgabe an die Probanden übereinstimmen und standardmäßig vom Programm berechnet werden (Endwert des Gesamtvermögens, in Tabelle 13 hervorgehoben). Aus diesem Grund beziehen sich die nachfolgenden Analysen primär auf diese Indikatoren. Die hier erarbeiteten Aufgabenanalysen und die Neuberechnung von intern validen Problemlösegütemaßen für das Szenario „DISKo" können in der Praxis vom diagnostischen Anwender kaum geleistet werden. Mit diesen zusätzlichen Indikatoren berechnete Analysen werden hier lediglich zu Forschungszwecken durchgeführt.

Die vorliegende Arbeit konzentriert sich - wie in Kapitel 6.4 angekündigt und erläutert - auf die in jedem Fall vorgeordnete Auswertung von Steuerungsleistungen. Die bei „DISKo" berechneten Indices der Verhaltensweisen und Strategien werden daher nur sporadisch berücksichtigt. Die in Kapitel 6.3 beschriebenen theoretischen Defizite der Verhaltensmaße und die zu ihrer Umsetzung notwendige willkürliche Datenreduktion sprechen gegen eine prioritäre Berücksichtigung dieser Indikatoren in der angewandten Diagnostik.

13.4 Zusammenfassung und Diskussion

Die deskriptiven Kennwerte weisen nach, daß die Steuerung des Szenarios „DISKo" für die Untersuchungsgruppe (zu) schwer war. Die in der Berliner Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen erarbeiteten Befunde für eine frühere Version des Systems „Schneiderwerkstatt" konnten für das System „DISKo" repliziert werden: Das vorgegebene Ziel der Maximierung des Gesamtvermögens läßt sich auch bei „DISKo" in die Subziele „Verkauf" und „Gewinnspanne pro Verkaufseinheit" zerlegen. Solange die Gewinnspanne - wie für die überwiegende Mehrheit der Untersuchungsgruppe - negativ ist, erwirtschaften die Probanden mit jedem verkauften Chip Verlust. Die bei der Bestimmung des Gesamtvermögens vorgenommene multiplikative Verrechnung der beiden Teilziele verhindert unter diesen Umständen eine Differenzierung zwischen Probanden mit einer positiven und einer negativen Steuerungsleistung. Als neuer Indikator wurden daher die z-transformierten Leistungen in den beiden Subzielen „Verkauf" und „Gewinnspanne pro Verkaufseinheit" *additiv* zu einem neuen Problemlösegütemaß zusammengefaßt. Dieser Indikator wurde zu Vergleichszwecken auch für das Szenario „Schneiderwerkstatt" berechnet, obgleich sich bei dieser Aufgabe keine derartigen Steuerungsprobleme ergaben.

Als weiteres Ergebnis der Aufgabenanalyse für „DISKo" konnte festgehalten werden, daß die Anzahl der Entscheidungstakte an einen negativen Steuerungserfolg in der Variable „Gesamtvermögen" geknüpft war. Verantwortlich für diesen Zusammenhang war vermutlich das - de facto fast unwiderrufliche - Simulations-Ereignis „desaströse Maschinenschäden", dessen Auftretenswahrscheinlichkeit mit zunehmender Anzahl an Entscheidungen stieg. Zusätzlich wurde daher eine Variante des neuen Indikators der Steuerungsleistung berechnet, bei dem lediglich die ersten acht - vergleichsweise weniger von dem Ereignis „Maschinenschäden" betroffenen - Bearbeitungstakte berücksichtigt wurden.

14. Prüfung der Voraussetzungsfreiheit der Steuerungsleistung

Im Theorieteil der vorliegenden Arbeit wurde ausgeführt, daß die Steuerungsleistung multipel bedingt sein kann. Neben der diagnostisch interessierenden Fähigkeit, die für die Systemsteuerung verantwortlich ist, kann möglicherweise auch die Computererfahrung und die Einstellung zur Arbeit mit Computern zum Steuerungsergebnis beitragen (siehe die Abschnitte 10.1.2 und 10.1.3). Falls die Computererfahrung und -einstellung gruppenspezifisch (z.B. alters- und geschlechtsspezifisch) variiert, könnten aus solchen Effekten aus einer bestimmten Fairneßperspektive Vorbehalte gegenüber dem diagnostischen Einsatz von Problemlöseszenarien erwachsen (siehe Abschnitt 10.1.1). Auch Vorwissenseffekte wurden in Erwägung gezogen (siehe Abschnitt 2.3.2.2). Wird durch die inhaltlichen Aufgabenmerkmale eines computergestützten Problemlöseszenarios Wissen aktiviert, welches in der Gruppe der Diagnostikanden im unterschiedlichen Ausmaß verbreitet ist, so kann es zu Vorwissenseffekten auf die Steuerungsleistung kommen. Diese Annahmen zu Voraussetzungen der Steuerungsleistung werden in den folgenden Abschnitten geprüft.

14.1 Effekte der Computererfahrung und des Alters auf die Steuerungsleistung

Zunächst wurde überprüft, ob die interindividuell unterschiedliche Computererfahrung (operationalisiert über die Skala „CErfahr") oder das Alter einen Einfluß auf die Steuerungsleistung in der „Schneiderwerkstatt" (Kriterium „Kapitalendwert") nahm. Dabei zeigte sich für die Computererfahrung ein schwacher positiver Zusammenhang ($r=.19$; $p=.05$ bei einseitiger Testung) und für das Alter ein nichtsignifikanter negativer Trend ($r=-.18$, n.s). Die Computererfahrung war zu $r=-.28$ negativ mit dem Alter korreliert ($p<.01$; Spearman-Rangkorrelation). Im Rahmen einer univariaten Varianzanalyse ergab sich eine borderline-signifikante Zwei-Weg-Interaktion (F (1,99) = 3.27; $p=.07$) zwischen den jeweils am Median dichotomisierten unabhängigen Variablen „Alter" sowie der „Computererfahrung" einerseits

und der Steuerungsleistung in der „Schneiderwerkstatt" andererseits. Abbildung 7 stellt die Ergebnisse des direkten Vergleichs der Alterseffekte innerhalb der Gruppen mit unterschiedlicher Computererfahrung dar. (Während die Analysen mit der normalisierten abhängigen Variable durchgeführt wurden, zeigen die Graphiken in den Abbildungen 7 und 8 aus Gründen der Anschaulichkeit die Werte für die nicht-transformierte Variable.) In der Gruppe mit hoher PC-Erfahrung (rechte Hälfte der Graphik in Abbildung 7) kam dem Alter keine Bedeutung für die Steuerungsleistung bei. Demgegenüber erzielten in der Gruppe der wenig PC-erfahrenen Teilnehmer die Älteren signifikant schlechtere Ergebnisse (F (1,99) = 4.96; $p < .05$).

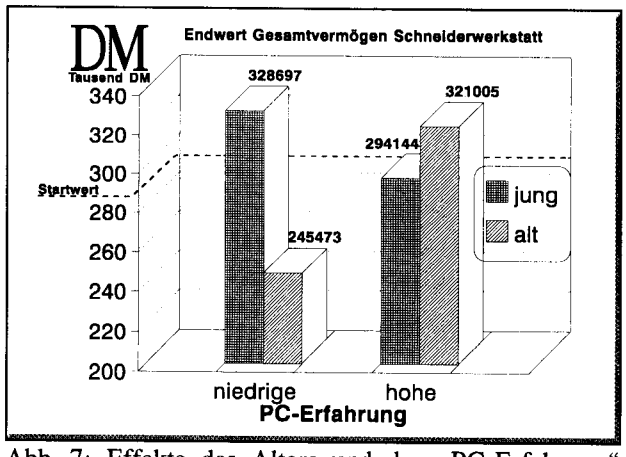


Abb. 7: Effekte des Alters und der „PC-Erfahrung" auf die Steuerungsleistung in der „Schneiderwerkstatt"

Für das Szenario „DISKo" zeigten sich in bezug auf die Steuerungsleistung (Kapitalendwert) keine entsprechenden Effekte, tendenziell ($r = -.11$, n.s.) ging hier ein größeres Ausmaß an Computererfahrung sogar mit einer *schlechteren* Leistung einher. Dieser Befund stellt die für die „Schneiderwerkstatt" aufgezeigten Effekte der Computererfahrung und des Alters auf die Steuerungsleistung allerdings nicht in Frage, wenn man die weiter oben (in Abschnitt 13.2.2) beschriebenen Probleme bei der „DISKo"-Steuerung in Rechnung stellt. Vermutlich aufgrund des de facto überwiegend unwiderruflichen Szenarienereignisses „Maschinenschäden größer als 95%", welches erst bei einer fortgeschrittenen Szenarienbearbeitung auftrat, galt, daß die Anzahl der Entscheidungstakte mit einiger Wahrscheinlichkeit an einen negativen Steuerungserfolg in der Variable „Gesamtvermögen" geknüpft war. Gerade die Teilnehmer mit einer ausgeprägten Computererfahrung zeigten aber eine große „Spielfreude": Der programminterne Parameter für die „Gesamtaktivität" war zu $r = .42$ mit der „Computererfahrung" und zu $r = .35$ signifikant mit der Anzahl der Bearbeitungstakte korreliert. Da mit einer zunehmenden Anzahl an Bearbeitungstakten das Risiko der desaströsen Maschinenschäden stieg, ergab sich aus der Computererfahrung im Szenario „DISKo" hinsichtlich des Steuerungserfolgs kein Vorteil. Dieser Befund dürfte sehr spezifisch für die vorliegende Situation (zu schweres Szenario mit einem unplausiblen und schwerwiegenden Ereignis „Maschinenscha-

den") gewesen sein. Der „DISKo"-programmintern berechnete Parameter zur Beurteilung der Verhaltensweisen und Strategien war zu r=.34 mit der Skala „Einstellung zur Arbeit mit Computern" und zu r=.47 mit der Skala „Computererfahrung" korreliert (p<.01, N =jeweils 100). Auch die um den Intelligenz- und Wissensanteil bereinigte Partialkorrelation zwischen dem „DISKo"-Verhaltensindikator und der Skala „Computererfahrung" blieb mit r=.38 statistisch bedeutsam. Dies läßt vermuten, daß auch bei dem Szenario „DISKo" unter „normalen Umständen" eher mit Effekten der Computererfahrung zu rechnen ist, die in der für die „Schneiderwerkstatt" beschriebenen Richtung verlaufen.

14.2 Effekte des allgemeinen Vorwissens und der Einstellung zur Arbeit mit Computern auf die Steuerungsleistung

Die Frage, ob ein bestimmter Anteil der Problemlöseleistung durch die Vorwissenskompatibilität, d.h. durch die Passung der „semantischen Einbettung" (Variablen-Etiketten, Rahmengeschichte) und dem entsprechenden Vorwissen der Teilnehmer bedingt wird, sollte durch die Zusammenschau der Steuerungsleistungen in den betriebswirtschaftlichen Szenarien und der Auswertung des Kenntnistests „Wirtschaft" geklärt werden. Dabei stellte sich heraus, daß hinsichtlich des Wirtschaftstests von den Teilnehmern überwiegend positive Leistungen erzielt wurden, so daß eine rechtsgipflige Verteilung vorlag. Dieser Befund bestätigt die in Abschnitt 12.2.1.1 dargestellte Annahme, daß alle Teilnehmer über allgemeine Kenntnisse über Wirtschaft verfügten und somit jeder Teilnehmer einen Zugang zu der Rahmengeschichte des Systems finden konnte. Diese Annahme hatte die Wahl eines wirtschaftlich eingekleideten Szenarios begründet. Die Abweichung von der Normalverteilung stellt ein Problem bei der Prüfung von Zusammenhangsbefunden dar. Auch nach der Normalisierung der negativ schiefverteilten Variablen blieb der Kolmogorov-Smirnov-Test auf Normalverteilung mit einem Signifikanzwert von p=.15 im kritischen Bereich. Selbst die Teilnehmer mit den relativ „schlechtesten" Ergebnissen lösten noch über 50% der 20 Items, im Durchschnitt wurden 81,4% der Fragen richtig beantwortet. Wahrscheinlich ist dies ein Grund dafür, daß das so gemessene allgemeine Vorwissen in keinem Zusammenhang zum Steuerungserfolg (Kapitalendwert) bei der „Schneiderwerkstatt" (r=.03) und bei „DISKo" (r=.06) stand (Spearman-Rangkorrelationen). In der Berliner Untersuchung zum Zusammenhang von Intelligenz, Wissen und Problemlösen hatte sich mit einer Korrelation von r=.38 ein mittelstarker Effekt des allgemeinen Wirtschaftswissen auf die Steuerungsleistung ergeben (siehe Abschnitt 9.1.3.3). Die Teilnehmer der Berliner Untersuchung waren

im Durchschnitt 18 Jahre alt und verfügten vermutlich über weniger Wirtschaftskenntnisse. Diese Überlegung basiert auf einem Vergleich mit Daten, die bei insgesamt 3274 Personen mit einer modifizierten Version des Kenntnistests „Wirtschaft" erhoben wurden (zur Stichprobe siehe Kersting und Beauducel, 1997). Während die 972 18jährigen, die vom Alter her den Teilnehmern der Berliner Untersuchung vergleichbar sind, 49 % der Items richtig lösen konnten (SD= 16,0%), konnte die 434 Personen starke Gruppe der über 28jährigen, die vom Alter her mit der Untersuchungsgruppe der vorliegenden Arbeit verglichen werden kann, im Durchschnitt 73% der Fragen korrekt beantworten (SD=15,6%). Dieser Vergleich führt zu der Hypothese, daß sich ein Haupteffekt des allgemeinen Vorwissens nur in einer Gruppe mit einem durchschnittlichen und annähernd normalverteilten Vorwissens-Niveau ergibt, ab einem gewissen Kenntnisniveau das allgemeine Vorwissen aber nur noch eine untergeordnete Rolle für die Steuerungsleistungen spielt (Schwellenmodell). Diese These wäre an einem geeigneten Datensatz zu überprüfen.

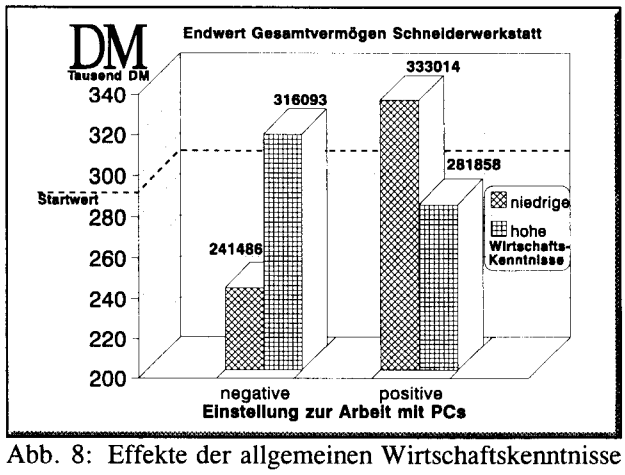


Abb. 8: Effekte der allgemeinen Wirtschaftskenntnisse und der Einstellung gegenüber der Arbeit mit Computern auf die Steuerungsleistung in der „Schneiderwerkstatt"

Das Ausbleiben eines Haupteffekts des allgemeinen Vorwissens besagt nicht, daß die Systemsteuerung hinsichtlich der Wirtschafts-Kenntnisse als voraussetzungsfrei gelten kann. Es zeigte sich vielmehr ein Interaktionseffekt (F (1,99)=5.71; $p<.05$) für die miteinander unkorrelierten Faktoren „Wirtschaftskenntnisse" und „Einstellung zur Arbeit mit Computern" auf die Steuerungsleistung in der „Schneiderwerkstatt" (siehe Abbildung 8). Betrachtet man den Kapitalendwert ausschließlich innerhalb der Gruppe der Teilnehmer mit einer relativ negativen Einstellung zur Arbeit mit Computern (linke Seite der Graphik in Abbildung 8), so führte innerhalb dieser Gruppe ein überaus fundiertes Wirtschaftswissen zu einer signifikant besseren Steuerungsleistung (F (1,99)=4.63; $p<.05$). Bei den Teilnehmern mit einer positiven Einstellung zur Arbeit mit Computern zeigten sich hingegen keine wissensabhängigen bedeutsamen Leistungsunterschiede. Die unabhängige Variable „allgemeines Vorwissen" zeigte auch in Interaktion mit der unabhängigen Variablen „Computererfahrung" einen Einfluß auf die Steuerungs-

leistung, die nach dem für die „Computereinstellung" geschilderten Muster verlief, die Signifikanzgrenze aber verfehlte (F (1,99) = 2.75; p=.10).

Die Analyse der entsprechenden Effekte beim Szenario „DISKo" ist aufgrund der weiter oben (Abschnitt 14.1) geschilderten unplausiblen Zusammenhänge zwischen dem „DISKo"-Kapitalendwert und der Computererfahrung wenig sinnvoll.

14.3 Zusammenfassung und Diskussion zur Voraussetzungsfreiheit der Steuerungsleistung

Die standardisierte Bearbeitung eines Problemlöseszenarios ist mit der Verwendung eines Computers zur Test*vorgabe* verbunden. Damit ergibt sich für die Diagnostikanden zusätzlich zur eigentlichen Problemlösung die Aufgabe der Interaktion mit dem Computer. Nach den vorliegenden Analysen kann es nicht ausgeschlossen werden, daß ein kleiner Varianzanteil der Steuerungsleistung durch die Computererfahrung mitbedingt ist, da sich für die Computererfahrung ein schwacher positiver Zusammenhang mit der Steuerungsleistung in der „Schneiderwerkstatt" ergab. Darüber hinaus galt, daß in der Gruppe der wenig PC-erfahrenen Teilnehmer die Älteren signifikant schlechtere Ergebnisse in der „Schneiderwerkstatt" erzielten. Für das Szenario „DISKo" konnten entsprechende Effekte auf die Steuerungsleistung nicht aufgezeigt werden. Allerdings ging die Computererfahrung bei diesem Szenario mit einer größeren Anzahl an Spielaktivitäten einher, was unter den gegebenen Umständen (siehe Abschnitt 13.2) bei „DISKo" die Wahrscheinlichkeit von fatalen „Maschinenschäden" erhöhte. Dieser ungewöhnliche Trend könnte die eigentliche Wirkung der Computererfahrung überdeckt haben. Zumindest für die Verhaltensbeurteilung konnten deutliche Effekte der Computererfahrung und -einstellung bei „DISKo" aufgezeigt werden. Das Verhalten von Personen mit einer eher hohen Computererfahrung und relativ positiven Einstellung zur Arbeit mit Computern wurde demnach postitiver beurteilt. Während die Computer*erfahrung* auch mit den Intelligenz- und Wissensmaßen konfundiert war, korrelierte die Computer*einstellung* lediglich mit dem für das Szenario „DISKo" berechneten Verhaltensmaß (r=.34, p<.01, N=100), nicht aber mit den Intelligenz- und Wissensmaßen.

Ein direkter Effekt des allgemeinen Vorwissens über die Wissensdomäne, aus der die semantische Einkleidung des Szenarios stammt, ließ sich nicht nachweisen. Lediglich innerhalb der Gruppe der Teilnehmer mit einer relativ negativen Einstellung zur Arbeit mit Computern ging ein hohes allgemeines Vorwissen mit einer besseren Steuerungsleistung in der „Schneiderwerkstatt" einher. Allerdings waren die Kenntnisse der Untersuchungsgruppe insgesamt überdurchschnittlich fundiert.

Für die diagnostische Praxis bedeuten die Befunde, daß die mit Hilfe von computergestützten Problemlöseszenarien gestellten Diagnosen nicht nur ein Indikator für eine *Fähigkeit* sind, sondern zu einem kleinen, unbekannten Anteil auch etwas über die Vorerfahrung der Diagnostikanden mit Computern, über die Kombination dieser Computererfahrung mit dem Alter und über die Einstellung der Testanden zur Arbeit mit Computern in Interaktion mit deren allgemeinen Vorwissen aussagen. Sollten diese zusätzlichen Steuerungsvoraussetzungen gruppenspezifisch (z.B. altersspezifisch, geschlechtsspezifisch) variieren, werden u.U. mit dem diagnostischen Einsatz von computergestützten Problemlöseszenarien bestimmte Personengruppen unter bestimmten Fairneßgesichtspunkten im schwachen Ausmaß diskriminiert (siehe Abschnitt 10.1). Außerdem ist die Interpretation der Steuerungsleistung als reiner Fähigkeitsindikator durch die Befunde etwas in Frage gestellt. In der Eignungsdiagnostik ist die Verwendung von computergestützte Problemlöseszenarien immer dann problematisch, wenn die zusätlichen Varianzquellen der Steuerungsleistung (z.B. Computererfahrung) nicht eignungsrelevant sind. In jedem Falle ist es aus diagnostischer Sicht unbefriedigend, daß die psychologisch unterschiedlichen Determinanten der Steuerungsleistung in der Pauschaldiagnose nicht weiter differenziert werden. Es empfiehlt sich daher, zur Absicherung der mit Hilfe computergestützter Problemlöseszenarien gestellten Diagnose stets auch die Computererfahrung, die Einstellung zur Arbeit mit Computern und das allgemeine Vorwissen über die Domäne, aus die Rahmengeschichte des Szenarios entlehnt ist, separat zu diagnostizieren.

15. Zur Konstruktvalidität der Steuerungsleistung

Das Verständnis der Problemlösefähigkeit ist - besonders auf der Konstruktebene - noch sehr lückenhaft (siehe Abschnitt 9.1 des Theorieteils), eine empirisch fundierte Elaboration eines Konstrukts „Problemlösen" existiert bislang nicht. Neben dem Postulat, daß mit den Steuerungsleistungen eine „neue" Fähigkeit, nämlich die Fähigkeit zum Problemlösen gemessen wird, steht die Vermutung, daß es sich bei der Steuerungsleistung um einen neuen Indikator für die etablierten Konstrukte Intelligenz und Wissen handelt. Die Konstruktvalidität berührt sowohl die Entscheidung, in welcher Situation Problemlöseszenarien als diagnostische Instrumente zum Einsatz kommen können als auch die Frage, wie die so gewonnenen diagnostischen Informationen zu interpretieren sind. Solange die Beziehung zwischen dem Indikator und dem Indizierten nicht näher bestimmt ist, kann das diagnostische Zeichen nicht eindeutig interpretiert werden. Entsprechend den Ausführungen im Abschnitt 9.1 wird ein Zusammenhang zwischen Intelligenz, Wissen und der Steuerungsleistung erwartet. Im folgenden werden die entsprechenden Analysen für die beiden Szenarien dargestellt.

Zunächst wurde für die Skalen der Allgemeinen Intelligenz und der Verarbeitungskapazität sowie für die Indikatoren des systemspezifischen Sachwissens über die „Schneiderwerkstatt", über „DISKo" und für ein Aggregat der beiden systemspezifischen Wissensindikatoren der Zusammenhang zur Steuerungsgüte auf bivariater Ebene berechnet. Tabelle 14 gibt die Korrelationen der entsprechenden Indikatoren mit der Steuerungsleistung in den Szenarien „Schneiderwerkstatt" und „DISKo" sowie für ein Aggregat der Steuerungsleistung in den beiden Szenarien für die Gesamtgruppe wieder.

Aggregiert wurden die z-transformierten Leistungen aufgrund des neuen Problemlösegütemaßes in beiden Szenarien, wobei für das Szenario „DISKo" das über die ersten acht Bearbeitungstakte berechnete Maß Berücksichtigung fand. Diese Indikatoren wurden aufgrund der Parallelitätswerte (siehe unten, Abschnitt 15.4) für die Aggregation der Steuerungsleistungen ausgewählt. Bei „DISKo" wurde neben der Steuerungsleistung auch der programminterne Parameter zur Bewertung der Verhaltensweisen und Strategien in der Korrelationsanalyse berücksichtigt.

Der bivariate Zusammenhang zwischen der Steuerungsleistung und der Intelligenz wird in Abschnitt 15.1 erläutert und diskutiert. Zusätzlich zu der Analyse für die Gesamtgruppe werden die Ergebnisse für die Teilgruppe berichtet, die den BIS-4 Test bearbeitet haben. Für diese Gruppe konnten die Zusammenhänge auch auf

Tab. 14: Korrelationen der Steuerungsleistung in den beiden Szenarien mit Intelligenz und systemspezifischem Sachwissen; für das Szenario „DISKo" wurde neben der Steuerungsleistung auch das Verhaltensmaß berücksichtigt

		Intelligenz		Sachwissen über das Szenario...		
		AI	K	„SWS"	„DISKo"	Aggregat
„SWS"	Kapitalendwert	.20*	.21*	.22*	.21*	.32**
	neues PLG	.23*	.22*	.22*	.22*	.29**
DISKo	Kapitalendwert	.00	.02	.14	.23*	.22*
	neues PLG	.21*	.26*	.31**	.32**	.37**
	neues PLG-8	.30**	.31**	.33**	.31**	.38**
	„Verhalten"	.32**	.33**	.23*	.11 p-w	.19 p-w
Aggregat: „SWS" (neues PLG) und DISKo (PLG-8)		.33**	.34**	.34**	.32*	.41**

* $p < .05$, ** $p < .01$, p-w part-whole korrigierte Korrelationen
Sachwissen:
für die „Schneiderwerkstatt": Test „WIS", Aggregat beider Skalen;
für „DISKo": Differenz der „richtigen" abzügl. der „falschen" „Testlauf"- Prognosen
„AI" und „K": Intelligenztestskalen „Allgemeine Intelligenz" u. „Verarbeitungskapazität"
„SWS": Problemlöseszenario „Schneiderwerkstatt"
PLG: Problemlösegütemaß; PLG-8: berechnet über die ersten acht Bearbeitungstakte
„Verhalten": programminterne Beurteilung der Verhaltensweisen und Strategien
N für den Zusammenhang...
der Intelligenzskalen mit „SWS"-Indikatoren: 103, mit „DISKo"-Indikatoren: 99
der Wissensskala Test WIS mit „SWS"-Indikatoren: 104, mit „DISKo"-Indikatoren 100
des Wissensindikators bei „DISKo" mit allen Steuerungsleistungsindikatoren: 100
für die Aggregate (Wissen und Steuerungsleistung) entspricht das N dem für „DISKo" genannten Wert

der Ebene der übrigen Intelligenzskalen berechnet werden. Das Verhältnis zwischen systemspezifischem Sachwissen und der Steuerungsleistung ist, ebenso wie die Frage des steuerungsbedingten Wissenserwerbs, Thema des Abschnitts 15.2. In Abschnitt 15.3 wird untersucht, welcher Anteil der Problemlösevarianz durch eine gemeinsame Berücksichtigung von Intelligenz und Wissen aufgeklärt werden kann. Mit der Frage, ob die für die beiden Szenarien bestimmten Steuerungsleistungen über den gemeinsamen Anteil von Intelligenz und Wissen hinaus generalisierbar sind, setzt sich der letzte Abschnitt (15.4) des Kapitels zur Konstruktvalidität der Steuerungsleistungen auseinander.

15.1 Intelligenz und Problemlösen

Für die „Schneiderwerkstatt" zeigte sich ein nur schwacher, aber statistisch bedeutsamer Zusammenhang zwischen den Skalen der Intelligenz und den Indikatoren der Steuerungsleistung (siehe Tabelle 14). Erwartungsgemäß unterschied sich das Korrelationsmuster zwischen Intelligenz und Steuerungsleistung für das „eigentliche" Gütemaß, den Kapitalendwert, nicht von dem Muster für das neu definierte Problemlösegütemaß. Substantielle Unterschiede in den korrelativen Außenbeziehungen der beiden Maße sind nur zu erwarten, falls die Steuerung des betriebswirtschaftlich eingekleideten Szenarios zu schwer ist und am Gesamtvermögen orientierte Parameter die Steuerungsleistung der Testanden aufgrund der überwiegend negativen Gewinnspanne nicht mehr intern valide abbilden. Dies war für das Szenario „DISKo" der Fall (siehe die Abschnitte 13.1 und 13.2). Äquivalent zu den Befunden für eine frühere, zu schwere Version der „Schneiderwerkstatt" (siehe oben, Abschnitt 7.3) galt in der vorliegenden Untersuchung für das Szenario „DISKo": in bezug auf das tradierte Problemlösegütemaß „Kapitalendwert" zeigte sich kein Zusammenhang zwischen Intelligenz und Steuerungsleistung. Erst mit dem neuen Problemlösegütemaß konnten die Zusammenhänge aufgezeigt werden. Damit hat die von Süß et al. (1993b) für die Berliner Erstuntersuchung vorgenommene Aufgabenanalyse und Definition eines neuen Problemlösegütemaßes neben der Bestätigung in der Berliner Wiederholungsuntersuchung eine weitere Replikation in einer unabhängigen Untersuchung mit einem anderen Szenario erfahren. Ohne diese Analyse wären die Zusammenhangsbefunde für das am Kapitalendwert orientierte „DISKo"-Problemlösegütemaß als Hinweis auf eine Dissoziation von Intelligenz und Steuerungsleistungen interpretiert worden.

Die Assoziation zwischen Intelligenz und Steuerungsleistung zeigte sich insbesondere dann, wenn bei der Berechnung des neuen Problemlösegütemaßes lediglich die ersten acht „DISKo"-Bearbeitungstakte berücksichtigt wurden. Das so gebildete Maß war vergleichsweise weniger von dem Szenarienereignis „Maschinenschäden" beeinflußt. Dieses Ereignis beeinträchtigte vermutlich - ebenso wie die infolge der Steuerungsprobleme auftretenden negativen Gewinnspannen - die interne Validität des ursprünglichen Maßes „Kapitalendwert".

Beachtlich sind die substantiellen Korrelationen zwischen den Intelligenztestleistungen und der programminternen Beurteilung der Verhaltensweisen und Strategien. Diese Verbindung deutet darauf hin, daß die mit diesem Verhaltensindikator erfaßten Leistungen zu einem entscheidenden Anteil im Fähigkeitsbereich und nicht (ausschließlich) im Motivations- und/ oder Temperamentsbereich zu verankern sind.

Um den Zusammenhang zwischen Intelligenz und Steuerungsleistung besser abschätzen zu können, wurden die z-transformierten Steuerungsleistungen der beiden

Szenarien aggregiert (jeweils das neue Problemlösegütemaß, für „DISKo" das über die ersten acht Bearbeitungstakte gerechnete Maß). Durch die Aggregation kann die Reliabilität der Messung der Steuerungsleistung gesteigert und die Symmetrie der Messungen erhöht werden (Wittmann, 1988). Auf dieser Aggregationsebene zeigten sich substantielle Korrelationen zwischen den Indikatoren der Intelligenz und der Steuerungsleistungen, die in ihrer Höhe den entsprechenden Befunden der Berliner Wiederholungsuntersuchung (Süß et al. 1991) entsprachen. Intelligenz und Problemlösen sind demzufolge in etwa der Höhe miteinander korreliert wie Aufgaben zu verschiedenen Dimensionen der Intelligenz (z.B. „Verarbeitungskapazität" und „Merkfähigkeit") untereinander.

Um ein auf den Zusammenhang zwischen einzelnen Dimensionen der Intelligenz und den Steuerungsleistungen fokussiertes Bild zu erhalten, wurden die Daten der 64 Testanden, die den BIS-4 Test bearbeitet hatten, gesondert analysiert. Tabelle 15 gibt die Korrelationen für die Intelligenzskalenleistungen wieder, die Problemlösegüte- und Wissensmaße entsprechen den für Tabelle 14 erläuterten Indikatoren.

Tab. 15: Korrelationen der Intelligenzleistungen im BIS-4 Test mit der Steuerungsleistung in den beiden Szenarien; für das Szenario „DISKo" wurde zusätzlich zur Steuerungsleistung auch der Verhaltensindikator berücksichtigt; Subgruppenanalyse: nur Probanden, die den BIS-4 Test bearbeitet haben

BIS-Skalen →:	K	E	M	B	V	F	N	AI
SWS-Kapitalendwert	.26*	.08	.37**	.16	.14	.33**	.24	.26*
SWS-PLG	.24	.15	.30*	.22	.15	.33**	.26*	.28*
DISKo-Kapitalendwert	.01	.05	-.11	-.05	-.06	.04	-.07	-.03
DISKo-PLG	.33**	.09	.21	.23	.17	.32*	.21	.26*
DISKo-PLG-8	.40**	.16	.35**	.40**	.30*	.41**	.32*	.39**
DISKo-Verhalten	.36**	.35**	.31*	.29*	.28*	.50**	.25*	.39**
Aggregat SWS-DISKo	.39**	.20	.40**	.39**	.29*	.45**	.37**	.42**

* $p < .05$, ** $p < .01$
Erläuterung der Abkürzungen der BIS-Skalen: siehe Tabelle 5;
Erläuterung der Indikatoren der Steuerungsleistung: siehe Tabelle 14;
N für den Zusammenhang der Intelligenzskalen mit SWS-Indikatoren: 63,
mit „DISKo"-Indikatoren sowie dem Aggregat: 61

Die Höhe der Korrelationen zwischen der „Allgemeinen Intelligenz" und der Steuerungsleistung in der „Schneiderwerkstatt" erwies sich als geringer als bei anderen Studien, auch kam der „Verarbeitungskapazität" in der vorliegenden Unter-

suchung keine herausragende Stellung gegenüber der „Allgemeinen Intelligenz" zu. Dieser Befund kann auf eine mögliche Varianzeinschränkung hinsichtlich der Intelligenztestleistungen der Untersuchungsgruppe hindeuten. Wie in Abschnitt 12.1 bereits erwähnt, muß davon ausgegangen werden, daß es sich bei der Untersuchungsgruppe um eine nach Intelligenz - und insbesondere nach der Verarbeitungskapazität - vorausgewählte Gruppe handelte. Entsprechende Restriktionen der Zusammenhangsbefunde für die Intelligenz sind daher wahrscheinlich, können hier aber nicht empirisch ermittelt werden. Von den operativen Intelligenztestleistungen stand außerdem noch die „Merkfähigkeit" im signifikanten Zusammenhang zur Steuerungsleistung in beiden Szenarien. Daß die Steuerung des „DISKo" Szenarios Anforderungen an die Merkfähigkeit stellt, erscheint nicht unplausibel. Sowohl die Orientierung innerhalb des Programms (z.B. die Klärung der Frage, in welchem Bearbeitungstakt sich das Szenario befindet, siehe oben, Abschnitt 13.2.2) als auch der Transfer der in den Testdurchläufen gesammelten Erfahrungen auf die Steuerung kann zu einer Beanspruchung der Merkfähigkeit führen. Für die „Schneiderwerkstatt" ist der Befund hingegen ungewöhnlich; in der Berliner Wiederholungsuntersuchung kam der Merkfähigkeit bei der multiplen Vorhersage der Steuerungsleistung sogar eine Suppressorwirkung zu. Bei der Interpretation der Ergebnisse kann zum einen das Alter der Probanden angeführt werden, welches höher war als das Alter der in anderen Untersuchungen häufig als Versuchspersonen agierenden Schüler und Studenten. Merkfähigkeitsleistungen lassen mit dem Alter nach, dieser Umstand könnte altersspezifische Effekte auf die Anforderungen von Leistungsaufgaben im allgemeinen und computergestützten Problemlöseszenarien im besonderen zeitigen. Andererseits könnte es sich bei dem Befund aber auch um einen Effekt der Darbietungsfolge handeln: Während die Skala „Merkfähigkeit" bei den 35 Teilnehmern, die die „Schneiderwerkstatt" als erstes Szenario steuerten, zu $r=.09$ mit dem neuen Problemlösegütemaß korreliert war, betrug die Korrelation für die Gruppe derjenigen 28 Personen, die vor der „Schneiderwerkstatt" bereits das Szenario „DISKo" gesteuert hatten, $r=.49$ ($p<.01$). Die Merkfähigkeit ist möglicherweise nicht direkt für die Steuerung, sondern indirekt für die Sicherung der Erfahrungen von einem Szenario zum nächsten verantwortlich. Dieser interessante Aspekt müßte in einer anderen Studie mit einem diesbezüglich geeigneteren Versuchsplan überprüft werden. Bei den inhaltsgebundenen Fähigkeiten überrascht die für die „Schneiderwerkstatt" schwer zu erklärende Bedeutung des figural-anschauungsgebundenen Denkens. Entsprechende Anforderungen werden zwar bei der Steuerung von „DISKo" durch die Möglichkeit zur Interpretation von Verlaufskurven der Variablen gestellt, die „Schneiderwerkstatt" sieht hingegen keinerlei Graphiken vor.

15.2 Systemspezifisches Wissen und Problemlösen

Die Steuerungsleistungen in beiden Szenarien konnten durch den jeweils szenarienspezifischen Wissenstest vorhergesagt werden (siehe Tabelle 14). Die lediglich moderate Höhe der Korrelation unterschätzt dabei möglicherweise den Zusammenhang zwischen Wissen und Problemlösen, da auf der Wissensseite nur das deklarative Sachwissen der Testanden berücksichtigt wurde. Süß et al. (1992) hatten die Hypothese aufgestellt, daß der Zusammenhang zwischen Sachwissen und Steuerungsleistung durch strategisches Handlungswissen und/oder quantitatives Sachwissen moderiert wird - beide Wissensformen waren in der vorliegenden Studie nicht berücksichtigt worden.

Beachtlich ist, daß darüber hinaus sowohl der mit Hilfe der beiden „WIS"-Skalen gemessene Umfang des systemspezifischen Sachwissens über die „Schneiderwerkstatt" als auch das programmintern bestimmte Sachwissen über „DISKo" (bestimmt anhand der Auswertung der semi-quantitativen Prognosen über die Effekte von Maßnahmen, die im Rahmen sogenannter „Testdurchläufe" getroffen wurden) mit verschiedenen Indikatoren der Steuerungsleistung *in beiden Szenarien* korrelierten. Da die Wissensindikatoren für die beiden Szenarien lediglich zu $r=.18$ (n.s., $N=100$) miteinander korreliert waren, kann eine Identität der beiden Wissensmaße als Grund für diese „Austauschbarkeit" ausgeschlossen werden. Die szenarienübergreifende prognostische Reichweite der beiden Wissensindikatoren könnte durch die Intelligenz als gemeinsame Drittvariable der Wissens- und Steuerungsmaße bedingt sein. In der Gesamtgruppe war der Zusammenhang zwischen den Leistungen im Wissenstest für die „Schneiderwerkstatt" und den Leistungen im Intelligenztest mit einer Korrelation in Höhe von $r=.52$ ($p<.01$, $N=103$) für die Skala „Verarbeitungskapazität" am deutlichsten. Aber auch die Skala „Allgemeine Intelligenz" war zu $r=.45$ ($p<.01$) mit dieser Wissensleistung korreliert. Der Nachweis des Zusammenhangs zwischen Intelligenz und Wissen stellt eine empirische Fundierung integrativer Konzepte von Intelligenz und Wissen - wie sie beispielsweise von Cattell und Horn erarbeitet wurden (z.B. Cattell, 1957, 1963, 1971; Horn, 1980) - dar. Neben den Verbindungen auf der Konstruktebene dürfte ein Teil der Korrelation aber auch auf die gemeinsame Methodenvarianz der beiden paper-pencil-Verfahren zurückzuführen sein. Der programmintern berechnete Wissensindikator für „DISKo", der sich auf computergestützt abverlangte Prognosen samt Rückmeldungen stützt und sich somit einer anderen Methode der Wissensdiagnostik bedient, erwies sich als weitgehend intelligenzunabhängig ($r=.03$ mit der „Verarbeitungskapazität" und $r=.00$ für die „Allgemeine Intelligenz", $N=99$).

Die Generalisierbarkeit des Wissensindikators für die „Schneiderwerkstatt" erschöpfte sich aber nicht in dessen Intelligenzanteil. Selbst wenn man die „Verarbei-

tungskapazität" aus der Leistung im WIS-Test auspartialisiert, blieb die Korrelation zwischen dem Sachwissen über die „Schneiderwerkstatt" und der mit Hilfe des neuen Problemlösegütemaßes bestimmten Steuerungsleistung in „DISKo" auf dem 5% Niveau signifikant.

Es ist zu vermuten, daß sich das Wissen über die „Schneiderwerkstatt" aufgrund der gemeinsamen betriebswirtschaftlichen Rahmengeschichte auch bei der Steuerung von „DISKo" nutzen läßt und vice versa. Dies ist plausibel, solange - wie in der vorliegenden Studie - lediglich semi-quantitatives Wissen abgefragt wird. Auf einem höheren Präzisionsniveau, auf der quantitativen Ebene, müßten sich die unterschiedlichen Gewichtungen der Variablenbeziehungen in den beiden Szenarien bemerkbar machen. Beschreibungen von Variablenzusammenhängen auf semi-quantitativer Ebene, wie sie im „Schneiderwerkstatt"-spezifischen Wissenstest verwendet wurden - z.B. *„eine Erhöhung des Lohns steigert die Arbeitsmotivation"* - können hingegen für betriebswirtschaftlich eingekleidete Aufgaben häufig szenarienübergreifend gültig sein. Dabei dürfte den Testanden ohne weiteres auch die Anpassung einzelner Variablenetiketten gelingen, so daß beispielsweise die Kenntnis des für die „Schneiderwerkstatt" gültigen Zusammenhangs *„eine Erhöhung des Hemdenpreises senkt die Nachfrage"* auch für die Festsetzung des Produktpreises in „DISKo" nützt.

Hinweise auf den Wissenstranfer von einem Szenario zum anderen ergeben sich auch aus einer Analyse der Effekte der Darbietungsfolge auf das Sachwissen über die „Schneiderwerkstatt". Univariate Methoden zeigten einen Haupteffekt (F (1,102) = 8.45; $p < .01$) der Darbietungsfolge auf den Umfang des verbalisierbaren Sachwissens. Obwohl die Analyse mit dem Aggregat der beiden jeweils z-transformierten WIS-Skalenleistungen gerechnet wurde, wird im folgenden der Anschaulichkeit halber der relative Anteil der im Durchschnitt über beide Skalen richtig gelösten Wissensitems referiert. Die 54 Testanden, für die das Szenario „Schneiderwerkstatt" die zuerst dargebotene Problemlöseaufgabe darstellte, verfügten im Anschluß an die Steuerung mit durchschnittlich 63,1 Prozent richtigen Lösungen über weniger systemspezifisches Sachwissen als die 50 Testanden, die bereits mit „DISKo" Steuerungserfahrungen sammeln konnten und die „Schneiderwerkstatt" als zweite Problemlöseaufgabe dargeboten bekamen (70,2%). Der zweistufige Faktor „Darbietungsabfolge" konnte 7,6% der Wissensvarianz binden.

Für den Wissensindikator bei „DISKo" ließ sich ein entsprechender Effekt nicht ausmachen, der Befund verlief hier tendenziell sogar gegenläufig, war aber statistisch nicht bedeutsam. Zumindest für den Wissenstest zur „Schneiderwerkstatt" galt somit, daß es sich um einen erfahrungssensitiven Test handelt. Dies hatte sich auch in der Berliner Untersuchung gezeigt, wo (1.) die Systemsteuerung, (2.) die wissensvermittelnde Instruktion und (3.) (mit geschlechtsspezifisch bedingten Einschränkungen) die Systemexploration zu einem Wissens*zuwachs* führten (Süß,

1996). Der Wissenstest zur „Schneiderwerkstatt" kann somit nicht nur als ein Indikator für das systemspezifische Vorwissen der Testanden, sondern auch als ein Indikator des Wissenserwerbs angesehen werden.

15.3 Vorhersage der Problemlöseleistungen durch Wissen und Intelligenz

Um differenzierte Aussagen über die Fähigkeit zur Steuerung komplexer Systeme treffen zu können, wurde über die bivariaten Analysen hinaus mit einer hierarchischen Regressionsanalyse geprüft, ob trotz der substantiellen Korrelation von Intelligenz und Wissen die zusätzliche Berücksichtigung des zweiten Prädiktors einen inkrementellen Varianzbeitrag zur Vorhersage der Problemlösegüte leistet. Die Analyse wurde für das Kriterium „Problemlösen" auf der Ebene der über beide Szenarien aggregierten Leistungen durchgeführt (siehe Tabelle 14). Eingeführt wurden zunächst die beiden Indikatoren für das systemspezifische Wissen und dann die „Allgemeine Intelligenz".

Tab. 16: Hierarchische Regression von Wissen und „Allgemeiner Intelligenz" auf die Steuerungsleistung
(Aggregat „SWS-PLG" und „DISKo-PLG8" laut Tab. 13)

	Kriterium: Problemlöseleistung						Kreuzvalidierung:	
↓ Prädiktoren ↓	r	R	B	β	R_{diff}		Gruppe	
Wissen („SWS")	.35**	.35**	.174	.196			1	2
Wis. („DISKo")	.32**	.43**	.452	.273	.08**	R	.47	.49
Allg. Intelligenz	.33**	.48**	.427	.245	.05*	R^2_{korr}	.16	.19
	p<.01; *p<.05; Intercept=-.012 N=99 R^2= .23 F (3,95)=9.71 R^2_{korr} .21					N	49	50

Erläuterung: SWS= „Schneiderwerkstatt"; „Wis." = Wissen

Tabelle 16 zeigt den unstandardisierten Regressionskoeffizienten (B), die Steigung, den standardisierten Regressionskoeffizienten (beta) sowie R, R^2 und das adjustierte R^2 nach der Berücksichtigung aller unabhängigen Variablen. (Die Höhe der Korrelationen zwischen den Prädiktoren wurde in Abschnitt 15.2 beziffert.). Nachdem alle Prädiktoren in die Regressionsgleichung aufgenommen worden waren, betrug R=.48 ($F(3,95)=9.71$, $p<.001$). Zum Ende des ersten Schritts, unter Berück-

sichtigung des Wissenstests für die „Schneiderwerkstatt", betrug R=.35 (F_{change} (1,97)=13,88, p<.001). Nach dem zweiten Schritt, nachdem auch der Indikator für das Sachwissen bei „DISKo" einkalkuliert war, betrug R=.43 (F_{change} (2,96) = 7,31, p<.01). Die Berücksichtigung der Wissenskala war mit einem signifikanten Zuwachs in R^2 verbunden. Nach dem letzten Schritt, bei dem neben den beiden Wissensindikatoren auch die Allgemeine Intelligenz berücksichtigt wurde, betrug R =.48 (F_{change} (3,95)=5,97,p<.05). Auch die durch die Aufnahme der „Intelligenz" in die Gleichung erzielte Steigerung des R^2 Wertes war statistisch bedeutsam.

Zur Kreuzvalidierung wurden die an der Gesamtstichprobe regressionsanalytisch geschätzten Gewichte zur Vorhersage in zwei Zufallsteilstichproben eingesetzt (geschichtete Zufallssplittung unter Berücksichtigung des Alters und der Dreiereinteilung des unterschiedlichen Berufserfolgspotentials der Teilnehmer, siehe oben, Abschnitt 12.3). Die multiple Korrelation von R=.48 hielt der Kreuzvalidierung stand. Eine nominell gleich hohe multiple Korrelation erhielt man auch, wenn man - bei ansonsten gleicher Analyse - anstelle der „Allgemeinen Intelligenz" die Skala „Verarbeitungskapazität" als zweiten Prädiktor berücksichtigte.

Die Allgemeine Intelligenz lieferte zustätzlich zum systemspezifischen Wissen einen substantiellen Beitrag zur Vorhersage der Steuerungsleistungen. Beide Prädiktoren erlaubten gemeinsam eine Vorhersage der Problemlöseleistungen, deren Güte angesichts der vermutlich vergleichsweise geringen Reliabilität dieser Maße beachtlich ist. Die Problemlöseleistungen waren somit in den beiden Konstrukten Intelligenz und Wissen fest verankert. Andererseits blieb ein bedeutsamer Teil der Varianz der Steuerungsleistungen unaufgeklärt. Im nächsten Abschnitt soll untersucht werden, ob die verbleibende systematische Varianz als Indikator einer konzeptuell eigenständigen „Problemlösefähigkeit" angesehen werden kann.

15.4 Zur Generalisierbarkeit der Problemlöseleistungen

Als letzter Aspekt der Konstruktvalidität sollte geprüft werden, in wie weit der Erfolg beim Problemlösen über das jeweils verwandte Problemlöseszenario hinaus generalisiert werden kann. Der Einsatz von Problemlöseszenarien wird erst in dem Maße eignungsdiagnostisch ergiebig, als die Gültigkeit der Ergebnisse nicht auf die einzelne Meßsituation beschränkt ist. Die Konzeptualisierung einer Problemlösefähigkeit setzt auf der empirischen Seite voraus, daß die thematisierten Leistungen mit verschiedenen und unabhängigen Operationalisierungen repliziert werden können. Zeitlich instabile Messungen sind der empirischen Fundierung eines Fähigkeitskonstrukts ebenso abträglich wie Leistungen, die lediglich aufgabenspezifisch

(hier: szenarienspezifisch) auftreten und nicht generalisiert werden können. In diesem Fall kann zwischen aufgabenspezifischer und konstruktbezogener Leistungsvarianz nicht unterschieden werden. Als ein Hinweis auf die Generalisierbarkeit kann in der vorliegenden Studie der Zusammenhang zwischen dem Steuerungserfolg in den beiden Szenarien „Schneiderwerkstatt" und „DISKo" gewertet werden. Die oben (Abschnitt 15.2) dargestellten Befunde für den Wissenstransfer deuten darauf hin, daß sich die beiden Szenarien zumindest von der Seite der inhaltlichen Aufgabenmerkmale her so ähnlich sind, daß die Gefahr einer Überlagerung generalisierbarer Problemlösefähigkeit durch unterschiedliche Wissensvoraussetzungen nicht besteht. Tabelle 17 zeigt die entsprechenden Korrelationen der bei den beiden Szenarien gezeigten Steuerungsleistungen für die verschiedenen Problemlösegütemaße.

Tab. 17: Korrelationen der Steuerungsleistungen in beiden Szenarien
N=100 * p<.05, ** p<.01

	„Schneiderwerkstatt"	
	Kapitalendwert	neues PLG
DISKo Kapitalendwert	.05	-.02
DISKo neues Problemlösegütemaß (PLG)	.23**	.35**
DISKo neues PLG über die ersten 8 Takte	.28**	.39**

Substantielle Zusammenhänge der Steuerungsleistungen zeigten sich nur, wenn für „DISKo" das neue Problemlösegütemaß zugrunde gelegt wurde. Dies ist ein weiterer Beleg dafür, daß es dem Indikator „Gesamtvermögen am Ende der Bearbeitungszeit" an interner Validität für die Steuerungsleistungen beim „DISKo"-Szenario mangelte. Auch die Beschränkung der Auswertung der „DISKo"-Steuerungsleistungen auf die ersten acht Bearbeitungstakte erhöhte nominell die Zusammenhänge zur „Schneiderwerkstatt". Dies lag - wie bereits angeführt - möglicherweise in dem Szenarienereignis „Maschinenschäden" begründet, welches in der hier verwendeten Version von „DISKo" mit zunehmender Anzahl an Bearbeitungstakten der Steuerbarkeit des Szenariums - und somit auch der Bildung intern valider Steuerungsindikatoren - abträglich war. Die deutlichsten Zusammenhänge zeigten sich, wenn auch für die „Schneiderwerkstatt" das neue Problemlösegütemaß gebildet wurde und somit die beiden Indikatoren größtmögliche Gemeinsamkeiten aufwiesen.

Selbst wenn man zunächst den Zusammenhang zwischen den Steuerungsleistungen in den beiden Szenarien isoliert betrachten würde, so wäre eine Korrelation in Höhe von maximal r=.39 für den Zusammenhang von Indikatoren des gleichen Konstrukts als niedrig zu bewerten. Zum Vergleich sind die entsprechenden Inter-

korrelationen von Intelligenzleistungen in unterschiedlichen Testverfahren heranzuziehen. Jensen (1984, S. 570) berichtet eine entsprechende durchschnittliche Interkorrelation von r=.77 (bzw. r=.86 nach Attenuationskorrektur) zwischen 30 verschiedenen Intelligenztestverfahren. Angesichts der in den vorherigen Abschnitten dieses Kapitels erläuterten Zusammenhänge zwischen der Steuerungsleistung einerseits sowie Wissen und Intelligenz andererseits, kann der Nachweis der Korrelation der Szenarien untereinander allerdings tatsächlich nur als notwendige, nicht aber als hinreichende Voraussetzung für die Annahme einer generalisierbaren Problemlösefähigkeit gelten. Möglicherweise war die Korrelation der Steuerungsleistungen in den beiden Szenarien ja lediglich eine Funktion der gemeinsamen Intelligenz- und Wissensanteile der Steuerungsleistungen. Der Argumentation von Süß (1996, S. 194 f.) folgend wird die Annahme eines neuen Konstrukts auch in der vorliegenden Arbeit nur dann als sinnvoll erachtet, wenn die Leistungen wiederholt (mit ein und demselben Szenario und/oder mit verschiedenen Szenarien) erbracht werden können und wenn der stabile Anteil der Steuerungsleistung nicht genauso gut duch eine Kombination von Leistungen in bereits etablierten Konstrukten wie Intelligenz und Wissen vorgesagt werden kann. *„Sonst sollte im Sinne der Sparsamkeit auf ein neues Konstrukt verzichtet werden, da es nichts erklären kann und nur neue kommunikative Probleme schafft.“* (Süß, ebd.). Zur Prüfung der Frage, ob die Steuerungsleistungen in den beiden Systemen systematische Varianz enthielten, die nicht durch Intelligenz und Wissen vorhergesagt werden kann, wurde die von Süß (1996, ebd.) vorgeschlagene Analyse der Residualvarianzen angewandt. Um eine optimale Vorhersagbarkeit der Steuerungsleistung durch Intelligenz und Wissen zu gewährleisten, sollte auch die Merkfähigkeit - die sich als guter Prädiktor erwiesen hatte, siehe Tabelle 15 - in der Prognose der Steuerungsleistungen Berücksichtigung finden. Aus diesem Grunde wurde die Analyse der Residualvarianzen auf die Probanden beschränkt, die den BIS-4 Test bearbeitet haben. Durch diese Beschränkung auf eine Subgruppe und durch die Datenausfällen auf Seiten des Problemlöseszenarios „DISKo“ (siehe Abschnitt 12.4) ergab sich eine Reduzierung der Fallzahl auf 61 Personen. Die für diese Subgruppe gültigen bivariaten Zusammenhänge sind in Tabelle 18 verzeichnet. Auf Seiten der Steuerungsleistungen wurden die Indikatoren gewählt, die am höchsten miteinander korrelierten. Dies war für beide Szenarien das neue Problemlösegütemaß, dessen Berechnung bei „DISKo“ auf die ersten acht Bearbeitungstakte beschränkt wurde. Somit handelte es sich um eine für das Problemlösen „idealisierte“ Analyse unter den optimalen Voraussetzungen möglichst intern valider und symmetrischer Problemlösegütemaße. Diese beiden Maße waren in der Gruppe der analysierten 61 Personen zu $r=.30$ ($p<.05$) miteinander korreliert.

Mit zwei hierarchischen Regressionsanalysen wurde für jedes Szenario die Steuerungsleistung durch das Aggregat der beiden systemspezifischen Wissenstests sowie

durch die Verarbeitungskapazität und die Merkfähigkeit vorhergesagt (vergleichbar dem in Abschnitt 15.3 für das Aggregat der Steuerungsleistung und die Gesamtgruppe ausführlicher beschriebenem Vorgehen). Zusätzlich wurden pro Regressionsanalyse die standardisierten Residuen berechnet.

Tab. 18: Hierarchische Regressionen von Wissen, Verarbeitungskapazität und Merkfähigkeit auf die Problemlöseleistung je Szenario mit anschließender Residuenberechnung (Teilgruppe mit BIS-4 Test: N=61)

	Kriterium: Problemlöseleistung									
	„Schneiderwerkstatt", neues Problemlösegütemaß					„DISKo", neues Problemlösegütemaß über 8 Takte				
↓Präd.	r	R	B	β	R_{diff}	r	R	B	β	R_{diff}
Wissen	.22[1]	.22[1]	.262	.192		.40**	.40**	.415	.323	
K	.24[1]	.28[1]	.004	.004	.04	.40**	.48**	.139	.151	.08*
M	.30*	.36*	.245	.283	.08[1]	.35**	.52**	.184	.226	.06
	Intercept=.033 R^2= .13 F (3,57)=2.84* R^2_{korr} .08					Intercept=-.075 R^2= .27 F (3,57)=6.9** R^2_{korr} .23				
Korrelation der PLGs	.30*									
Korrelation der Residuen	.16 (n.s., p=.224)									

* p<.05; ** p<.01; [1]p<.10

Subgruppenanalyse: nur Probanden, die den BIS-4 Test bearbeitet haben

Präd.= Prädiktoren:
Wissen= Aggregat der beiden szenarienspezifischen Wissensindikatoren
K= Skala Verarbeitungskapazität; M=Skala Merkfähigkeit

„PLG" Problemlösegütemaß siehe Text

B=unstandardisierter Regressionskoeffizient, β=standardisierter Regressionskoeffizient, B, β, R, R^2 und das adjustierte R^2 für das jeweilige Gesamtmodell (alle Prädiktoren)

Die Korrelation der Steuerungsleistungen in den beiden Szenarien schrumpfte nach Auspartialisierung der auf die Prädiktoren zurückzuführenden Varianz von r=.30 auf einen nicht signifikanten Zusammenhang von r=.16 zusammen. Es ergab sich somit auch in der vorliegenden Studie kein Hinweis auf eine eigenständige „Fähigkeit zum Problemlösen". Die mit den Szenarien reliabel erfaßten Informationen wurden durch Intelligenz und Wissen hinreichend repräsentiert.

Ein inhaltlich gleich zu interpretierender Befund ergab sich auch, wenn man die Korrelation von r=.30 (N=61; p<.05) zwischen der Steuerungsleistung in der „Schneiderwerkstatt" und der bei „DISKo" erstellten *Verhaltensbeurteilung* aus der Perspektive der dargestellten Analyse der Residualvarianzen betrachtete (Vorgehen

wie in Tabelle 18 dargestellt). Auch die Erhebung dieses „prozeßorientierten" Maßes erübrigt sich aus psychometrischer Perspektive, wenn man Intelligenz („Verarbeitungskapazität" und „Merkfähigkeit") sowie das Wissen berücksichtigt. (Die nicht-signifikante Korrelation zwischen der „DISKo"-Verhaltensbeurteilung und der Steuerungsleistung (neues Problemlösegütemaß) in der „Schneiderwerkstatt" betrug nach Auspartialisierung der genannten Prädiktoren noch r=.14.)

15.5 Zusammenfassung und Diskussion

Entgegen Dörner's Diktum, daß der Zusammenhang zwischen Intelligenz und Problemlösen so gering sei, *„daß er für jegliche Prognostik oder Diagnostik ohne Wert"* ist (Dörner, 1989, S. 46), konnten auch in der vorliegenden Studie Intelligenz und Wissen als wesentliche und hinreichende Voraussetzungen der systematischen Varianz der Steuerungsleistung in beiden Szenarien bestätigt werden. Voraussetzung hierfür war die interne Validität der Indikatoren der Steuerungsleistung. Diesbezüglich konnte für das Szenario „DISKo" die Notwendigkeit und Korrektheit der in der Berliner Untersuchung (Süß et al. 1991) erarbeiteten Aufgabenanalyse sowie der dort geleisteten Definition eines neuen Problemlösegütemaßes repliziert werden. Den systemspezifisch erhobenen Wissensindikatoren kam in bezug auf die Steuerungsleistung eine szenarienübergreifende prognostische Reichweite zu. Für den Wissenstest zur „Schneiderwerkstatt" ließ sich anhand der Analyse der Folgeeffekte eine gewisse Erfahrungssensitivität nachweisen.

Die Annahme eines neuen Konstrukts, das Postulat einer über Intelligenz und Wissen hinausgehenden Problemlösefähigkeit, welche durch die Steuerungsleistungen an den Szenarien indiziert wird, erwies sich - wie schon in der Berliner und Mannheimer Untersuchung (Abschnitt 9.1.1.1) - als empirisch unbegründet.

Die Tatsache, daß die Steuerungsleistungen sich mit Ausnahme ihres Intelligenz- und Wissensanteils als systemspezifisch erwiesen, stellt den diagnostischen Einsatz der Szenarien deutlich in Frage. Die mit dem Szenario gemessene Leistung verweist zu einem kleinen Teil auf Fähigkeiten aus dem Konstruktbereich Intelligenz und Wissen, die mit entsprechenden Instrumenten präziser erfaßt werden können, und zum anderen Teil auf die spezifischen Szenarien selbst. Der eignungsdiagnostische Einsatz von computergestützten Problemlöseszenarien bedeutet somit, daß der Wahl eines Szenarios eine erhebliche ergebnisdeterminierende Wirkung zukommen kann, da das Abschneiden der Kandidaten bei einem Szenario - sieht man von den Intelligenz- und Wissensanteilen ab - nicht überzufällig mit dem Abschneiden in einem anderen Szenario zusammenhängt.

16. Retrograde und konkurrente Kriteriumsvalidierung

Hauptanliegen des Empirie-Teils der vorliegenden Arbeit ist es, einen Beitrag zur Klärung der Kriteriumsvalidität von Problemlöseszenarien zu leisten. Dabei geht es vor allem darum zu prüfen, ob diese neuen diagnostischen Instrumente gegenüber den herkömmlichen Verfahren der Intelligenz- und Wissensdiagnostik eine *inkrementelle* Validität aufweisen. Im Mittelpunkt der Kriteriumsvalidierung steht die *Vorhersage* beruflicher Leistungen, zuvor wird in Abschnitt 16.1 die retrograde und in Abschnitt 16.2 die konkurrente Kriteriumsvalidität anhand der Kritieren „Laufbahnprüfung", „dienstliche Beurteilung" und „Laufbahnstatus" untersucht.

16.1 Retrograde Kritieriumsvalidierung: Laufbahnprüfung gehobener Dienst und dienstliche Beurteilung

Zur retrograden Kriteriumsvalidierung wurde die Abschlußnote der Laufbahnprüfung sowie das Gesamturteil der dienstlichen Beurteilung herangezogen. Aufgrund der im Vergleich der Bundesländer teilweise bestehenden Unterschiede in den verwendeten Beurteilungsskalen sowie aufgrund möglicher länderspezifischer Unterschiede in der Auslegung des jeweiligen Beurteilungsmaßstabs (Milde/Strenge) wurden die Werte pro Bundesland z-transformiert[16].

Die zur retrograden Kriteriumsvalidierung herangezogene Laufbahnprüfung am Ende der Ausbildung/des Studiums dient der Feststellung, ob der Beamte von seinen Kenntnissen und Fähigkeiten her für die jeweilige Polizeilaufbahn befähigt ist. Für 99 Personen lag die Abschlußnote der Laufbahnprüfung zum gehobenen Dienst, für 58 Personen zusätzlich die Abschlußnote der Laufbahnprüfung zum mittleren Dienst vor. Die Häufigkeitsdifferenz erklärt sich zum einen durch 23 Personen, die als sogenannte „Direkteinsteiger" in den gehobenen Dienst keine Laufbahnprüfung für den mittleren Dienst absolviert hatten, zum anderen dadurch, daß für 18 Personen

[16] Die beiden Beamten des Landes Bremen wurden dabei der „Gruppe Niedersachsen" zugeschlagen. Für ein Bundesland wurden unterschiedlich skalierte Werte übermittelt, so daß zwei Gruppen gebildet werden mußten.

die entsprechenden Werte nicht ermittelt werden konnten. Aufgrund dieses möglicherweise systematischen Datenverlusts hinsichtlich der Laufbahnprüfung für den mittleren Dienst wurde auf eine weitere Berücksichtigung dieses Indikators verzichtet. Für siebzig der 99 Personen mit Daten zur Laufbahnprüfung „gehobener Dienst" konnte der Zeitpunkt der Laufbahnprüfung in Erfahrung gebracht werden. Diese Laufbahnprüfung fand demnach im Durchschnitt zehn Jahre und einen Monat vor der hier thematisierten Studie statt (Median=10; SD=4,50). Da die z-transformierten Werte wenig anschaulich sind, eine untransformierte Zusammenschau der Werte verschiedener Bundesländer aber nicht gerechtfertigt erscheint, beschränken sich die folgenden Angaben zur Verteilung auf die größte länderhomogene Gruppe der 58 niedersächsischen Beamten. 3,5% dieser Beamten erzielten in der Laufbahnprüfung die Note „ausreichend", 58,6% die Note „befriedigend" und 37.9% die Note „gut" (Mittelwert=2,65; Median=3; SD=0,55). Die geringe Streuung der in der „Laufbahnprüfung" erzielten Noten begrenzt die Vorhersagbarkeit dieses Kriteriums. Die Variable „Laufbahnprüfung gehobener Dienst" erfüllte die Voraussetzung der Normalverteilung nicht (siehe oben, Abschnitt 12.4). Die in Tabelle 21 für diese Variable berichteten Prädiktor-Kriterium- Zusammenhänge wurden daher nur auf Rangdatenniveau berechnet (Spearman-Rangkorrelationen).

Als weiteres retrogrades Kriterium wurde die dienstliche Beurteilung berücksichtigt. Dieser Beurteilung liegen formalisierte Beurteilungssysteme zugrunde, die sich über die Bundesländer hinweg und innerhalb der Bundesländer möglicherweise auch über die Zeit hinweg unterscheiden, denen es aber gemeinsam ist, über die berufliche Eignung des zu Beurteilenden Auskunft zu geben. Beispiele für die der Beurteilung u.a. zugrunde gelegten Befähigungsmerkmale finden sich weiter unten (Kapitel 17, Tabelle 22). Die dienstliche Beurteilung wird regelmäßig erstellt, aus diesem Grunde variierte die Anzahl der pro Beamten vorliegenden Beurteilungen in Abhängigkeit von den Dienstjahren. Auf das Gesamturteil der letzten Beurteilung konnte für 95 Personen zurückgegriffen werden. Für 69 Personen dieser Gruppe lagen Angaben zum Erstellungszeitpunkt vor, die letzte Beurteilung erfolgte demnach im Durchschnitt ein Jahr und 11 Monate vor der hier thematisierten Untersuchung (Median=2; SD=1,55). Eine zusätzliche vorletzte Beurteilung konnte bei 85 Personen berücksichtigt werden, diese Beurteilung wurde - den Angaben von 61 Personen dieser Gruppe zufolge - im Durchschnitt vier Jahre und ein Monat vor der Untersuchung erstellt (Median=4; SD=2,11). Nur bei 20 Personen war eine dritte (vorvorletzte) dienstliche Beurteilung verfügbar, auf diese Information wurde aufgrund der geringen Gruppengröße verzichtet.

Für die 40 Polizisten, die sich aktuell in der Aufstiegsbewerbung befanden (siehe oben, Abschnitt 12.1) konnte außerdem noch eine zusätzliche Beurteilung berücksichtigt werden, die in Form eines relativ aktuellen „Eignungsberichts" vorlag.

Dem Vorgehen bei der Note der Laufbahnprüfung entsprechend, beschränken sich die Angaben zu den Kennwerten der Verteilungen der drei Beurteilungen in Tabelle 19 auf die größte länderhomogene Gruppe der niedersächsischen Polizisten. Die für die Gesamtgruppe berechneten z-transformierten Werte sind wenig anschaulich, eine untransformierte Zusammenschau der Werte verschiedener Bundesländer mit verschiedenen Beurteilungsskalen erscheint wenig sinnvoll. Die Verteilung der Punktwerte für die in der Tabelle analysierten niedersächsischen Polizeibeamten zeigt, was auch für die Gesamtgruppe gilt: die Beurteilungsskala (in Niedersachen 1-15 Punkte) wird nicht ausgenutzt; es zeigte sich eine deutliche Präferenz für vier positive Beurteilungskategorien (11 bis 14 Punkte), die verbal mit der Notenstufe „gut" und „sehr gut" verknüpft sind. Eine selektionsbedingte Varianzeinschränkung war nur für die Eignungsberichte zu erwarten. Eignungsberichte werden nur für die Aufstiegsbewerber und somit für eine nach Leistung (selbst-)selektierte Gruppe erstellt. Die dienstliche Beurteilung umfasst hingegen auch „normale" Polizeibeamte, für die nicht von vornherein mit einer eingeschränkten Varianz der Beurteilungen gerechnet werden mußte. Das Differenzierungspotential der Beurteilungen und der Eignungsberichte ist durch die eingeschränkte Varianz in Frage gestellt.

Tab. 19: Deskriptive Kennwerte zur Verteilung der Punktkategorien der Beurteilung
Subgruppenanalyse: Nur Beamte aus Niedersachsen

	Gesamtbeurteilung in Punkten (in den Zellen: Häufigkeit in Prozent)						M=Mittelwert SD=Standard-abweichung		
		„gut"			„sehr gut"				
	<11	11	12	13	14	15	M	SD	N
Eignungsbericht			12.5	65	20	2.5	13.1	.65	40
letzte Beurteil.	1.75[1]	7.0	31.6	35.1	22.8	1.75	12.7	1.3	57
vorletzte Beurteil.	4.1[2]	10.2	51	24.5	10.2		12.2	1.1	49

[1)] Eine Person mit dem Punktwert „6"; [2)] eine Person mit dem Punktwert „9"

Tabelle 20 zeigt die Spearman-Ranginterkorrelationen der Beurteilungen. Zumindest die Korrelationen der letzten dienstlichen Beurteilung mit den übrigen beiden Indikatoren waren substantiell, so daß eine Aggregation der Skalen gerechtfertigt erschien. Bei den Testanden, für die zu allen drei Indikatoren Angaben vorlagen, wurden alle drei Skalenwerte aggregiert. Für die Polizisten, für die aufgrund der geringeren Anzahl an Dienstjahren erst eine Beurteilung erstellt wurde sowie für Teilnehmer, für die aufgrund der fehlenden Aufstiegsbewerbung kein Eignungsbericht vorlag, wurden die fehlenden Werte vor der Aggregation durch den einzig

Tab. 20: Spearman-Rangkorrelationen der Beurteilungen

	DB1	DB2	EB
Letzte dienstl. Beurteil. (DB1)		N=85	N=40
vorletzte dienstl. Beurteil. (DB2)	.62**		N=32
Eignungsbericht (EB)	.62**	.38*	

** p<.01; * p<.05; N: siehe grauhinterlegte Felder

validen Wert bzw. durch den Mittelwert aus den zwei validen Werten substituiert. Während die Variablen für die Einzelbeurteilungen (letzte und vorletzte) sowie für den Eignungsbericht nicht normalverteilt waren (in Tabelle 20 sind deshalb die Spearman-Rangkorrelationen berichtet), erwies sich die Variable des aggregierten Wertes gemäß des Kolmogorov-Smirnov-Tests als normalverteilt.

In Tabelle 21 sind die korrelativen Zusammenhänge für die Prädiktoren mit den drei zusammengefaßten dienstlichen Beurteilungen sowie mit der Laufbahnprüfung aufgeführt. Die negativen Vorzeichen der Korrelationen in der Reihe „Laufbahnprüfung" bedeuten lediglich, daß dieses Kriterium mit seiner Orientierung an Schulnoten anders gepolt ist als die dienstliche Beurteilung. Statistisch bedeutsame Zusammenhänge ergaben sich nur für die Intelligenzskalen sowie - in bezug auf die Laufbahnprüfung - für die allgemeinen Wirtschaftskenntnisse. Weder die Steuerungsleistung noch der bei „DISKo" zusätzlich berechnete Verhaltensindikator stand in einem statistisch bedeutsamen Zusammenhang mit den berücksichtigten Kriterien. Die aufgrund der Intelligenztests erzielte Treffsicherheit konnte in einem regre-

Tab. 21: Korrelation verschiedener Prädiktoren mit der durchschnittlichen Beurteilung (DB) sowie der Laufbahnprüfung für den gehobenen Dienst (LB)

	Intelligenz		AW	Wissen		„SWS"			„DISKo"		
	AI	K	Wirt	Dis.	Sw.	Ka.	PL	Ag.	Ka.	PL	P8
DB	.17[1]	.30**	.10[R]	.04	.12	.05	.02	.10	-.04	.07	.15
LB[R]	-.36**	-.20*	-.20*	-.00	-.14	-.15	-.08	-.10	-.16	-.04	.00
N	98 / 97			95		99/98			95		

N=Gruppengröße: vor dem Schrägstrich: für „DB", nach dem Schrägstrich: für „LB"
[R] Rangkorrelationen; * p<.05; ** p<.01; [1]p<.10;
Intelligenz: „AI"=Allgemeine Intelligenz; „K"=Verarbeitungskapazität (Skalen)
„AW": Allgemeines Vorwissen, „Wirt"= Wirtschaftskenntnisse
„Wissen": Systemspezifisches Wissen:
Sachwissen über „DISKo" („Dis."), die „Schneiderwerkstatt" („Sw.")
„SWS": Szenario „Schneiderwerkstatt"; „DISKo" = Szenario „DISKo"
„Ka.": Kapitalendwert; „PL"=neues Problemlösegütemaß;
„P8": PL über die ersten 8 Takte; „Ag"=Aggregat („SWS-PL" und „DISKo-P8")

sionsanalytischen Vorgehen durch die zusätzliche Berücksichtigung eines weiteren Prädiktors oder mehrerer Prädiktoren (Steuerungsleistung und / oder systemspezifisches Wissen) nicht signifikant gesteigert werden.

16.2 Konkurrente Kriteriumsvalidierung: Laufbahnstatus

Die Berufstätigkeit im Polizeivollzug umfaßt Ämter des mittleren, gehobenen und höheren Dienstes. Der prozentual kleinste Anteil der Beamten des höheren Polizeivollzugsdienst wird nach dem Prinzip der Bestenauslese überwiegend aus dem gehobenen Dienst rekrutiert. Die Zulassung zum höheren Dienst kann - bei allen weiter unten (Kapitel 17) dargestellten Vorbehalten gegenüber Karriereindikatoren im öffentlichen Dienst - somit als ein Erfolgsindikator für den Polizeiberuf angesehen werden, auch wenn nicht alle Polizeibeamten des gehobenen Dienstes sich um den Aufstieg in den höheren Dienst bemühen. Während die Beförderung *innerhalb* einer Laufbahn früher oder später fast immer erfolgt, stellt der Aufstieg von einer Laufbahn in die nächsthöhere ein seltenes und herausragendes berufliches Ereignis dar. Äußeren Rahmenbedingungen, beispielsweise der dienststellenbezogene Beförderungssituation, kommt bei der Aufstiegsentscheidung eine vergleichsweise geringere Bedeutung bei. Auch das leistungsunabhängige „Hilfskriterium" Dienstalter ist hier weniger ausschlaggebend als bei Beförderungen innerhalb einer Laufbahn.

Von den Untersuchungsteilnehmern waren 36 als Ratsanwärter für die Ausbildung zum höheren Dienst zugelassen. Zwanzig dieser Beamten waren bereits zum Zeitpunkt der Untersuchung Ratsanwärter, 16 Personen wurden erst nach der Erhebung zur Ausbildung an der Polizeiführungsakademie zugelassen. Bezogen auf die Validierung stellt der Laufbahnstatus bei der zuerst genannten Gruppe ein Kriterium für eine konkurrente, bei der anderen Gruppe ein Kriterium für eine prädiktive Validierung dar. Eine solche Differenzierung würde allerdings die Gruppengröße mit ohnehin ungleicher Aufteilung auf die beiden Ausprägungen des dichotomem Kriteriums überstrapazieren. Es wurde daher eine Analyse auf der Ebene der Gesamtgruppe vorgenommen, deren Ergebnisse konservativ als Hinweise auf die konkurrente Validierung interpretiert werden. Da der höhere Dienst günstigere Besoldungsstufen vorsieht als der gehobene Dienst entspricht das Kriterium „Laufbahnstatus" einer dichotomisierten Variante des häufig verwendeten Kriteriums „Einkommen".

Der Laufbahnstatus (gehobener versus höherer Dienst) kann nicht für die Validierung der Intelligenztests herangezogen werden, da diese Tests bei der untersuchten Gruppe als ein Entscheidungskriterium für die Zulassung zum höheren Poli-

zeivollzugsdienst gelten. Dies bedingt zwangsläufig eine Korrelation zwischen Intelligenz und Laufbahnstatus, die nicht als Nachweis der Kriteriumsvalidität verbucht werden kann. Die punkt-biseriale Korrelation zwischen der „Allgemeinen Intelligenz" und dem Laufbahnstatus betrug $r_{pbis}=.51$ ($N=103$; $p<.01$). Die Tatsache, daß der Laufbahnstatus nicht unabhängig ist von der Intelligenz betrifft auch die Kriteriumsvalidierung der übrigen Variablen. Als Validitätsnachweis können nur solche Zusammenhänge gewertet werden, die nach statistischer Kontrolle der Einflußnahme der Intelligenz noch bedeutsam waren. Es wurden die Korrelationen zwischen allen in Tabelle 21 aufgelisteten Prädiktoren (mit Ausnahme der Intelligenzskalen) und dem Laufbahnstatus bestimmt, wobei die Skalenleistung „Allgemeine Intelligenz" auspartialisiert wurde. Lediglich für das systemspezifische Sachwissen über die „Schneiderwerkstatt" sowie für das Aggregat der beiden systemspezifischen Wissensindikatoren (Sachwissen über die „Schneiderwerkstatt" und über „DISKo") ergab sich mit $r=.32$ (Wissen über die „Schneiderwerkstatt"; $p<.01$, $N=100$) und $r=.29$ (Aggregat der Wissensindikatoren; $p=.29$; $N=96$) ein Zusammenhang zum Kriterium, der auch nach Auspartialisierung der „Allgemeinen Intelligenz" noch auf dem 5% Niveau gegen den Zufall abgesichert war.[17]

16.3 Zusammenfassung und Diskussion

Lediglich die Intelligenzskalen konnten einen bedeutsamen Beitrag zur zeitlich rückwärtsgewandten „Vorhersage" einiger Kriterien der Bewährung sowohl in der Ausbildung als auch im Beruf leisten. Für den Erfolg in der Laufbahnprüfung zeigte sich außerdem eine schwache Beziehung zu den durchschnittlich ein Jahrzehnt später erhobenen Wirtschaftskenntnissen. Beachtlich ist, daß das systemspezifische Sachwissen einen gegenüber der diesbezüglich kontaminierten Intelligenz inkrementellen Beitrag zur „Vorhersage" der konkurrenten hierarchischen beruflichen Positionierung leisten konnte. Für die Indikatoren der Problemlöseleistung ergaben sich keinerlei Hinweise auf eine retrograde oder konkurrente Kriteriumsvalidität. Bei der Interpretation der Ergebnisse gilt es zu beachten, daß die Kriterien über ein nur mäßiges Differenzierungspotential verfügten.

---

[17] Trotz der dichotomen Kriteriumsvariable wurden Produkt-Moment Korrelationen bestimmt, da zunächst nur die Stärke der Beziehung, nicht jedoch ihre Richtung, analysiert werden sollte. In diesem Falle kann auch die Produkt-Moment Korrelation interpretiert werden. Die Berechnung der punkt-biserialen Korrelation stellt lediglich eine Vereinfachung der Formel zur Berechnung der Produkt-Moment Korrelation dar (siehe Bortz, 1989, S. 270).

17. Prädiktive Kriteriumsvalidierung

Im Mittelpunkt der vorliegenden Studie stand eine - zeitlich nachgeordnete - Vorgesetztenbefragung über die von den Polizisten im beruflichen Alltag gezeigten Leistungen. Die Befragung mußte dabei inhaltlich so gestaltet werden, daß sie ein relevantes und realitätsnahes Bewährungskriterium darstellt. Um die Relevanz des gewählten Kriteriums nachvollziehen zu können, muß man mit den Instrumentarien der Personalsteuerung *im öffentlichen Dienst* vertraut sein. Während es sich bei anderen Berufsgruppen empfiehlt, berufliche Leistungen und Erfolg über sogenannte „harte" Kriterien zu operationalisieren, stellen solche Kriterien wie z.B. das Gehalt oder die hierarchische Stellung im öffentlichen Dienst vergleichsweise schwache Indikatoren beruflicher Leistung dar. Die Bezahlung richtet sich hier nach der Dienstpostenbewertung, nicht nach der individuellen Leistung. Die Beförderung innerhalb einer Laufbahn hängt u.a. von formalen Rahmenbedingungen (Anzahl freier Planstellen) ab und wird de facto darüber hinaus vom Dienstalter wesentlich mitbestimmt. Demgegenüber ist die im öffentlichen Dienst überaus bedeutsame *dienstliche Beurteilung* ein vergleichsweise unmittelbarerer Indikator der beruflichen Leistung. Positive Beurteilungen sind darüber hinaus in der Regel eine notwendige - obgleich nicht hinreichende - Voraussetzung für Beförderungen. Ein wesentliches Merkmal dienstlicher Beurteilungen ist, daß es sich dabei gemäß der Rechtsprechung nicht um mechanische Einzelmessungen, sondern um einen Akt „wertender Erkenntnis" handelt. Studien zur Vorhersage des Berufserfolgs im öffentlichen Dienst müssen sich an den dort herrschenden Gegebenheiten orientieren, auch wenn die Vorhersagbarkeit der Kriterien mit deren Subjektivitätsgrad abnimmt. Es macht keinen Sinn, psychometrisch hochwertigere, aber für den tatsächlichen Berufserfolg im öffentlichen Dienst irrelevante Kriterien zu bemühen. Die verwendeten Kriterien müssen die Grundlage der Berufserfolgsentscheidungen im öffentlichen Dienst repräsentieren. Es galt daher, die Praxis der dienstlichen Beurteilungen mit einem Fragebogen nachzuvollziehen *und* eine möglichst hohe psychometrische Qualität zu gewährleisten. Die hier zu Forschungszwecken eingeholte Beurteilung sollte sich durch die Akualität, durch die bessere Erläuterung der erfragten Merkmale sowie durch die Verwendung von Skalen (anstelle von Single-Items) positiv von der „originalen" dienstlichen Beurteilung (Abschnitt 16.1) unterscheiden. Außerdem sollte durch die Anonymisierung dem Milde-Effekt entgegengewirkt werden. Die vorliegende Untersuchung beruht - wie die meisten Studien zur Validität eignungsdiagnostischer Instrumente - auf Vorgesetztenbeurteilungen. Hossiep (1995, S. 51)

berichtet die Ergebnisse mehrerer Auszählungen, denen zufolge in 70% bis 72% der berufsbezogenen Validitätsstudien Beurteilungsdaten als Kritierien für die Verfahrensüberprüfung herangezogen werden. Dabei werden im „wirklichen Leben" ebenso wie in den Studien häufig die Beurteilungen verschiedener Personen miteinander verglichen, obgleich diese Beurteilungen (in großen Organisationen: zwangsläufig) von jeweils anderen Beurteilern ausgesprochen wurden. Die dienstliche Beurteilung erschöpft sich - entgegen der Erwartung von Personen, die nicht mit dem öffentlichen Dienst vertraut sind - nicht in einem Urteil darüber, wie der Beamte die Aufgaben seines *konkreten* Dienstposten erfüllt hat. Um einen dienstpostenübergreifenden Vergleich zu ermöglichen, erstreckt sich die dienstliche Beurteilung vielmehr explizit auch auf die Beurteilung sogenannter „allgemeiner Fähigkeiten". Die für die Mehrheit der Untersuchungsgruppe gültige Beurteilungsrichtlinie für den Polizeivollzugsdienst des Landes Niedersachsen sieht z.B. die Befähigungsmerkmale „Geistige Beweglichkeit" und „Merkfähigkeit" vor (vgl. Tabelle 22). Daneben gibt es auch Befähigungsmerkmale, die im nicht-kognitiven Bereich anzusiedeln sind, beispielsweise die „Kontaktfähigkeit". Da es in der vorliegenden Studie um die Validierung von Instrumenten zur *Fähigkeitsdiagnostik* geht, standen bei der Befragung die kognitiven Aspekte im Vordergrund. Für den „Berufserfolg" sind aber natürlich nicht nur kognitive Aspekte entscheidend. Deshalb wird in der Studie zusätzlich auch das Kriterium „Kooperationsfähigkeit" berücksichtigt, welches bei der empirischen Kontrolle der Aussagekraft von Instrumenten zur *Fähigkeitsdiagnostik* allerdings lediglich zur diskriminanten Validierung genutzt werden kann.

Um die Relevanz der durchgeführten Vorgesetztenbefragung für das interessierende Kriteriumsverhalten (berufliche Leistung und Berufserfolg) weiter abzusichern, wurden die Vorgesetzten bei allen erhobenen Beurteilungen zusätzlich um eine direkte Einschätzung der Bedeutung des jeweiligen Merkmals für die erfolgreiche Bewährung des beurteilten Mitarbeiters gebeten. Die entsprechenden Ergebnisse (siehe Abschnitt 17.2.2) ermöglichen eine empirisch fundierte Abschätzung der Relevanz der in der vorliegenden Studie verwendeten Kriterien.

Tab. 22: Merkmale der Befähigungsbeurteilung (laut der Beurteilungsrichtlinie für den Polizeivollzugsdienst des Landes Niedersachsen vom 4.1.1996)

Auffassungsgabe	Selbstständigkeit	Durchsetzungsvermögen
Merkfähigkeit	Konzeptionelles Arbeiten	Kritikfähigkeit
Denk- u. Urteilsfähigkeit	Organisatorische Fähigk.	Kooperationsfähigkeit
Geistige Beweglichkeit	Schriftl. Ausdrucksfähig.	Führungsfähigkeit
Entscheidungsfähigkeit	Mündl. Ausdrucksfähig.	Zuverlässigkeit
Aufgeschlossenheit	Kontaktfähigkeit	Belastbarkeit
Einfallsreichtum	Verhandlungsgeschick	

17.1 Beurteiler und Beurteilte

17.1.1 Deskriptive Angaben zu den beurteilten Personen

Wie bereits konstatiert und begründet (Abschnitt 12.1) wurden für die Untersuchung Personen aus unterschiedlichen Ebenen der hierarchisch strukturierten Organisation „Polizei" berücksichtigt, wobei für die höchste Stufe, den höheren Dienst, lediglich sogenannte „Anwärter" gewonnen werden konnten. Ein Teil dieser Subgruppe im Umfang von 26 Personen[18] hatte zum Zeitpunkt der Kriterienerhebung ihre Berufstätigkeit für ein Studium an der Polizeiführungsakademie unterbrochen. Entsprechend konnte für diese Gruppe keine Befragung zum *aktuellen beruflichen* Verhalten vorgenommen werden. Die alternativ durchgeführte Befragung der Dozenten an der Polizeiführungsakademie kann und soll nicht mit der Beurteilung beruflicher Leistungen gleichgesetzt werden. Zum einen konnten die Dozenten lediglich das Verhalten im Studium beurteilen, welches sich deutlich von der hier interessierenden *beruflichen* Tätigkeit unterscheidet und kein direktes Berufserfolgskriterium darstellt. Zum anderen dürften die Ergebnisse der Dozentenbefragung aufgrund der kurzen „Arbeitsbeziehung" zwischen Dozent und Student (Durchschnitt=10,4 Monate, Median=12 Monate; SD=2,02) ein qualitativ minderwertiges Kriterium abgeben (siehe Abschnitt 9.2.1, Seite 144). Der Mindestwert für die Erstellung tauglicher Erfolgskriterien (eine „Arbeitsbeziehung" von ca. zwei Jahren Dauer) wird damit deutlich unterschritten. Aufgrund dieser unzureichenden Dauer der Arbeitsbeziehung und aufgrund der Konzentration auf die Vorhersage des *Berufserfolges*, beschränkt sich die folgende Analyse auf die Beamten, die zum Befragungszeitpunkt im aktiven Polizeidienst waren. Nimmt man entsprechend die 26 Studenten der Führungsakademie aus der Gesamtgruppe aus, verbleiben 78 Untersuchungsteilnehmer. Für 73 dieser Personen lagen Befragungsergebnisse vor, dies entspricht einer Rücklaufquote von 93,6%. Unter den 73 Personen zwischen 28 und 57 Jahren (Median =36; SD=6,45) waren fünf Polizistinnen (6,8%). Die Kriminalpolizei war mit 37 Beamten annähernd gleich häufig repräsentiert wie die Schutzpolizei (36 Personen, inkl. Wasserschutzpolizei). Es handelte sich um Polizisten aus Niedersachsen (63%), Schleswig-Holstein (27,4%) und Nordrhein-Westfalen (9,6%). 97% der Teilnehmer machten Angaben zur Dauer ihrer Berufstätigkeit. Jede dieser Personen war demnach seit mindestens zehn Jahren Polizist (Median=18; SD=6,66).

---

[18] In Abschnitt 16.2 wurde zur konkurrenten Validierung die Aufstiegs*zulassung* als Kriterium herangezogen. 36 Teilnehmer hatten eine Aufstiegs*zulassung*, aber nur 26 Personen studierten bereits zum Zeitpunkt der Befragung.

17.1.2 Deskriptive Angaben zu den beurteilenden Personen

Die Beurteilung wurde überwiegend (87,7 %) von dem unmittelbaren Linienvorgesetzten der Untersuchungsteilnehmer abgegeben. Aus diesem Grunde wird der Begriff „Vorgesetztenbeurteilung" benutzt, obwohl es sich in den übrigen Fällen (11,3%) um Beurteilungen durch Kollegen der beurteilten Personen handelte. Die Beurteiler kannten die zu beurteilenden Personen im Durchschnitt seit sieben Jahren und vier Monaten (Median=5 Jahre; SD=sechs Jahre). Hauptsächlich hatte jeder Beurteiler den Fragebogen (siehe unten) für *eine* Person auszufüllen, lediglich drei Vorgesetzte hatten mit zwei (zweimal) bzw. vier zu beurteilenden Personen mehrere Bogen zu bearbeiten. Insgesamt wurden die 73 Polizisten also von 68 Personen beurteilt. Aus einer Laborperspektive betrachtet wäre es sicherlich wünschenswert, wenn alle Teilnehmer von *einem* Vorgesetzten beurteilt worden wären. Die Praxis in großen Organisationen sieht aber anders aus, und Personalauswahlverfahren sollen auch den Erfolg in großen Organisationen vorhersagen, bei denen verschiedene Personen an Karriereentscheidungen mitwirken. Mit 95,6 % war der größte Anteil der Beurteiler männlich, das Alter der Beurteiler variierte zwischen 33 und 59 Jahren (Median=48; SD=7,13). Alle Beurteiler waren Polizisten, etwas mehr als der Hälfte der Beurteiler waren Angehörige des höheren Polizeivollzugsdienstes. Die Beurteiler wurden nicht darüber informiert, welche Ausprägungen die zu beurteilte Person auf den Prädiktorvariablen erzielt hatte.

17.2 Die Befragung zu spezifischen Aspekten der beruflichen Leistung

Die Befragung der Beurteiler vollzog sich über den Postweg. Der in Abbildung 10 auszugsweise und im Anhang (Abbildung 15) vollständig dargestellte Fragebogen und ein Informationsblatt (siehe weiter unten und Abbildung 9) wurden gemeinsam mit einem Anschreiben, mit der Einverständniserklärung der zu beurteilenden Person sowie mit einem frankierten und adressierten Rückumschlag postalisch an die Beurteiler übersandt. Das Anschreiben erhielt eine Information zum Forschungsprojekt sowie die Bitte um Mitarbeit. Außerdem wurde im Begleitbrief die Anonymisierung der Daten zugesichert, und es wurde darauf hingewiesen, daß die Beurteilung ausschließlich zur Validierung der Auswahlverfahren erhoben wird. Auf dem Fragebogen war die Code-Nummer des zu Beurteilten eingetragen, zusätzlich war der Name des Beurteilten mit einer - vor der Rücksendung zu entfernenden - Haft-Notiz vermerkt. Wie in Abschnitt 17.1.1 erwähnt, betrug die Rücklaufquote 93,6%.

Die Beurteilungen wurden durchschnittlich ein Jahr und siebeneinhalb Monate nach der Prädiktorerhebung abgegeben (Median = zwanzig Monate; SD = 4,5 Monate).

Das im Anhang (Abbildung 15) abgebildete Erhebungsinstrument lieferte insgesamt Informationen zu drei Themenkomplexen: (1) Angaben zur beurteilenden und zur beurteilten Person sowie zu dem dienstrechtlichen Verhältnis zwischen beiden, (2) Angaben zur Bedeutung der jeweils erfragten Fähigkeits- und Leistungsdimension für die erfolgreiche Bewährung im konkreten Arbeitsgebiet der beurteilten Person sowie (3) Angaben zur Qualität der im Beruf gezeigten Fähigkeiten und beruflichen Leistungen / Verhaltensweisen der beurteilten Person. Da es um die Validierung von Intelligenztests und computergestützten Problemlöseszenarien ging, wurden solche im Berufsalltag gezeigten Fähigkeiten, Leistungen und Verhaltensweisen thematisiert, welche den Konstrukten „Intelligenz" und „Problemlösefähigkeit" zugeordnet werden können. Der Berufserfolg stellt ein komplexes Kriterium dar. Komplexe Kriterien bedürfen nach Jäger (1986, S. 281 f.) der Zerlegung in Varianzkomponenten, und als Validitätskriterium für Instrumente zur Fähigkeitsdiagnostik können nur die kognitiven Komponenten gelten. Um Daten für eine diskriminante Validierung zu gewinnen wurde außerdem noch die im Beruf gezeigte „Kooperationsfähigkeit" berücksichtigt. Dem Beurteilungsbogen war ein Informationsblatt mit Definitionen der drei interessierenden Verhaltensdimensionen beigefügt (siehe Abbildung 9), welches ein einheitliches Verständnis der Beurteiler sichern sollte. Die Dimensionsbeschreibungen wurden verschiedenen Literatur-Quellen entnommen und teilweise für den hier verwendeten Zweck aufbereitet.

Dem ersten Abschnitt mit der Bitte um Angaben zur beurteilenden und zur beurteilten Person sowie zu dem Verhältnis zwischen beiden folgten zunächst Fragen auf der Ebene der allgemeinen Fähigkeiten, die den Beurteilern mit den in Abbildung 9 dargestellten Definitionen erläutert worden waren. Die Beurteiler sollten zunächst auf einer vierstufigen Skala für jede der drei Fähigkeiten (Intelligenz, Problemlösefähigkeit und Kooperationsfähigkeit) einschätzen, wie wichtig diese Fähigkeit für die erfolgreiche Bewährung im Arbeitsgebiet des Beurteilten ist. Anschließend sollten die Beurteiler für jede Fähigkeit angeben, ob der Beurteilte im Vergleich zu seinen Kollegen der gleichen Laufbahn hinsichtlich der beschriebenen Fähigkeiten eher (1.) zu den unteren 25%, zum (2.) schlechteren oder (3.) besseren „Mittelfeld" (jeweils 25%) oder zu den (4.) oberen 25% der Vergleichsgruppe gehört. (Single-Items zur Fähigkeitsausprägung, statistische Kennwerte: siehe Abschnitt I der Tabelle 23.)

Schließlich wurde noch ein ipsativer Vergleich der Fähigkeiten verlangt. Die Fähigkeiten wurden dazu paarweise geordnet dargeboten. Zu entscheiden war jeweils, welche der beiden genannten Fähigkeiten bei der beurteilten Person höher ausgeprägt ist. Dieses Itemformat hat sich nicht bewährt. So wurden die Paarvergleiche teilweise inkonsistent beantwortet. Als inkonsistent galt z.B. die Aussage

eines Beurteilers, die Problemlösefähigkeit sei höher ausgeprägt als die Intelligenz und die Intelligenz höher als die Kooperationsfähigkeit sowie schließlich die Kooperationsfähigkeit sei höher als die Problemlösefähigkeit. Auch der mit diesem Itemformat verbundene „Wahlzwang" muß im nachhinein als problematisch gewertet werden. Die Tatsache, daß die Beantwortung dieser Items (ipsativer Vergleich) von mehreren Beurteiler ausgespart wurde (was sich bei den übrigen Items nicht ereignete), ist möglicherweise auf den Wahlzwang zurückzuführen. Die Angaben zum ipsativen Vergleich blieben aus den genannten Gründen bei der hier dargestellten Auswertung unberücksichtigt.

Nachdem mit den ersten Fragen die Ausprägungen der Fähigkeiten ganz allgemein beurteilt werden sollten, ging es in den folgenden 15 Items um konkrete im Berufsalltag gezeigte Verhaltensweisen oder Leistungen, die als Indikatoren der Konstrukte „Intelligenz", „Problemlösefähigkeit" und „Kooperationsfähigkeit" gewertet wurden. Die Befragung gliederte sich dabei stets in zwei Teile. Zunächst ging es darum einzuschätzen, wie bedeutend die jeweilige Leistung für die erfolgreiche Bewährung im Arbeitsgebiet des Beurteilten ist. Die Antwortkategorien umfassten die vier Stufen von *„geringe"* (Bedeutung) über *„eher geringe"*, und *„eher hohe"* bis zu *„hohe"* (Bedeutung). Mit dem zweiten Teil des Items wurde gefragt, wie die beschriebene Leistung bei der zu beurteilenden Person ausgeprägt war. Die Beurteiler sollten den zu beurteilenden Polizisten mit Laufbahn-Kollegen vergleichen und überlegen, welchen Rangplatz er hinsichtlich seiner Leistungen einnimmt. Die einzelnen Punkte der sechsstufigen Ratingskala wurden verbal charakterisiert. Auf die Frage nach der Ausprägung der Leistung gab es Antwortkategorien von (1) „sehr schwache" Ausprägung über (2) „schwache", (3) „eher schwache", (4) „eher hohe", (5) „hohe" bis (6) „sehr hohe" Ausprägung. Erläuternd wurde angemerkt, daß ein Eintrag in der jeweiligen Kategorie bedeutet, der Polizist gehöre hinsichtlich der beschriebenen Leistung zur „weit unterdurchschnittlichen Gruppe", „zur leistungsschwachen Gruppe", „zum schlechteren Durchschnitt", „zum besseren Durchschnitt", „zur leistungsstarken Gruppe" oder zur „weit überdurchschnittlichen Spitzengruppe". Abbildung 10 zeigt an einem der „Problemlösefähigkeit" zugeordneten Item beispielhaft den Aufbau der Fragen in diesem Teil des Beurteilungsbogens. Dieser Itemtyp wird im Abschnitt II der Tabelle 23 unter der Überschrift „Skalen zu spezifischen beruflichen Leistungen oder Verhaltensweisen" behandelt.

Insgesamt wurde mit fünf Items nach Indikatoren für die im Berufsalltag gezeigte Problemlösefähigkeit und mit drei Items nach der im Berufsalltag gezeigten Kooperationsfähigkeit gefragt. Sieben Items zielten auf die Beurteilung von im Berufsalltag gezeigten Verhaltensweisen und Leistungen, die der Intelligenz zugeordnet waren. Von diesen sieben Items thematisierten vier Aspekte der „Verarbeitungskapazität" und je ein Item die „Gedächtnisleistungen", den „Einfallsreichtum" und

Intelligenz

Unter Intelligenz sei hier die Fähigkeit zur Erfassung und Herstellung von Bedeutungen, Beziehungen und Sinnzusammenhängen verstanden. Zur Intelligenz gehört die Fähigkeit

- ✓ Aufgaben zu lösen, die schlußfolgerndes Denken und das Heranziehen, Verfügbarhalten und sachgerechte Beurteilen von Informationen verlangen
- ✓ leichte Aufg. zu lösen, bei denen es vor allem auf ein hohes Arbeitstempo ankommt
- ✓ sich sprachliches, numerisches oder figurales Material aktiv einzuprägen und kurzfristig zu reproduzieren oder wiederzuerkennen
- ✓ möglichst viele und verschiedene „brauchbare" Ideen zu produzieren
- ✓ zum Rechnen und allgemein die Fähigkeit zum erfolgreichen Umgang mit Zahlen
- ✓ zum erfolgreichen Umgang mit sprachlichem Material
- ✓ zum erfolgreichen Umgang mit figuralem Material (z.B. Zeichnungen und Grafiken)

Problemlösefähigkeit

Bei der Problemlösefähigkeit geht es nicht so sehr um Einzelfähigkeiten, sondern um die Koordination der Einzelfähigkeiten, um eigenständige Lösungskonstruktionen, rückmeldungsabhängige Strategieanpassungen und um längerdauernde Verarbeitungsprozesse. Eine hohe Problemlösefähigkeit zeigt sich bei der Bewältigung unklarer Situationen, die komplex und vernetzt sind. Entscheidungen haben hier Haupt- und Nebenwirkungen, die möglicherweise erst später sichtbar werden. Solche Probleme sind grundsätzlich nicht statisch, sondern dynamisch, sie verändern sich in der Zeit.
Bei der Problemlösefähigkeit geht es um die Fähigkeit

- ✓ sich durch die Suche nach Informationen ein Bild von der Lage zu erarbeiten und die Informationen aufzunehmen und zu integrieren
- ✓ die Zielvorgaben zu konkretisieren und gegebenenfalls in Teilziele zu zerlegen
- ✓ die einzelnen Teilziele untereinander auszubalancieren
- ✓ strategisch zu denken und sich für bestimmte Maßnahmen und deren Abfolge zu entscheiden
- ✓ die Entscheidungen immer wieder flexibel an die erhaltenen Rückmeld. anzupassen
- ✓ durch ein geschicktes Selbstmanagement, Zeitdruck und Frustrationen zu bewältigen.

Kooperationsfähigkeit

Die Kooperationsfähigkeit zeigt sich im Umgang mit anderen. Es geht um die Fähigkeit, mit anderen Personen Kontakt aufzunehmen, sich auf ihre Gefühle und Bedürfnisse einzustellen und mit ihnen partnerschaftlich zusammenzuarbeiten.
Es geht um die Fähigkeit

- ✓ auf andere Menschen zuzugehen
- ✓ rasch Anschluß zu finden
- ✓ Einfühlungsvermögen und Anpassungsbereitschaft zu zeigen
- ✓ die eigene Wirkung auf andere realistisch einzuschätzen
- ✓ sich nicht auf Kosten anderer durchzusetzen
- ✓ die Zusammenarbeit zu fördern
- ✓ Konflikte zu schlichten und Konsensfähigkeit zu zeigen
- ✓ sich auf unterschiedliche Menschen einzustellen.

Abb. 9: Für die Beurteiler erstellte Definition der einzelnen Fähigkeiten

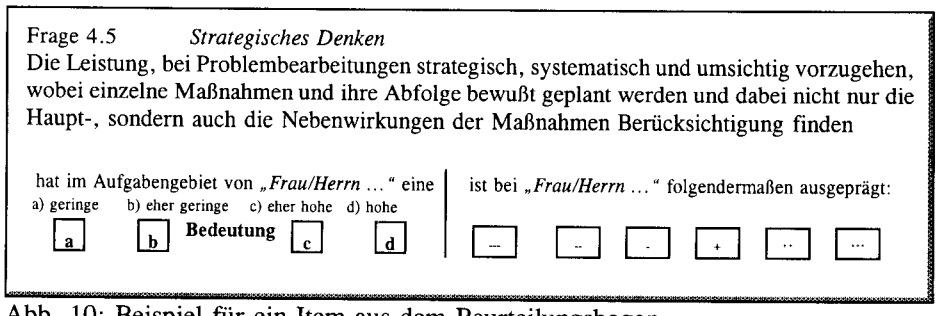


Abb. 10: Beispiel für ein Item aus dem Beurteilungsbogen

die „Effizienz bei der Erledigung von Routinetätigkeiten“. Die Itemreihung wurde per Zufall bestimmt, wobei allerdings das unmittelbare Aufeinanderfolgen von zwei Items zum gleichen Konstrukt unterbunden wurde. Die im Vergleich zur „Problemlösefähigkeit“ größere Anzahl an Items für das Kriteriumsverhalten „Intelligenz“ ergab sich aus der deutlicheren Binnendifferenzierung dieses Konstrukts. Neben den am Arbeitsplatz gezeigten Leistungen der „Verarbeitungskapazität“ sollten auch die Leistungen der „Allgemeinen Intelligenz“ Berücksichtigung finden.

17.2.1 Kriteriumsverhalten: Skalenbildung und Skalenkennwerte

Zunächst wurden die Items zu konkreten beruflichen Verhaltensweisen oder Leistungen theoriegeleitet zu Skalen zusammengefaßt. Die drei Items zur „Kooperationsfähigkeit“, die fünf Items zur „Problemlösefähigkeit“ und die vier Items zur „Verarbeitungskapazität“ wurden aggregiert. Als Skala für die am Arbeitsplatz gezeigten Leistungen der „Allgemeinen Intelligenz“ wurde je ein Item zu „Gedächtnisleistungen“, den „Einfallsreichtum“ und die „Effizienz bei der Erledigung von Routinetätigkeiten“ mit dem Mittelwert der Skala „Verarbeitungskapazität“ (die somit als Einzelitem dieser Skala berücksichtigt wurde) aggregiert.

In Tabelle 23 sind die so entstandenen Kriterienmaße mit einigen zugeordneten statistischen Kenngrößen abgebildet. Die interne Konsistenz (Cronbach's alpha) war mit Werten über .80 für alle Skalen zufriedenstellend. Der über den Skalenmittelpunkten liegende Durchschnittswert aller Skalen verdeutlicht, daß der bei Vorgesetztenbeurteilungen in der Regel auftretende Milde-Effekt auch in der vorliegenden Studie zu beobachten war. Die theoretischen Annahmen bei der Itemformulierung und die Höhe der Spearman-Rangkorrelationen[19] der Beurteilungen zur allgemeinen

[19] Aufgrund der nicht-normalverteilten Urteile zu den Einzel-Items (Abschnitt I in Tabelle 23) wurden die Zusammenhänge nur auf Rangdatenniveau berechnet.

Tab. 23: Skalen der Vorgesetztenbeurteilung

I. Single-Items zur Fähigkeitsausprägung, allg. Beschreibung; 4-stufige Skala								
	Items	M	Med	SD	Var.	Min	Max	
a) Intelligenz	1	3.41	3	.64	.41	2	4	
b) Problemlösen	1	3.36	3	.65	.43	2	4	
c) Kooperationsverh.	1	3.38	4	.76	.57	2	4	
II. Skalen zu spez. berufl. Leistungen oder Verhaltensweisen; 6-stufige Skala								
	Items	M	Med	SD	Var.	Min	Max	α
a1) Allg. Intelligenz	4*	4.83	4.75	.67	.45	3	6	.81
a2) Verarbeitungskap.	4	4.77	4.75	.73	.53	2.75	6	.87
b) Problemlösen	5	4.64	4.80	.77	.59	2.60	6	.85
c) Kooperationsverh.	3	4.74	4.67	1.0	1.0	2.33	6	.90

N=73 „M"=Mittelwert; „Med" = Median; „SD"=Standardabweichung;
„Var."=Varianz; „α"=Cronbach's alpha
*) Mittelwert der Skala „Verarbeitungskapazität" (a2) als Item berücksichtigt

Fähigkeitsausprägung (Single-Items in Abschnitt I der Tabelle 23) mit den Beurteilungen der dem jeweiligen Fähigkeitskonstrukt entsprechenden einzelnen Leistungen und Verhaltensweisen (Skalen in Abschnitt II der Tabelle 23) ließen mit r=.64 (Ia und IIa1 laut Tabelle 23), r=.55 (Ib und IIb) sowie r=.61 (Ic und IIc) eine weitere Aggregation dieser beiden Skalen gerechtfertigt erscheinen. Zwecks Gleichgewichtung und aufgrund der ungleichen Skalierung (vierstufige versus sechsstufige Skala) wurden die Beurteilungen vor der Aggregation z-transformiert. Die Skala „Verarbeitungskapazität" wurde nicht mit dem Einzelitem zur allgemeinen Einstufung der „Intelligenz" aggregiert, da die den Beurteilern zur Bearbeitung dieses Einzelitems an die Hand gegebene Beschreibung der Intelligenz (siehe Abbildung 9) explizit alle Intelligenzdimensionen umfaßte. Die Skala „Verarbeitungskapazität" wurde für die folgenden Auswertungen nur noch insofern berücksichtigt als das Skalenmittel gleichgewichtet mit den Items zur Beurteilung der „Merkfähigkeit", des „Einfallsreichtums" und der „Bearbeitungsgeschwindigkeit" zur Bildung der Skala „Allgemeine Intelligenz" herangezogen wurde.

Tabelle 24 zeigt die Interkorrelationen der drei Kriteriumsvariablen. (Lediglich das aggregierte Maß der Urteile über die „Kooperationsfähigkeit" erfüllte die Voraussetzungen der Normalverteilung nicht und wurde normalisiert.) Die Urteile über

Tab. 24: Interkorrelationen der Beurteilungen

Urteile üb. d. im Beruf gezeigte...	Int.	Probl.	Koop
Intelligenz (Ia und IIa laut Tab. 23)			
Problemlösef. (Ib u. IIb lt. Tab.23)	.71**		
Kooperationsf. (Ic u. IIc lt. Tab. 23)	.32**	.51**	

** $p < .01$; N: 73

die im beruflichen Alltag gezeigte Intelligenz waren mit r=.71 deutlich mit den Urteilen über die im Berufsalltag gezeigte Problemlösefähigkeit korreliert. Überlappungen zwischen Intelligenz und Problemlösen wurden in Abschnitt 9.1.2.1 bereits postuliert. Die hohe Korrelation der beiden Kriteriumsvariablen „Intelligenz" und „Problemlösen" indizierte die Bildung einer weiteren Kriteriumsvariablen, die das Aggregat dieser beiden z-transformierten Variablen darstellt. Somit ergaben sich insgesamt vier Kriteriumsvariablen: die drei in Tabelle 24 aufgeführten sowie das Aggregat aus den beiden in Tabelle 24 genannten Variablen „Intelligenz" und „Problemlösen". Im Gegensatz zu anderen Validierungsstudien, in denen u.a. Single-Items als Kriterien verwendet wurden, handelt es sich bei allen vier hier verwendeten Kriteriumsvariablen um Skalen, die mehrere Items umfassen und eine befriedigende interne Konsistenz aufweisen.

Diese vier Kriteriumsvariablen standen in keinem statistisch bedeutsamen Zusammenhang mit dem Aggregat der drei zusammengefaßten dienstlichen Beurteilungen (siehe Abschnitt 16.1). Die Spearman-Rangkorrelationen für den Zusammenhang der vier Kriteriumsvariablen mit der Laufbahnprüfung gehobener Dienst variierten zwischen -.24 ($p < .05$) und -.35 ($p < .01$). Der ausbleibenden Zusammenhang zwischen den (zeitlich zurückliegenden) dienstlichen Beurteilungen und der einige Zeit nach der Prädiktorerhebung eingeholten Beurteilung der im Berufsalltag gezeigten Fähigkeiten und Leistungen muß aufgrund des Zeitabstandes und aufgrund der inhaltlichen Unterschiede zwischen beiden Beurteilungen (z.B. spezifische Aspekte [Intelligenz, Problemlösen und Kooperationsfähigkeit] auf der einen und „Gesamturteil" auf der anderen Seite) nicht überraschen. Gleichwohl könnte man durch diesen Befund die Relevanz der Kriterien in Frage gestellt sehen. Deshalb wird im folgenden Abschnitt über die Relevanz der erfragten Merkmale berichtet.

17.2.2 Anforderungsanalytische Fundierung / Relevanz der erfaßten Kriterien

Um die Relevanz der durchgeführten Vorgesetztenbefragung für das interessierende Kriteriumsverhalten (berufliche Leistung) abschätzen zu können, wurden die Vorgesetzten bei allen erhobenen Beurteilungen zusätzlich um eine Einschätzung der Be-

deutung der jeweiligen Fähigkeit / Leistung für die erfolgreiche Bewährung im Arbeitsgebiet des beurteilten Mitarbeiters gebeten. Die Skalenbildung für diese Bedeutsamkeitseinschätzungen erfolgten analog zu dem Vorgehen für die Beurteilung der Leistungs- bzw. Fähigkeitsausprägung (Abschnitt 17.2.1). Lediglich bei der abschließenden Aggregation der Single-Items zur Bedeutsamkeit der auf allgemeiner Ebene beschriebenen Fähigkeiten einerseits und den Items zur Bedeutsamkeit der auf der spezifischen Ebene beschriebenen Leistungen oder Verhaltensweisen andererseits konnte aufgrund der identischen Itemformate auf die vorherige z-Transformation verzichtet werden. Tabelle 25 informiert über Kennwerte der Variablen.

Allen in der Kriterienstudie berücksichtigten Leistungen oder Verhaltensweisen wurde von den Beurteilern bescheinigt, daß sie für die erfolgreiche Bewährung im Arbeitsgebiet des Beurteilten eine „eher hohe" (Skalenausprägung „3") bis „hohe" (Skalenausprägung „4") Bedeutung haben. Für akademische Diskussionen über die praktische Relevanz der gewählten Kriterien bleibt daher wenig Raum.

Bei der vergleichenden Analyse der Relevanz der einzelnen Fähigkeiten ist zu berücksichtigen, daß die „Allgemeine Intelligenz" einzelne Merkmale umfaßt, die sich hinsichtlich ihrer Relevanz im Urteil der Vorgesetzten deutlich unterscheiden. Dies ließ sich anhand der Einzelitems einschätzen, bei denen spezifische Intelligenzleistungen beschrieben wurden. Denjenigen Leistungen, die der „Verarbeitungskapazität" zuzuschreiben sind, wurde eine höhere Bedeutung für den Arbeitsplatz beigemessen als den Leistungen im Kontext der Problemlösefähigkeit (Mittelwerte: siehe Tabelle 24; $t=2.47$; $p<.05$). Demgegenüber ist die Leistung, einfache Routinetätigkeiten mit hohem Arbeitstempo und gleichzeitig mit hoher Sorgfalt zu erledigen nach Ansicht der Vorgesetzten relativ unbedeutender für das Arbeitsgebiet der Beurteilten ($M=2.85$, $SD=.97$). Schließt man das Urteil über die Relevanz des Arbeitstempos und der Sorgfalt aus der Skala zur Relevanz der „Allgemeinen Intelligenz" aus, erhöht sich die interne Konsistenz dieser Skala auf .60.

Tab. 25: Relevanzeinschätzungen (Skalen)

	Einstufung der Bedeutsamkeit; Beschreibung auf...						
	allg. Ebene[1]		spezifische Ebene[2]			Aggregat	
	M	SD	M	SD	α	M	SD
Allg. Intelligenz	3.62	.54	3.14	.44	.45	3.38	.40
Verarbeitungskap.			3.36	.43	.56		
Problemlösefähigkeit	3.82	.42	3.22	.49	.69	3.52	.38
Kooperationsfähigk.	3.75	.52	3.11	.75	.79	3.43	.58

N=73; „M" = Mittelwert; „SD" = Standardabweichung; „α" = Cronbach's alpha;
[1)] Single-Item; [2)] Anzahl der Items: siehe Tabelle 23

Die Relevanzeinschätzungen zeigen, daß die in der vorliegenden Arbeit der prädiktiven Kriteriumsvalidierung zugrundegelegten Fähigkeiten und Leistungen für die erfolgreiche Bewährung im beruflichen Alltag eines Polizisten bedeutsam sind. Dieser Sachverhalt findet in der absoluten Höhe der Relevanzurteile seinen klaren Ausdruck. Damit ist eine wesentliche Voraussetzung für die Kriteriumsvalidierung, nämlich der Nachweis der Bedeutsamkeit der erhobenen Kriterienmaße, erfüllt. Die Ergebnisse sind eine empirische Bestätigung der Annahme, daß die Beschreibungen des Konstrukts „Problemlösen" Fähigkeiten betreffen, die auch im Berufsalltag von Personen in bestimmten Aufgabenbereichen vorkommen (siehe Kapitel 3.1). Allerdings korrespondierten auch die Konstruktannahmen zur „Intelligenz" mehr („Verarbeitungskapazität") oder minder („Bearbeitungsgeschwindigkeit") deutlich mit den im Berufsalltag eines Polizisten gestellten Anforderungen.

17.3 Intelligenz, Wissen und Problemlösefähigkeit als Prädiktoren beruflicher Leistung

In den folgenden vier Abschnitten werden die Ergebnisse zur Vorhersage beruflicher Leistung durch Intelligenz, Wissen und Problemlösen sowie zum Verhältnis der Prädiktoren untereinander berichtet. In Abschnitt 17.3.1 wird zunächst die Vorhersageleistung der Einzelprädiktoren betrachtet, wobei zusätzlich die Partialkorrelationen für die statistische Kontrolle eines oder zwei weiterer Prädiktoren Beachtung finden. Die Steigerung der Vorhersageleistung durch eine Kombination der Prädiktoren im Sinne einer multiplen Regression wird in Abschnitt 17.3.2 thematisiert. Gerechnet wurde eine hierarchische Regressionsanalyse, bei der aufgrund von theoretischen Annahmen über die Prädiktorreihenfolge entschieden wurde. Eine von theoretischen Setzungen unabhängige Analyse der relativen Vorhersageanteile der einzelnen Prädiktoren leistet die Kommunalitätenanalyse, über deren Ergebnisse in Abschnitt 17.3.3 berichtet wird. Den Abschluß dieses Kapitels bildet ein Strukturgleichungsmodell zur Vorhersage der beruflichen Leistungen in Abschnitt 17.3.4.

17.3.1 Zur Vorhersagekraft der Einzelprädiktoren

In der Tabelle 26 sind für einzelne Prädiktoren die korrelativen Zusammenhänge mit dem im Durchschnitt ein Jahr und siebeneinhalb Monate nach der Prädiktorerhebung abgegebenen Urteil über verschiedene Aspekte der im Berufsalltag gezeigten

Tab. 26: Korrelation einiger Prädiktoren mit dem Vorgesetztenurteil über die im Berufsalltag gezeigten Fähigkeiten/Leistungen in den Dimensionen Intelligenz („Int."), Problemlösen („Probl.") und dem Aggregat („Aggr.") dieser beiden Urteile sowie (zur diskriminanten Validierung) dem Urteil über die im Berufsalltag gezeigte Kooperationsfähigkeit („Koop.")

	Kriterien: Urteil über die im Beruf gezeigte...				
↘ **Prädiktoren**	Int.	Probl.	Aggr.	Koop.	N
Skala Allgemeine Intelligenz („AI")	.43**	.28*	.39**	.07	72
Nach statistischer Kontrolle von... „SWS"-Gesamtvermögen	.39**	.21[1]	.33**		
Sachwissen über „SWS"	.38**	.22[1]	.32**		
„SWS" und Sachwissen üb. „SWS"	.35**	.17	.28*		
„SWS" Gesamtvermögen	.31**	.37**	.37**	.19	73
Nach statistischer Kontrolle von... Skala Allgemeine Intelligenz („AI")	.21[1]	.31**	.29*		
Sachwissen über „SWS"	.26*	.33**	.32**		
„AI" und Sachwissen über „SWS"	.18	.28*	.25*		
„DISKo" Gesamtvermögen	-.01	-.17	-.10	-.09	69
„DISKo" neues PLG über 8 Takte	.16	.33**	.26*	.08	69
Nach statistischer Kontrolle von... Skala Allgemeine Intelligenz („AI")	.02	.26*	.16		
Sachwissen über „DISKo"	.14	.34**	.26*		
Sachwissen über „SWS"	.10	.28*	.21[1]		
„AI" und Sachwissen über „SWS"	-.02	.23[1]	.12		
„DISKo" Verhaltensbeurteilung	.26*	.17	.23*	.11	69
Nach statistischer Kontrolle von... Skala Allgemeine Intelligenz („AI")	.02	.02	.02		
Sachwissen über „SWS"	.29*	.26*	.30*	.06	73
Nach statistischer Kontrolle von... Skala Allgemeine Intelligenz („AI")	.21[1]	.20[1]	.23[1]	* $p<.05$; ** $p<.01$; [1] $p<.10$	
„SWS"-Gesamtvermögen	.24*	.19	.23[1]		
„AI" und „SWS"-Gesamtvermög.	.18	.15	.18		

Fähigkeiten und Leistungen aufgeführt. Für die Fälle, in denen sich überzufällige Zusammenhänge aufzeigen ließen, sind in der Tabelle zusätzlich Partialkorrelationen verzeichnet. Die Analyse der in der Tabelle dargestellten Partialkorrelationen vermittelt einen Eindruck von der Prädiktionskraft der Variablen nach statistischer Kontrolle eines oder zwei weiterer Prädiktoren.

Zunächst fällt auf, daß trotz der Ähnlichkeit ($r=.71$, siehe Tabelle 24) der Urteile über die im Beruf gezeigten Leistungen/Fähigkeiten aus dem Bereich der Intelligenz und aus dem Bereich des Problemlösens eine Differenzierung zwischen den beiden Kriterienarten möglich war, d.h. die Kriteriumsvalidität der Verfahren variierte in Abhängigkeit von dem jeweils vorhergesagtem Kriterium. Jeder Prädiktor hatte dabei seine individuelle Vorhersagestärke für den theoretisch ähnlichsten Kriterienbereich. Beispielsweise betrug die Validität der Skala „Allgemeine Intelligenz" $r=.43$, wenn das Urteil über die im Berufsalltag gezeigten Leistungen und Fähigkeiten aus dem Bereich „Intelligenz" als Erfolgsmaß herangezogen wurde. Gemessen am Urteil über Leistungen und Fähigkeiten aus dem Bereich „Problemlösen" belief sich die Validität der Intelligenztests hingegen auf $r=.28$. Die Differenz zwischen den beiden Korrelationen verfehlte nur knapp die Signifikanzgrenze ($t=1.81$, $p<.10$; Berechnung nach Cohen & Cohen, 1975, S. 53 f.). Umgekehrt konnte mit der Steuerungsleistung in den beiden Szenarien das Urteil über die im Beruf gezeigten Fähigkeiten und Leistungen aus dem Bereich des Problemlösens nominell besser vorhergesagt werden als das Urteil über die intelligenzbezogenen beruflichen Leistungen. Das für „DISKo" über die ersten acht Takte berechnete Problemlösegütemaß stand beispielsweise mit $r=.33$ in einem signifikant engeren Zusammenhang mit dem Vorgesetztenurteil über die im Beruf gezeigten Fähigkeiten und Leistungen aus dem Bereich des Problemlösens als mit dem entsprechenden Urteil über die im Berufsalltag gezeigten intellektuellen Fähigkeiten und Leistungen ($r=.16$; $t=2.05$, $p<.05$; Berechnung nach Cohen & Cohen, 1975, S. 53 f.).

Hinsichtlich des Szenarios „DISKo" ist anzumerken, daß sich in bezug auf die Steuerungsleistungen das „ursprüngliche" instruktionsgemäße Problemlösegütemaß, der Kapitalendwert, als vollständig unbrauchbar für die Prognose der Kriterien erwiesen hat. Dieser Befund reiht sich in die Kette der übrigen Belege für die Tatsache ein, daß es diesem Indikator für die hier analysierte Gruppe an interner Validität mangelte (siehe Abschnitt 13.2). Im „normalen" Anwendungsfall wären unter Umständen schwerwiegende Personalentscheidungen aufgrund dieses invaliden „Problemlösegütemaßes" getroffen worden. Erst die Berücksichtigung des neuen, intern validen Problemlösegütemaßes zeigt die Vorhersagekraft der bei „DISKo" erzielten Steuerungsleistungen. (Tabelle 26 verzeichnet das „optimale", über die ersten acht Bearbeitungstakte berechnete Maß. Mit dem über alle Bearbeitungstakte berechneten neuen Problemlösegütemaß konnten lediglich auf dem 10% Niveau signifikante Zu-

sammenhänge erzielt werden.) Hinzuweisen ist auf die Vorhersagekraft des bei „DISKo" standardmäßig berechneten Parameters zur Beurteilung der Verhaltensweisen und Strategien der Testanden. Dieser Zusammenhang war allerdings im wesentlichen auf die Intelligenz zurückzuführen. Die um den Intelligenzanteil bereinigte Partialkorrelation zwischen diesem Prädiktor und den Kriterien betrug lediglich r=.02. Die vermeintliche Prozeßdiagnostik führte also im vorliegenden Fall in ihrem verwertbaren, kriteriumsrelevanten Teil zum gleichen Ergebnis wie die „Endproduktdiagnostik" der Intelligenztests, ohne allerdings die Höhe der Validität des Intelligenztests zu erreichen. Die bei „DISKo" geleistete Verhaltensbeurteilung und die Beurteilung der Strategien der Testanden konnte auf der Grundlage der prädiktiven Kriteriumsvalidität weitgehend dem Intelligenzkonstrukt subsumiert werden. Das semi-quantitative Sachwissen über Variablenzusammenhänge im System „DISKo" erwies sich nicht als prädiktiv valide.

Im Gegensatz zum Szenario „DISKo" konnten bei der „Schneiderwerkstatt" mit dem den Zielvorgaben entsprechenden Standardproblemlösegütemaß, nämlich dem Endwert des Gesamtvermögens, die intendierten Kriterien prognostiziert werden.

Als prognostisch valide erwies sich auch das Sachwissen über die „Schneiderwerkstatt". Allerdings verfehlten die Partialkorrelationen zweiter Ordnung hier die Signifikanzgrenze. Das in der Tabelle 26 berücksichtigte Aggregat der beiden Skalen des Wissenstests zur „Schneiderwerkstatt" erzielte dabei eine nominell niedrigere Treffsicherheit als die Skala „Zusammenhänge verbal" alleine. Im Gegensatz zum systemspezifischen Wissen konnten die Kriterien mit dem ebenfalls zur Prädiktion erhobenen Wirtschaftskenntnissen nicht überzufällig vorhergesagt werden.

In ihren Relevanzeinschätzungen (siehe oben, Abschnitt 17.2.2) haben die Vorgesetzten zum Ausdruck gebracht, daß für den Erfolg im polizeilichen Berufsalltag *sowohl* Fähigkeiten und Leistungen aus dem Bereich der Intelligenz *als auch* aus dem Bereich der Problemlösefähigkeit bedeutsam sind. Demzufolge kann es nicht Aufgabe der Eignungsdiagnostik in diesem Bereich sein, Personen auszuwählen, die ihren Leistungsschwerpunkt ausschließlich entweder in der einen oder in der anderen Dimension haben. Gefragt sind vielmehr Auswahlverfahren, die eine Vorhersage beider Leistungsbereiche erlauben. Die Analysen in den nachfolgenden Abschnitten beschränken sich dementsprechend auf das aggregierte Kriteriumsmaß. Die Validität der Prädiktoren bei der Vorhersage dieses aggregierten Kriteriums läßt sich in Tabelle 26 in der Spalte mit den korrelativen Zusammenhänge zwischen Prädiktor und dem Aggregat der Urteile über die beruflichen Leistungen in den Dimensionen Intelligenz und Problemlösen ablesen. Diesbezüglich erwies sich die Skala „Allgemeine Intelligenz" als nominell bester Prädiktor. Berechnungen, die mit der Skala „Verarbeitungskapazität" durchgeführt wurden, zeigten vergleichbare, wenngleich nominell niedrigere Validitätskoeffizienten als sie für die Skala „Allgemeine

Intelligenz" berichtet wurden. Die Vorhersagekraft der Problemlöseleistung ließ sich nominell im stärkeren Ausmaß auf deren Intelligenz- und Wissensanteil zurückführen als umgekehrt die Vorhersagekraft der Intelligenz auf deren Problemlöse- und Wissensanteil (siehe hierzu Abschnitt 17.3.3). Die mit den Szenarien reliabel erfaßten kriteriumsrelevanten Informationen wurden also auch durch die Variablen Intelligenz und Wissen weitgehend repräsentiert.

Um einen Indikator für die diskriminante Validität zu gewinnen, wurden die Vorgesetzten um eine Einschätzung der Kooperationsfähigkeit der Polizisten gebeten. Erwartungsgemäß korrelierte kein Prädiktor bedeutsam mit diesem Kriterium.

17.3.2 Multiple Regressionsanalyse

In einem weiteren Analyseschritt wurde mit Hilfe der multiplen Regressionsanalyse geprüft, ob durch die Kombination der Prädiktoren die Validität der Vorhersage inkrementell gesteigert werden kann und welche Prädiktorkombination am aussagekräftigsten ist. Im Rahmen der Überlegung, welche Prädiktoren in der Regressionsgleichung berücksichtigt werden sollten, stellte sich heraus, daß die Vorhersagekraft der Problemlöseleistung durch die Berücksichtigung der Steuerungsleistung bei der „Schneiderwerkstatt" ausreichend repräsentiert war, und die zusätzliche Einbeziehung der „DISKo"-Steuerungsleistung die Validität dieser Prädiktorgruppe nicht verbesserte. Auf der Ebene des Sachwissens konnte sich die Analyse auf den ersten Aufgabentyp des Wissenstests zur „Schneiderwerkstatt" („Relationen/Zusammenhänge verbal"; siehe Abschnitt 12.2.2.1) beschränken. Durch die zusätzliche Einbeziehung des zweiten Aufgabentyps dieses Wissenstests (Variableneigenschaften) und/oder des Wissensindikators aus dem Programm „DISKo" konnte die Vorhersage des Kriteriums durch die Prädiktorgruppe „Wissen" nicht gesteigert werden. Die Intelligenzleistungen wurden in Form der Skala „Allgemeine Intelligenz" in die Prädiktion einbezogen. Diese Skala korrelierte bei der analysierten Gruppe im Umfang von 72 Personen zu $r = .26$ ($p < .05$) mit dem Kapitalendwert in der „Schneiderwerkstatt" und zu $r = .25$ ($p < .05$) mit der berücksichtigten Wissensskala. Die beiden zuletzt genannten Prädiktoren waren zu $r = .36$ ($p < .01$) miteinander assoziiert.

Wie im vorherigen Abschnitt erläutert, wurden die Analysen für das Kriterium „berufliche Leistungen" auf der Ebene der über beide Leistungsdimensionen (Intelligenz und Problemlösen) aggregierten Vorgesetztenbeurteilungen durchgeführt. In der Analyse wurde zunächst die Skala „Allgemeine Intelligenz", dann der Indikator für das systemspezifische Wissen über die „Schneiderwerkstatt" sowie schließlich die Steuerungsleistung (Endwert Gesamtvermögen) bei der „Schneiderwerkstatt" in die Vorhersagegleichung eingeführt.

Tab. 27: Hierarchische Regression von Intelligenz, Wissen und Problemlösen auf das Kriterium (Urteil berufliche Fähigkeiten/Leistungen, die den Bereichen Intelligenz und Problemlösen zugeordnet werden können)

	Kriterium: berufliche Leistung					Kreuzvalidierung:		
↓ Prädiktoren ↓	r	R	B	β	R_{diff}		Gruppe	
							1	2
Allg. Intelligenz	.39**	.39**	.635	.285				
Wissen („SWS“)[1]	.33**	.46**	.127	.187	.07*	R	.53	.49
Kapitalend. „SWS“	.36**	.50**	.389	.215	.04	R^2_{korr}	.21	.17
	p<.01;*p<.05; Intercept=-1.30 N=72 R^2= .25 F (3,68)=7.53 R^2_{korr} .22					N	34	38

[1] Wissen über d. „SWS“: Skala „Zusammenhänge verbal“; „SWS“ = „Schneiderwerkstatt“
β=standardisierter, B=unstandardisierter Regressionskoeffizient;
B, β, R, R2 und das adjustierte R2 für das jeweilige Gesamtmodell

Tabelle 27 zeigt den unstandardisierten Regressionskoeffizienten (B), die Steigung, den standardisierten Regressionskoeffizienten (beta) sowie R, R^2 und das adjustierte R^2 nach der Berücksichtigung aller unabhängigen Variablen. Zum Ende des letzten Schritts, nach Aufnahme aller Prädiktoren in die Regressionsgleichung betrug die multiple Korrelation R=.50 ($F(3,68)=7.53$, $p<.001$). Im ersten Schritt wurde die Variable „Allgemeine Intelligenz“ aufgenommen. Sie wies die höchste bivariate Korrelation auf. Zum Ende des ersten Schritts betrug R=.39 (F_{change} $(1,70)=12,36$, $p<.001$). Im zweiten Schritt wurde die Skala „Variablenrelationen“ des Wissenstests für die „Schneiderwerkstatt“ als Prädiktor einbezogen, dadurch stieg das R von .39 auf .46 an (F_{change} $(2,69)=5,25$, $p<.05$). Dies bedeutet eine zusätzliche Varianzaufklärung von fast 6%. Nach dem dritten Schritt, nachdem auch der Indikator für die Steuerungsleistung bei der „Schneiderwerkstatt“ berücksichtigt worden war, betrug R=.50 (F_{change} $(3,68)=3,54$, $p<.10$). Die durch die Aufnahme des dritten Prädiktors (hier: Steuerungsleistung) erzielte Steigerung der Vorhersagegenauigkeit war lediglich auf dem 10% Niveau statistisch bedeutsam.

Zur Prüfung der Stichprobenunabhängigkeit der multiplen Korrelation wurde eine Kreuzvalidierung durchgeführt, indem die an der Gesamtgruppe ermittelte Gleichung auf die Daten von zwei Zufallsteilstichproben angewandt wurde (geschichtete Zufallssplitung unter Berücksichtigung des Alters und der Dreiereinteilung des unterschiedlichen Berufserfolgspotential der Teilnehmer, siehe oben, Abschnitt 12.3). Die an der Gesamtgruppe gewonnenen Regressionsgewichte waren zur Vorhersage in den beiden Zufallsteilstichproben annähernd gleich gut geeignet (siehe Tabelle 27).

Das systemspezifische Wissen lieferte zusätzlich zur Allgemeinen Intelligenz einen substantiellen Beitrag zur Prädiktion des Kriteriums. Die nach Berücksichtigung der Allgemeinen Intelligenz und des Wissens verbleibende systematische Varianz der Problemlöseleistung konnte nur noch marginal zur Steigerung der Vorhersagegenauigkeit beitragen. Die Reihenfolge der Berücksichtigung der Prädiktoren in der hierarchischen Regression erfolgte aufgrund der in Abschnitt 9.1 dargelegten theoretischen Überlegungen zur Konstruktvalidität, die in den in Kapitel 15 berichteten Analysen eine empirische Bestätigung gefunden hatten (siehe Tabelle 18). Demzufolge kann die systematische Varianz der bei der Bearbeitung von Problemlöseszenarien erzielten Steuerungsleistungen im wesentlichen auf Intelligenz und Wissen zurückgeführt werden. Da die mit Problemlöseszenarien geleisteten Messungen nichts indizieren, was nicht bereits durch andere Messungen indiziert wird, wurde die Etablierung eines neuen Konstrukts „Problemlösen" aus Gründen der bei der Theoriebildung zu beachtenden Sparsamkeit („parsimony") zurückgewiesen. Vor diesem Hintergrund wurden der Intelligenz und dem Wissen in der hierarchischen Regressionsanalyse Priorität eingeräumt. Rein statistisch erklärten die Steuerungsleistung und das Wissen teilweise die gleiche Varianz des Kriteriums. Dreht man die Schrittfolge zwei und drei der dargestellten Regressionsanalyse entgegen den theoretischen Annahmen um und führt als zweiten Block die Steuerungsleistung ein, so führt das im zweiten Schritt berücksichtigte Problemlösegütemaß zu einer signifikanten Steigerung des R^2 Wertes, während der im dritten Block durch die Berücksichtigung des Wissens erzielte Zuwachs der Vorhersagekraft an der Signifikanzgrenze scheitert.

Die relative Bedeutung der einzelnen Prädiktoren soll im nächsten Abschnitt mit Hilfe einer Kommunalitätenanalyse (Kerlinger und Pedhazur, 1973) untersucht werden, die den Vorteil einer von theoretischen Setzungen unabhängigen Betrachtung der überlappenden Varianzanteile bietet. Zuvor sei aber noch eine einschränkende Anmerkung zur Bedeutung von multiplen Regressionskoeffizienten für die diagnostische Praxis ergänzt. Speziell im eignungsdiagnostischen Bereich scheint eine gegenläufige Beziehung zu existieren zwischen der Häufigkeit, mit der multiple Regressionsanalysen berechnet werden und der Seltenheit, mit der die Ergebnisse von multiplen Regressionsanalysen in praktische Entscheidungsregeln umgesetzt werden. Dies liegt zum einen daran, daß die multiple Korrelation im besonderen Maße von stichprobenbedingten Zufälligkeiten beeinflußt wird, womit der Festschreibung von Gewichtungsregeln Grenzen gesetzt sind. Zum anderen spielt in der Praxis aber auch nicht selten der Aufwandsgesichtspunkt eine (zu große) Rolle. Eine komplizierte Gewichtungsprozedur findet aus diesem Grunde - insbesondere wenn die Testauswertung ohne EDV-Unterstützung stattfindet - häufig keine Akzeptanz. Schließlich ist auch die Interpretation eines im Sinne der Ergebnisse einer multiplen

Regressionsanalyse gewichteten Aggregatwertes problematisch, da der Bedeutungsgehalt der einzelnen Komponenten sich verändert, wenn - wie in der multiplen Regression - die Varianz der übrigen Variablen auspartialisiert wird. (Siehe hierzu z.B. die im Rahmen einer kritischen Diskussion der Interpretation von Strukturgleichungsmodellen gegebenen Ausführungen zur Bedeutung von Regressionskoeffizienten bei Holling (1993).) Sofern die Ergebnisse der multiplen Regressionsanalyse aber nicht in Entscheidungsregeln umgesetzt werden, sind diese Daten auf einem Abstraktionsniveau angesiedelt, welches für die Praxis von bestenfalls sekundärer Bedeutung ist. Die Analyse gibt dann vor allem Auskunft über die maximal mögliche Vorhersagbarkeit des Kriteriums durch die Prädiktoren, nicht aber über die Treffsicherheit des tatsächlich angewandten diagnostischen Verfahrens. Daraus erfolgt zum einen, daß die Praxis sich verstärkt um gewichtete Prognosen bemühen sollte, zum anderen aber auch, daß die Beschreibung der Validität sich auf die in der Praxis tatsächlich verwandten Maße beziehen sollte. Der einfache Korrelationskoeffizient liefert häufig konkretere Informationen über die Vorhersagegüte der in der Praxis eingesetzten Instrumente. Alternativ zur Variablengewichtung im oben dargestellten Ansatz der Regressionsanalyse wurden zusätzlich auch einfache Aggregate aus jeweils zwei der drei Prädiktoren Intelligenz (Skala „Allgemeine Intelligenz") Wissen (Skala „Variablenrelationen") und Problemlösen (Kapitalendwert der „Schneiderwerkstatt") gebildet. Aggregiert wurden jeweils z-transformierte Leistungen. Die durch die z-Transformation erzielte Gleichgewichtung läßt sich in der Praxis leicht nachstellen, wenn dort zur Aggregation die ohnehin häufig verwendeten standardisierten Leistungskennwerte herangezogen werden. Die einfache Korrelation zwischen diesen Maßen und dem Vorgesetztenurteil über die im Berufsalltag gezeigten Leistungen/Fähigkeiten aus dem Bereich der Intelligenz und des Problemlösens betrug $r=.47$, $r=.46$ und $r=.42$ für die Aggregate „Intelligenz und Problemlösen", „Intelligenz und Wissen" sowie „Wissen und Problemlösen". Mit dem einfachen Aggregat aller drei Prädiktoren konnte eine - auf dem 1% Niveau statistisch bedeutsame - Vorhersage in Höhe von $r=.49$ erzielt werden.

17.3.3 Kommunalitätenanalyse

Bei der Kommunalitätenanalyse (Kerlinger und Pedhazur, 1973) werden die relativen Vorhersageanteile der einzelnen Prädiktoren unabhängig von einem theoretischen Modell der Variablenbeziehungen bestimmt. Die bei der Vorhersage insgesamt aufgeklärte Varianz läßt sich mit der Kommunalitätenanalyse separieren in *spezifische* Varianz, die nur durch jeweils einen der drei Prädiktoren „Intelligenz", „Wissen" und „Problemlösen" vorhergesagt werden konnte sowie in konfundierte

Varianz, die durch zwei oder drei der Prädiktoren erklärt wurde. In Tabelle 28 sind die entsprechenden Varianzanteile wiedergegeben. (Für Anwendungsbeispiele zur Kommunalitätenanalyse siehe z.B. Schrader & Helmke 1990, Marjoribanks, 1994.)

Tab. 28: Kommunalitätenanalyse: Relative Bedeutung von Intelligenz, Wissen und Problemlösen bei der Prognose des Vorgesetztenurteils über die im Beruf gezeigten Fähigkeiten und Leistungen aus den Bereichen „Intelligenz" und „Problemlösen"

Spezif. Varianz:	Intelligenz: 7,32%		Wissen: 2,95%	Pl.:3,91 %
Konfundierte Varianz:	I+Pl=2,49%	I+W=2,02%	W+Pl=3,07%	
	Intelligenz (I), Wissen (W) u. Problemlösen (Pl): 3,17 %			
Erklärte Gesamtvarianz:	24,93%			

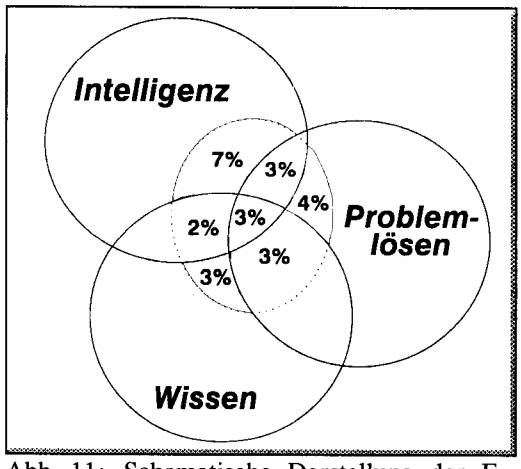


Abb. 11: Schematische Darstellung der Ergebnisse der Kommunalitätenanalyse zur relativen Bedeutung der Prädiktoren. (Die Größe der Varianzanteile wird nur durch die gerundeten Prozentangaben, nicht durch die Größenverhältnisse repräsentiert.)

Abbildung 11 stellt eine schematische Darstellung der Ergebnisse der Kommunalitätenanalyse dar. Die Darstellung veranschaulicht lediglich das Prinzip der spezifischen und konfundierten Varianzen; es wurde kein Wert auf eine graphische Repräsentation der Größenverhältnisse der Varianzanteile gelegt. Der innere kleine Kreis veranschaulicht den Anteil von 24,93% der Kriteriumsvarianz, die durch die drei Prädiktoren vorhergesagt werden konnte (siehe Tabelle 27). Relativ gesehen stellt die Intelligenz mit dem größten Anteil an spezifischer Varianz einen unverzichtbaren Prädiktor dar. Der Anteil der spezifischen Varianz der übrigen beiden Prädiktoren ist hingegen jeweils nur halb so groß. Die übrigen Prädiktoren können daher lediglich als eine sinnvolle Ergänzung der durch die Intelligenz geleisteten Prädiktion angesehen werden. Dabei reicht *ein* weiterer Prädiktor - wie die Regressionsanalyse im vorherigen Abschnitt gezeigt hat - zur signifikanten Steigerung der Vorhersageleistung aus. Die insgesamt aufgeklärte Kriteriumsvarianz konnte zu 84% durch die Kombination von Intelligenz und Wissensleistungen vorhergesagt werden. Es bleibt zu vermuten, daß der Anteil der

konfundierten Varianz noch größer wäre, falls auf Seiten der Intelligenz eine Binnendifferenzierung in die einzelnen Intelligenzkomponenten und auf Seiten des Wissens eine Berücksichtigung anderer Wissensbereiche - wie z.B. strategisches Handlungswissen und/oder quantitatives Sachwissen - möglich gewesen wäre.

17.3.4 Strukturgleichungsmodell zur Vorhersage der Problemlöseleistung und des Berufserfolgs durch Intelligenz und Wissen

Die theoretischen Annahmen der vorliegenden Arbeit wurden im letzten Abschnitt der Datenanalyse in einem Strukturgleichungsmodell spezifiziert. Dabei wurden die postulierten Beziehungen zwischen den Variablen Intelligenz, Wissen, Problemlösen und Berufsleistung untersucht. Als Annahme wurde formuliert, daß Intelligenz eine Voraussetzung zum Erwerb und zur Anwendung von Wissen ist. Zusätzlich wurde die Problemlöseleistung durch Intelligenz und Wissen erklärt. Schließlich wurde davon ausgegangen, daß die Berücksichtigung von Intelligenz und Wissen die berufliche Leistung hinreichend erklärt und die Annahme einer eigenständigen Vorhersagekraft des Problemlösens nicht notwendig ist.

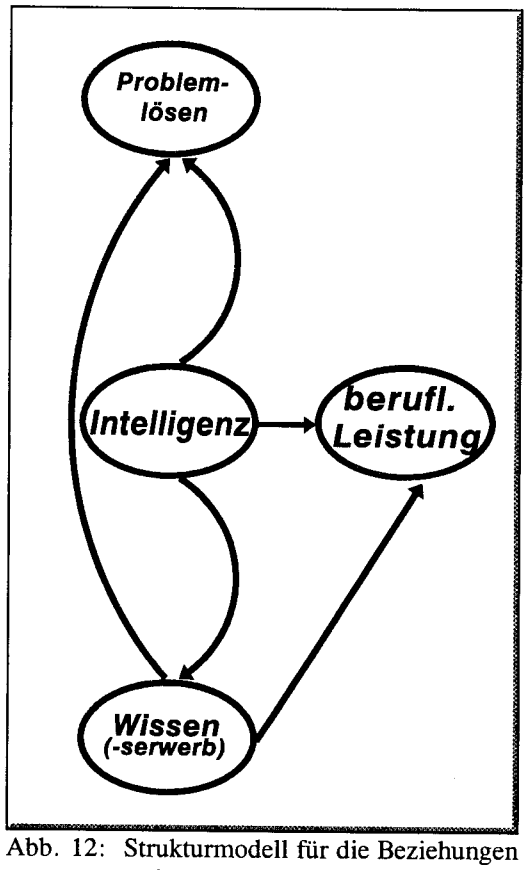


Abb. 12: Strukturmodell für die Beziehungen zwischen Intelligenz, Wissen, Problemlösen und beruflicher Leistung

Das in Abbildung 12 dargestellte theoretische Strukturmodell umfaßt die genannten Beziehungen der Konstrukte. Die Methode der Analyse eines linearen Strukturgleichungsmodells wurde eingesetzt, um über die in den bisher berichteten Analysen betrachtete Koexistenz der Variablen hinaus empirische Hinweise auf die theoretisch postulierte Sukzession der Variablen zu erhalten. (Zu Koexistenz- und Sukzessionsgesetzen siehe z.B. Groeben und West-

meyer, 1975). Das formulierte „Kausalmodell" nimmt den Begriff der „Kausalität" mit Hodapp (1984) oder Schneider (1985) lediglich im Sinne der hypothesenkennzeichnenden Vorhersage-Asymmetrie der Variablenbeziehungen in Anspruch.

Die theoretischen Variablen wurden durch beobachtete Variablen spezifiziert. Während das Konstrukt „Intelligenz" nur durch einen Indikator, nämlich durch die Skala „Allgemeine Intelligenz" definiert wurde, standen für die übrigen Variablen jeweils zwei empirische Repräsentationen zur Verfügung: Die Problemlösefähigkeit wurde durch die Steuerungsleistung in den beiden Szenarien definiert. Dabei wurde für „DISKo" der intern valideste Indikator, das auf die ersten acht Takte beschränkte neue Problemlösegütemaß ausgewählt. Für die „Schneiderwerkstatt" wurde ebenfalls das neue Problemlösegütemaß herangezogen, da dieser Indikator eine größere Nähe zu dem „DISKo"-Indikator aufwies als der Kapitalendwert (siehe oben, Tabelle 17), und somit die Kombination dieser beiden Variablen für die Definition eines gemeinsamen Konstrukts geeignet ist. Zur Spezifikation von „Wissen" wurden die beiden Wissensskalen aus dem systemspezifischen Wissenstest zur „Schneiderwerkstatt" herangezogen. Die Variable „berufliche Leistung" wurde durch das Urteil der Vorgesetzten über die im Berufsalltag gezeigten (1.) intellektuellen und (2.) problemlösenden Fähigkeiten/Leistungen definiert. Dabei wurde a priori festgesetzt, daß diese beiden Indikatoren das gleiche Gewicht erhalten.

Das lineare Strukturgleichungsmodell wurde mit Hilfe des Programms LISREL (PC-Version 8.12a, Jöreskog und Sörbom, 1993) berechnet. Das Modell zeigte eine gute Übereinstimmung mit den Daten (siehe Abbildung 13). Aufgrund der kleinen Stichprobe und aufgrund des Umstandes, daß als Ausgangsbasis der Berechnungen die Korrelationsmatrix (nicht die Kovarianzmatrix) diente, wurde als Schätzmethode die Methode der ungewichteten kleinsten Abweichungsquadrate benutzt. Der „Goodness-of-fit"-Index (GOF) betrug .99 bzw .97 für den „Adjusted Goodness of Fit" (AGOF). Insgesamt bestätigen die geschätzten Koeffizienten die theoretischen Annahmen über die Bedeutung von Intelligenz und Wissen für das Problemlösen sowie die Annahmen über die Bedeutung von Intelligenz, Wissen und Problemlösen als Prädiktoren beruflicher Leistung.

Der im Vergleich zur Intelligenz nominell nahezu gleich hohe standardisierte Pfadkoeffizient für den Zusammenhang zwischen Wissen und beruflicher Leistung (0.36 gegenüber 0.35) ist vor dem Hintergrund der indirekten Effekte zu interpretieren. Zerlegt man die aufgeklärte Varianz in die Einzelkomponenten und betrachtet den indirekten, d.h. durch Intelligenz vermittelten, Effekt des Wissens auf die Berufsleistung, so betrug dieser standardisierte Effekt 0.12. Hinsichtlich der standardisierten totalen Effekte ging der nominell stärkste Einfluß von der Intelligenzvariablen aus (0.47 gegenüber 0.36 für Wissen). Die Bedeutung der Intelligenz zeigte sich sowohl durch die direkten Pfade zum Kriterium und zum Problemlösen

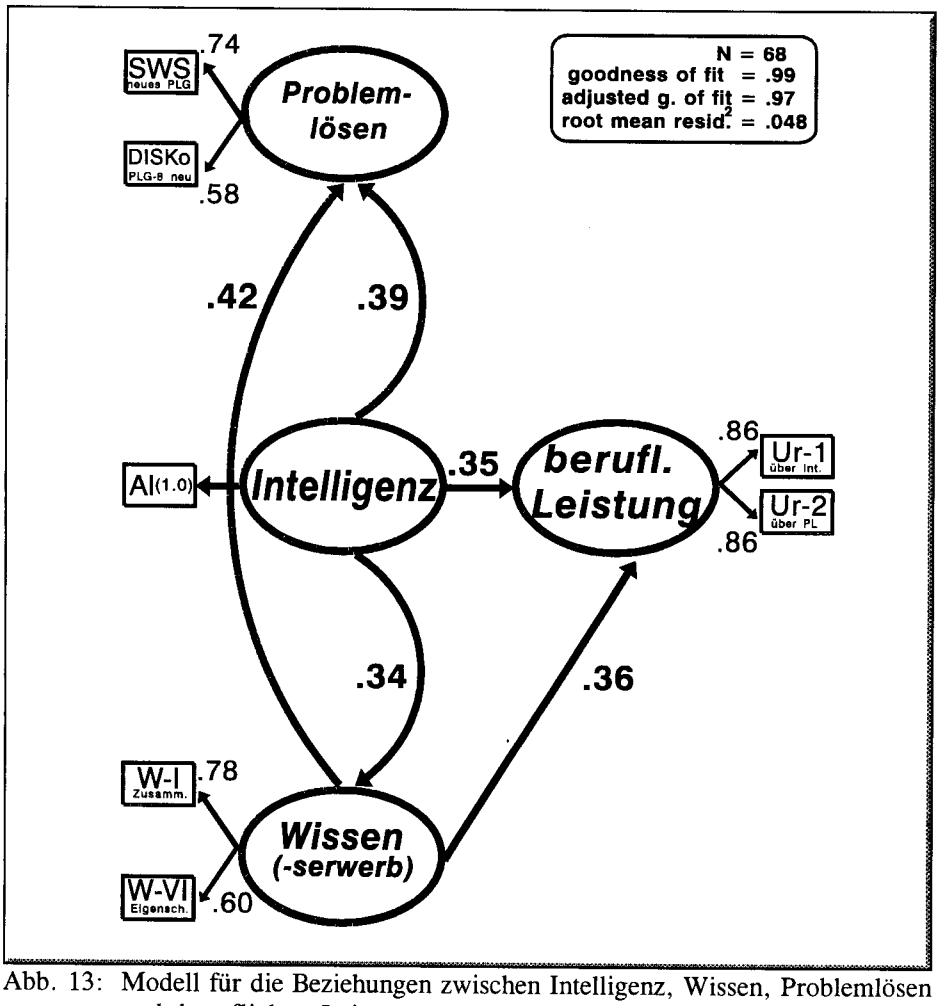


Abb. 13: Modell für die Beziehungen zwischen Intelligenz, Wissen, Problemlösen und beruflicher Leistung; verzeichnet sind die vollstandardisierten LISREL-Schätzer der Einflußfaktoren

als auch durch die zusätzlichen indirekten Pfade über das Wissen zum Problemlösen und zur beruflichen Leistung.Die Daten widersprechen weder der Annahme, daß Intelligenz eine Voraussetzung zum Erwerb und zur Anwendung von Wissen ist noch der Annahme, daß Intelligenz und Wissen gemeinsam das problemlösende Handeln sowohl im Problemlöseszenario als auch im Berufsalltag bestimmen. Allerdings bedarf das Modell einer Replikation an einem weiteren Datensatz mit einer größeren Stichprobe. Für ein Modell mit einem zusätzlichen direktem Pfad zwischen der Problemlösefähigkeit und der beruflichen Leistung ergab sich kein deutlich besserer Fit (GOF ebenfalls .99, und AGOF=.98).

17.4 Zusammenfassung und Diskussion

Sowohl die in der Literatur zur Kennzeichnung der Problemlösefähigkeiten und der Problemlöseleistungen benutzten Beschreibungen als auch die Beschreibungen von Fähigkeiten und Leistungen aus dem Kontext des Konstrukts „Intelligenz" sind nach Einschätzung der befragten Vorgesetzten für die erfolgreiche Bewährung im beruflichen Alltag eines Polizisten des gehobenen Dienstes relevant. Die im Berufsalltag gezeigten Fähigkeiten und Leistungen aus beiden Bereichen konnten mit den Prädiktoren „Intelligenz", „Wissen" und „Problemlösen" vorhergesagt werden. Bei dem Szenario „DISKo" erwies sich allerdings nicht die standardmäßig über den Gesamtkapitalendwert bestimmte Steuerungsleistung, sondern nur ein modifiziertes Problemlösegütemaß als prädiktiv valide. Tendenziell stand die Kriteriumsvalidität der Intelligenztests und der computergestützten Problemlöseszenarien in Abhängigkeit von dem jeweils vorhergesagtem Kriterium (Urteil über intelligenzbezogene oder über problemlösebezogene berufliche Fähigkeiten und Leistungen). Dabei konnte jeder Prädiktor den theoretisch ähnlichsten Kriterienbereich nominell am besten vorhersagen. Allerdings waren sowohl die Prädiktoren untereinander als vor allem auch die beiden Kriterien untereinander korreliert. Für die weiteren Analysen wurden die Vorgesetztenurteile über die im Beruf gezeigten intellektuellen Fähigkeiten und Leistungen mit den Urteilen über die im Berufsalltag gezeigten problemlösenden Fähigkeiten und Leistungen aggregiert. Diese Aggregation erfolgte, da beiden Leistungsbereichen von den Vorgesetzten eine hohe berufliche Relevanz zugesprochen wurde und die isolierte Betrachtung nur eines Kriteriums den Anforderungen des Berufsalltags somit nicht gerecht geworden wäre. Als bedeutsamster und unverzichtbarer Prädiktor dieses Kriterienaggregats erwies sich die Intelligenz mit einer singulären prädiktiven Kriteriumsvalidität von $r = .39$. Die Vorhersagekraft des Problemlösens ließ sich im stärkeren Ausmaß auf deren Intelligenz- und Wissensanteil zurückführen als umgekehrt die Vorhersagekraft der Intelligenz auf deren Problemlöse- und Wissensanteil (siehe hierzu Abschnitt 17.3.3). Die mit den Szenarien reliabel erfaßten kriteriumsrelevanten Informationen wurden größtenteils bereits durch die Variablen Intelligenz und Wissen weitgehend repräsentiert.

Die Prognose beruflicher Leistungen konnte inkrementell gesteigert werden, indem zusätzlich zu dem besten Einzelprädiktor, dem Intelligenztest, ein weiterer Prädiktor Berücksichtigung fand. Der Befund einer Steigerung der Prognosekraft durch die Berücksichtigung eines weiteren Prädiktors ist überaus praxisrelevant. Hossiep (1997) führt ein Berechnungsbeispiel für das gehobene mittlere Management (jährlicher Personalkostenaufwand DM 350.000) an. Bei realistischen Annahmen zu den Rahmenbedingungen (Anzahl der untersuchten Bewerber, voraussichtliche Betriebszugehörigkeit in Jahren, Verfahrenskosten usw.) schlägt sich eine Validitätssteige-

rung um 0,05 bereits in einem Nutzenzuwachs von ca. DM 140.000 nieder. Für die Steigerung der Vorhersageleistung über den durch Intelligenz abgedeckten Varianzbereich hinaus erwiesen sich die Variablen „Wissen" und „Problemlösen" statistisch als funktional weitgehend äquivalent. Diese funktionale Austauschbarkeit wurde darauf zurückgeführt, daß die systematische Varianz der Steuerungsleistung sich im wesentlichen auf Intelligenz- und Wissensleistungen zurückführen ließ (siehe auch Abschnitt 15.4). Eine einfache Aggregation der Intelligenz- und Wissensleistungen steigerte den Zusammenhang zum Kriteriumsverhalten auf r=.46. Die insgesamt aufgeklärte Kriteriumsvarianz konnte zu 84% durch die Kombination von Intelligenz und Wissensleistungen vorhergesagt werden. Es bleibt zu vermuten, daß der verbleibende Anteil systematischer Kriteriumsvarianz bei einer differenzierteren Intelligenz- und Wissensmessung ebenfalls durch Indikatoren dieser Konstrukte aufgeklärt werden könnte. Die theoretischen Annahmen zum Zusammenhang von Intelligenz, Wissen und Problemlösen im Kontext der Prädiktion beruflicher Leistungen wurden schließlich in einem Strukturmodell spezifiziert. Die theoretisch formulierten Zusammenhänge stimmten mit den beobachteten Zusammenhängen hinreichend überein. Demzufolge bildet Intelligenz eine Voraussetzung zum Erwerb und zur Anwendung von Wissen. Intelligenz und Wissen sind sowohl eine Voraussetzung des Problemlösens als auch der beruflichen Leistungen.

18. Abschließende Diskussion / Ausblick

Ziel der vorliegenden Arbeit war es, die Möglichkeiten einer professionellen Fähigkeitsdiagnostik mit Hilfe computergestützter Problemlöseszenarien auszuloten und den theoretischen Anspruch kritisch zu würdigen sowie die empirische Datenlage zu sichten und teilweise zu ergänzen. Die drei Hauptargumente für den diagnostischen Einsatz von computergestützten Problemlöseszenarien konnten einer kritischen Prüfung überwiegend nicht standhalten.

(1) Die Behauptung, (1a) zwischen bestimmten beruflichen Anforderungen (z.B. Managementanforderungen) einerseits und den Anforderungen bei der Bearbeitung computergestützter Problemlöseszenarien andererseits bestünde eine besonders hohe Korrespondenz (ökologische Validität) entbehrt ebenso der Grundlage wie die Behauptung (1b), die Szenarien würden bestimmte Realitätsbereiche *simulieren*. Auf theoretischer Ebene ist anzumerken, daß die für eine sinnvolle Benutzung des Simulationsbegriffs notwendigen Voraussetzungen, z.B. Modellanalysen, Spezifikationen von Abbildungsvorschriften und Modellvalidierungen, für die meisten Szenarien nicht einmal ansatzweise erfüllt sind. Ähnliches gilt hinsichtlich des Anspruchs auf ökologische Validität. Der Nachweis ökologischer Validität erfordert u.a., daß die bei der Steuerung eines spezifischen Szenarios gestellten Anforderungen präzise definiert werden. Diese Voraussetzung erfüllen die meisten Szenarien nicht. Dieser Mangel geht häufig mit der fehlenden Beschreibung der formalen Aufgabenmerkmale von Szenarien einher. In der vorliegenden Studie ergab sich, daß Fähigkeiten und Leistungen, die im Kontext der Problemlöseforschung beschrieben werden, zwar von berufserfahrenen Personen als berufserfolgsrelevant erachtet wurden, daß dies aber auch für Fähigkeiten und Leistungen gilt, die im Kontext der Intelligenzforschung postuliert werden. Daß den Fähigkeitsbeschreibungen eine Relevanz für den Berufserfolg zugesprochen wird, bedeutet nicht, daß auch den zugeordneten Messungen eine entsprechende Relevanz zukommt.

(2) Auch das zweite Argument für den diagnostischen Einsatz von Problemlöseszenarien, die vermeintlich hohe Akzeptanz von Problemlöseszenarien, muß aus theoretischen Gründen und aufgrund empirischer Daten deutlich in Frage gestellt werden. In der hier vorgestellten Untersuchung wurden auch Daten zur vergleichenden Bewertung der Akzeptanz von Problemlöseszenarien und Intelligenztests erhoben. Anhand einer an anderer Stelle (Kersting, 1998) dokumentierten Auswertung dieser Daten ließ sich zeigen, daß die Teilnehmer die Akzeptanz dieser Aufgaben

als Instrumente der Personalauswahl differenziert beurteilten. Die Problemlöseszenarien fanden vor allem unter dem Gesichtspunkt *„positives Erleben/Spaß"* eine hohe Akzeptanz. Intelligenztests übertrafen die Problemlöseszenarien hingegen bezüglich der *„Kontrollierbarkeit"* im Sinne einer höheren wahrgenommenen Qualität der Messung. Die Akzeptanzurteile variierten in Abhängigkeit von Person- und Situationsmerkmalen. Bei einer Meßwiederholung im Anschluß an eine Rückmeldung über das Verfahrensergebnis ergaben sich hinsichtlich der dritten geprüften Akzeptanzdimension, der *„Anforderungsnähe"/„face validity"* veränderte Akzeptanzeinstufungen. Insgesamt konnten keine stichhaltigen Anhaltspunkte dafür gefunden werden, daß bei der Personalauswahl Problemlöseszenarien eine grundsätzlich höhere Akzeptanz finden als Intelligenztests. Aus differential-psychologischer Perspektive ist es - insbesondere für den Bereich der Eignungsdiagnostik - anzuzweifeln, daß einem Verfahren ein einziger, person- und situationsunabhängiger Akzeptanzwert zugeordnet werden kann. Pauschale Behauptungen über die vermeintlich hohe Akzeptanz von Problemlöseszenarien als Instrumente der Personalauswahl können daher als Argument für deren diagnostischen Einsatz nicht überzeugen.

(3) Das dritte Argument für den diagnostischen Einsatz von Problemlöseszenarien, demzufolge computergestützte Problemlöseszenarien eine Erweiterung der auf Intelligenztests basierenden Fähigkeitsdiagnostik leisten, ist differenziert zu bewerten. Eine Erweiterung in dem Sinne, daß ein neues Konstrukt, eine neue Fähigkeit (z.B. „Problemlösen", „operative Intelligenz") diagnostiziert werden könnte oder sollte, ist empirisch nicht zu legitimieren. Diese Aussage konnte in der vorliegenden Studie in zwei Analysen bestätigt werden. Sowohl der systematische Anteil gemeinsamer Varianz der Steuerungsleistung in beiden eingesetzten Szenarien als auch - in etwas eingeschränkter Form - die kriteriumsrelevante Varianz der Steuerungsleistung ließ sich mit den Intelligenz- und Wissensindikatoren hinreichend empirisch abbilden. Mit diesem Ergebnis erfährt der entsprechende Befund von Süß (1996) eine konzeptionelle Replikation. Auch die Hoffnung, die Intelligenzdiagnostik durch Problemlöseszenarien in Richtung einer „Prozeßdiagnostik" zu erweitern, erscheint empirisch unberechtigt. In der vorliegenden Studie wurde ein solches Maß der Prozeßdiagnostik berücksichtigt, nämlich der bei dem Szenario „DISKo" automatisch gebildete Indikator zur Beurteilung der Verhaltensweisen und Strategien. Die kriteriumsrelevante Varianz dieser Variable wurde durch die herkömmliche Intelligenzmessung hinreichend repräsentiert.

Computergestützte Problemlöseszenarien bieten also keinen Zugang zu einem neuen Konstrukt oder zu einer validen Prozeßdiagnostik. Eine andere Bewertung muß das Erweiterungsargument allerdings erfahren, wenn man bei der Erweiterung der mit Intelligenztests geleisteten Fähigkeitsdiagnostik durch computergestützte Problemlöseszenarien an die Wissensdiagnostik denkt. Auch in der vorliegenden

Studie zeigte sich, daß die Steuerungsleistung insofern über die Intelligenz hinausgeht als bei der Systembearbeitung nicht nur Intelligenz, sondern zusätzlich auch Wissen erforderlich ist. Durch die Berücksichtigung eines kontentvaliden systemspezifischen Wissenstests konnte die prädiktive Validität der Intelligenztests inkrementell gesteigert werden. Statistisch waren die Variablen Wissen und Steuerungsleistung hinsichtlich ihres Beitrags zur Steigerung der Vorhersageleistung der Intelligenztests austauschbar. Auch wenn zur Zeit nichts für die Annahme eines neuen Konstrukts „Problemlösen" spricht, so kann die Auseinandersetzung mit der Steuerungsleistung bei Problemlöseszenarien sich doch als fruchtbar für mögliche Fortschritte auf dem Gebiet der Fähigkeitsdiagnostik erweisen. Als ein Resultat dieser Auseinandersetzung steht die Erinnerung an das etablierte Konstrukt „Wissen" und die Überlegung, die Intelligenzdiagnostik durch eine Wissensdiagnostik zu ergänzen. Problemlöseszenarien könnten bei der Wissensdiagnostik hilfreich sein, falls bei der Bearbeitung der Szenarien - begünstigt durch das szenarienkennzeichnende Merkmal des Feedbacks - meßbare Wissenserwerbsprozesse stattfinden und falls es gelingt, diesen Wissenserwerb zuverlässig zu messen.

Gerade im Kontext der beruflicher Leistungen und ihrer Vorhersage kommt dem „Wissen" eine große Bedeutung bei. Schmidt, Hunter und Outerbridge (1986, zitiert nach Schmidt, 1992, S. 1178 f.) berechneten auf der Basis der metaanalytisch gewonnenen korrigierten Korrelationen zwischen „Intelligenz", „berufsspezifischem Wissen" (operationalisiert über kontentvalide Messungen), „berufsspezifischem Leistungspotential" (erhoben über Arbeitsproben), „Berufserfahrung" und „berufliche Leistungen" (Vorgesetztenbefragung) ein Pfadmodell der Berufsleistung. Diesem Modell zufolge wirkt Intelligenz vor allem indirekt, nämlich über das berufsspezifische Wissen, auf die berufliche Leistung. Diese Auffassung ist mit der von Süß et al. (1993a, S. 192f.) im Kontext der Problemlöseforschung vorgebrachten und in der vorliegenden Arbeit zugrundegelegten These vereinbar, daß intellektuelle Fähigkeiten eine notwendige Voraussetzung für den Erwerb, die Anwendung und die Modifikation von Wissen sind. Die Überlegungen und die Daten sprechen dafür, Wissen nicht nur indirekt (durch die Messung der Intelligenz), sondern verstärkt auch direkt bei der Prognose beruflicher Leistungen zu berücksichtigen. Bislang wird Wissen lediglich in Form der Überprüfung spezifischer Fachkenntnisse bei der Eignungsdiagnose berücksichtigt; auch dies ist nach Schuler (1996, S. 139) zur Zeit aber noch eher unüblich. Diese Praxis steht im Gegensatz zu den empirischen Befunden, die dem Wissen eine hohe Kriteriumsvalidität beimessen. Dye, Reck und McDaniel (1993, zitiert nach Schuler, ebd.) berichten für Fachkenntnisse eine durchschnittliche Validität von $r = .45$. Eine Ursache für die eignungsdiagnostische Vernachlässigung von Wissen im allgemeinen und Fachkenntnissen im besonderen dürfte darin liegen, daß Wissen nicht den Grad der Voraussetzungsfreiheit erfüllt,

den man in der Eignungsdiagnostik erwartet. Das weiter oben angeprochene Pfadmodell der Berufsleistung hatte gezeigt, daß neben der Intelligenz auch die Berufserfahrung das berufsspezifische Wissen bestimmt. Wissen, welches erst im Beruf erworben wird, läßt sich diagnostisch nicht für die Auswahl von Berufs*einsteigern* nützen. Diagnostisch interessant ist eigentlich nicht das Wissen selbst, welches rasch veraltet, sondern die Fähigkeit zum Wissenserwerb und zur Wissensnutzung. Diese Fähigkeit ist auf begrifflicher Ebene im Intelligenzkonstrukt beheimatet, findet auf der Ebene der Operationalisierungen in herkömmlichen Intelligenzaufgaben aber kaum Berücksichtigung. Die Entwicklung von Lerntests (für einen Überblick siehe Guthke und Wiedl, 1996) dürfte in diesen Defiziten der Intelligenztestaufgaben ihr fundamentum in re haben. Die Verwandtschaft zwischen Lerntests und Problemlöseszenarien wurde bereits theoretisch und empirisch ausgelotet (Beckmann, 1994; Beckmann und Guthke, 1995). Auf der Meßebene könnten die Problemlöseszenarien eventuell tatsächlich eine sinnvolle Erweiterung der herkömmlichen Intelligenzdiagnostik eröffnen - allerdings nicht als Tests, sondern als Ausgangsbasis für eine relativ voraussetzungsfreie Wissensdiagnostik. Die Bearbeitung eines sorgfältig konstruierten Szenarios könnte als Anwendungsfeld intellektueller Fähigkeiten über Prozesse der systematischen Hypothesentestung Gelegenheit zur Aneignung und zum Ausbau von Wissen bieten. Das Problemlöseszenario ermöglicht es den Diagnostikanden, Informationen auswählen, zu generieren, zu verknüpfen und zu strukturieren. Da alle Probanden Gelegenheit zur Systemsteuerung und somit Zugang zur Erfahrungsgrundlage erhalten, ist die anschließende Diagnose des systemspezifischen Wissens vergleichsweise voraussetzungsfreier als der Einsatz herkömmlicher Kenntnistests. Für die hier skizzierte diagnostische Perspektive wäre allerdings ein psychometrisch befriedigender, erfahrungssensitiver und valider systemspezifischer Wissenstest notwendig. Während ein solcher Test für die „Schneiderwerkstatt" existiert (Kersting, 1991; Kersting und Süß, 1995), stehen vergleichbare Konstruktionsbemühungen und vor allem Dokumentationen sowie Bewährungsnachweise für Wissenstests zu anderen Szenarien überwiegend noch aus.

Diese zuletzt geschilderte Perspektive der Nutzung von Problemlöseszenarien zur Provokation von testbaren Wissenserwerbsprozessen geht als Ausblick bewußt über den empirischen Beitrag der vorliegenden Arbeit hinaus. In der vorliegenden Studie wurde der systemspezifische Wissenstest zur „Schneiderwerkstatt" nur einmal - nach der Steuerung - vorgegeben, die Wissensdiagnose bei „DISKo" erfolgte ebenfalls nicht unabhängig von der Systemsteuerung. Es ist im nachhinein nicht möglich zu entscheiden, ob es sich bei dem prognostisch validen Wissensanteil um systemspezifisches Vorwissen oder - eignungsdiagnostisch reizvoller - um erworbenes Wissen handelte. Die Klärung dieser Frage bleibt weiteren Untersuchungen vorbehalten. Für entsprechend orientierte Studien bieten sich insbesondere die im Kontext

der Problemlöseforschung weniger beachteten abstrakt eingekleideten Szenarien an.

Festgehalten werden kann, daß die Berücksichtigung der Variable „Wissen" in der vorliegenden Studie mit einer inkrementellen Steigerung der intelligenztest-basierten Vorhersageleistung verbunden war und daß eine darüber hinausgehende zusätzliche Berücksichtigung der Steuerungsleistung die Vorhersage nicht mehr zufallskritisch bedeutsam verbessern konnte. Statistisch hätte allerdings auch die direkte Berücksichtigung der Steuerungsleistung zu einer inkrementellen Steigerung der Vorhersage beruflicher Leistungen durch Intelligenztests geführt. Die verfügbaren empirischen Daten lassen es offen, ob die Steuerungsleistung oder ob die Wissensleistung als Ergänzung zur Intelligenzmessung bei der Vorhersage beruflicher Leistungen Berücksichtigung finden sollte. Aus theoretischer Sicht ist diesbezüglich u.a. zu berücksichtigen, daß die systematische Varianz der Steuerungsleistung sich durch Intelligenz und Wissen aufklären läßt. Die Prädiktorkombination von Intelligenz- und Steuerungsleistungen ist daher ebenso im Sinne der Konstrukte „Intelligenz" und „Wissen" zu interpretieren wie die Prädiktorkombination von Intelligenz- und Wissensleistungen. Die beiden sinnvoll differenzierbaren Konstrukte „Intelligenz" und „Wissen" werden in der Steuerungsleistung aber zu einem nicht mehr differenzierbaren Konglomerat verschmolzen, welches zusätzlich möglicherweise noch durch weitere Situations- und Personmerkmale bestimmt wird (z.B. Computererfahrung). Dies führt dazu, daß bei der diagnostischen Verwendung von Steuerungsleistungen die Beziehung zwischen dem Indikator und dem Indizierten nicht näher bestimmt werden kann und damit das diagnostische Zeichen nicht eindeutig interpretierbar ist. Gegen die unmittelbare diagnostische Nutzung der Steuerungsleistung spricht auch der Umstand, daß dieses Maß vermutlich relativ leicht manipuliert werden kann (siehe Abschnitt 10.2). Hinsichtlich des konkurrenten Kriteriums „Laufbahnstatus" erwies sich in der vorliegenden Arbeit das Wissen, nicht aber die Steuerungsleistung als valide.

Man wird weiterhin wissenschaftlich kontrovers über die für die Fähigkeits- und Berufseignungsdiagnostik geeigneten Ergänzungen von Intelligenztests diskutieren. Hingegen sollte unter Experten Einigkeit bestehen, daß es zum gegenwärtigen Stand der Forschung im Regelfall nicht zu rechtfertigen ist,

(1) von einer durch die Steuerung von computergestützten Problemlöseszenarien diagnostizierbaren „Problemlösefähigkeit" zu sprechen und damit eine Fähigkeit zu meinen, die nicht (besser) durch Intelligenz- und Wissenstests erfaßt werden kann,

(2) im Rahmen der Personalauswahl von Personen ohne vorherige Berufserfahrung bei der Fähigkeitsdiagnostik auf den Einsatz von Intelligenztests zu verzichten.

Intelligenztests erzielen höhere Validitätswerte und sind für differentialdiagnostische Entscheidungen besser geeignet als Problemlöseszenarien. Die Frage, ob Intelligenztests eine treffsichere Vorhersage beruflichen Erfolgs erlauben, hat sich

längst erledigt. Daß sie es tun, ist ein Nebenergebnis einer kaum noch zu überschauenden Anzahl von Studien (siehe z.B. Hunter & Hunter, 1984; Schmidt, 1988; 1992). Lediglich Fragen wie die nach dem Funktionsmechanismus der intelligenztestbasierten Prognosen (z.B. direkt oder über Wissen vermittelt) oder die Frage nach der sinnvollen Ergänzung und Erweiterung der Intelligenztestaufgaben lassen einen zusätzlichen Einsatz von computergestützten Problemlöseszenarien in der Eignungsdiagnostik sinnvoll erscheinen. Im günstigsten Falle stellen computergestützte Problemlöseszenarien *add-on technologies* dar: sie setzen - z.B. als Ausgangsbasis für eine relativ voraussetzungsfreie Wissensdiagnostik - auf das, was prognostisch möglich ist, noch eines drauf; es sind aber nach dem aktuellen Stand der Forschung keinesfalls *substitute technologies*, die obsolete Verfahren durch effektivere ersetzen könnten. Zahlreiche schwerwiegende Bedenken gegen den diagnostischen Einsatz von Problemlöseszenarien bleiben bestehen oder werden durch die vorliegenden Ergebnisse sogar noch genährt. Insbesondere die durch inadäquate Problemlösegütemaße hervorgerufen Schwierigkeiten können nicht hoch genug angesiedelt werden. Der in der Berliner Erstuntersuchung (Süß et al, 1993b) ermittelte Befund, demzufolge das Standardproblemlösegütemaß unter den Umständen einer überforderten Probandengruppe keine interne Validität aufwies und die Konstruktion eines neuen Problemlösegütemaßes erforderte, konnte in der vorliegenden unabhängigen Studie für ein anderes Szenario („DISKo") repliziert werden. Da in Abschnitt 7.2 gezeigt wurde, daß die Überforderung der Steuerer beinahe zum charakterisierenden Merkmal des Einsatzes von Problemlöseszenarien geworden ist, weckt dieser Befund erneut über die aktuellen Daten hinausgehende Bedenken an der internen Validität von Problemlösegütemaßen. Die für jeden Anwendungsfall notwendigen Aufgabenanalysen der Szenarien, wie sie in der Berliner und in der vorliegenden Untersuchung geleistet wurden, düften in der diagnostischen Praxis eher die Ausnahme darstellen. In mittlerweile zwei unabhängigen Studien ergaben sich aufgrund der Überforderung der Problemlöser - und diese Überforderung kann nicht als untypisch gelten - eklatante Probleme mit dem Standardmaß der Beurteilung der Problemlösegüte. Unter diesen Umständen unterscheidet sich die diagnostische Bewertung von Personen aufgrund dieser Standardproblemlösegütemaße nicht vom Kaffeesatzlesen. Jeder Anbieter von diagnostisch eingesetzten Problemlöseszenarien hat den Nachweis zu erbringen, daß die Problemlösegütemaße für die jeweilige Zielgruppe des Verfahrens intern und extern valide sind. Besondere Merkmale der Zielgruppe sind aufzulisten, dies gilt insbesondere für erforderliche Erfahrungsvoraussetzungen (z.B. Vertrautheit mit Computern, mit der Rahmengeschichte des Szenarios usw.). Deutliches Mißtrauen ist auch gegenüber sogenannten Verhaltensmaßen angebracht, deren Ableitung und Bewertung diagnostischen Standards in der Regel nicht genügt und die häufig gegen das Transparenzgebot verstoßen. Nicht geprüft werden konnte in der

- **Voraussetzungen**
  - Analyse und Dokumentation der formalen Aufgabenmerkmale
  - Analysen zur Verfälschbarkeit der Steuerungsleistung
  - Schwierigkeitsanalysen /Sicherung der internen Validität der Problemlösegütemaße/Sicherung der Steuerbarkeit der Szenarien durch die jeweilige diagnostische Zielgruppe (u.a. zur Vermeidung der Provokation von extremen Emotionen [z.B. Frustration] und deren Auswirkungen)
  - weitgehende Freiheit von Zufallseinflüssen
  - Angaben zur Zielgruppe (und deren [Erfahrungs-]Voraussetzungen)
  - Analysen zu geschlechts- und altersspezifischen Steuerungsleistungen

- **Untersuchungsdurchführung**
  - Steigerung der Durchführungsobjektivität durch eine direkte Interaktion der Diagnostikanden mit dem Rechner
  - Lernphase zur Computerbedienung
  - Standardisierung der Durchführungszeiten und Bearbeitungstakte zur Vereinheitlichung der pro Person interpretierbaren Datenbasis/ Wissenserwerbsgelegenheiten
  - einfache, eindeutige, kontrollierbare und (z.B. gegenüber Tippfehlern) korrigierbare Eingabemöglichkeiten für die Diagnostikanden
  - Transparenz hinsichtlich der Handlungsmöglichkeiten
  - eindeutige Zielvorgabe für die Steuerungsaufgabe
  - Transparenz hinsichtlich der Bewertungskriterien
  - empfehlenswert ist die wiederholte Steuerung ein und desselben Szenarios mit wechselnden Startwerten (u.a. zur Steigerung der Reliabilität)

- **Begleitende Diagnostik:**
  zusätzlicher Einsatz von Verfahren zur Messung...
  - der Computererfahrung und der Einstellung gegenüber Computern
  - des allgemeinen Vorwissens über die Domäne, aus der die semantische Einkleidung, die Rahmengeschichte des Szenarios stammt
  - des systemspezifischen Wissens
  - der Intelligenz

- **Auswertung**
  - standardisiert bestimmte Problemlösegütemaße, die sich unmittelbar an der Zielvorgabe orientieren
  - nachweislich intern valide Problemlösegütemaße
  - möglichst unverfälschbare Problemlösegütemaße
  - nachweislich reliable Problemlösegütemaße
  - im Falle mehrerer Problemlösegütemaße: Vorgabe von konfiguralen oder integralen Bewertungsrichtlinien
  - nachvollziehbar dokumentierte systemspezifische Validitätsnachweise

Abbildung 14: Einige Voraussetzungen für den Einsatz von computergestützten Problemlöseszenarien zur Fähigkeitsdiagnostik

vorliegenden Studie die Befürchtung, daß Mädchen und Frauen bei der Steuerung computergestützter Problemlöseszenarien - möglicherweise aufgrund eines im Geschlechtervergleich geringeren Umfangs an Computererfahrung und aufgrund größerer Vorbehalte gegenüber der Arbeit mit Computern - schlechter abschneiden als Jungen und Männer. Bei den überwiegend männlichen Teilnehmern der vorliegenden Untersuchung zeigte sich ein schwacher positiver Zusammenhang zwischen der Computererfahrung und der Steuerungsleistung in der „Schneiderwerkstatt" sowie zwischen der Computererfahrung und der Verhaltensbewertung beim Szenario „DISKo". Die bei „DISKo"-programmintern berechnete Beurteilung der Verhaltensweisen und Strategien variierte außerdem in Abhängigkeit von der Einstellung der Diagnostikanden zur Arbeit mit Computern. Gegen den diagnostischen Einsatz von Problemlöseszenarien spricht, daß die Leistungen bei der Steuerung der „Schneiderwerkstatt" und „DISKo" sich mit Ausnahme ihres Intelligenz- und Wissensanteils als systemspezifisch erwiesen und nicht generalisiert werden konnten. Die Ergebnisse einer allein auf Szenarien gestützten Personalauswahl würden somit wesentlich Anteil dadurch bestimmt, welches Szenario jeweils zum Einsatz kommt. Schließlich bleibt auch der Einwand der vergleichsweise mäßigen Reliabilität der mit Problemlöseszenarien geleisteten Messungen bestehen. Mit der vorliegenden Studie konnte zwar ein Baustein zur Kriteriumsvalidierung von zwei Problemlöseszenarien beigetragen werden, das schwerwiegende Problem der völlig unzureichenden Anzahl methodologisch befriedigender und hinreichend dokumentierter Studien zur Validität der einzelnen Szenarien ist damit aber nicht behoben.

Sofern computergestützte Problemlöseszenarien trotz der genannten Bedenken im Rahmen der (berufsbezogenen) Fähigkeitsdiagnostik eingesetzt werden, sollten zumindestens die in Abbildung 14 genannten Voraussetzungen für eine möglichst *standardisierte* Anwendung und Auswertung erfüllt werden. Dies gilt auch, falls die Szenarien lediglich zur Ergänzung der Intelligenztests appliziert werden oder wenn der Einsatz der Szenarien lediglich die latente Funktion der Provokation von Wissenserwerbmöglichkeiten erfüllt. Den von Dörner (1992) entworfenen Regeln der »Mikrowelt-Philosophie« kommt zumindest bei diagnostischen Fragestellungen keine Bedeutung bei (siehe Kapitel 5).

Ein Vergleich der in Abbildung 14 aufgelisteten Voraussetzungen mit den Merkmalen der in der Praxis angebotenen Szenarien und Einsatzbedingungen der Szenarien zeigt, daß selbst für den weiter oben skizzierten eingeschränkten diagnostischen Einsatz von Problemlöseszenarien nur sehr wenige Systeme überhaupt ernsthaft in Frage kommen. Zu hoffen bleibt, daß in Zukunft trotz stetig neuer Gestaltungsmöglichkeiten von Computerprogrammen der Evaluation von computergestützten Problemlöseszenarien gegenüber der Konzeption immer neuer Problemlöseszenarien und Szenarienvarianten Priorität eingeräumt wird.

19. Literatur

Althoff, K. (1977). Zusammenhänge zwischen Ergebnissen von Eignungstests und beruflicher Bewährung. *Schriftenreihe der Polizei-Führungsakademie*, 6-27.

Althoff, K. (1984). Zur prognostischen Validität von Intelligenz- und Leistungstests im Rahmen der Eignungsdiagnostik. *Psychologie und Praxis - Zeitschrift für Arbeits- und Organisationspsychologie, 28*, 144-148.

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Anderson, J. R. (1981). *Cognitive skills and their acquisition*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge: Harvard University Press.

Andresen, N. & Schmid, U. (1993). Zur Invarianz von Problemlösestilen über verschiedene Bereiche. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40*, 1-17.

Arbinger, R. (1991). Wissensdiagnostik. In Kh. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends. Band 9* (S. 80-108). Weinheim: Beltz.

Atwood, M.E. & Polson, P.G. (1976). A process model for water jug problems. *Cognitive Psychology, 8*, 191-216.

Badke-Schaub, P. & Tisdale, T. (1995). Die Erforschung menschlichen Handelns in komplexen Situationen. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 43-56). Göttingen: Verlag für Angewandte Psychologie.

Baron, J. (1985). What kinds of intelligence components are fundamental? In S.F. Chipman, J.W. Segal & R. Glaser (Eds.), *Thinking and learning skills* (Vol. 2, Research and open questions, pp. 365-390). Hillsdale, NJ: Erlbaum.

Barthel, E. (1988). *Nutzen eignungdiagnostischer Verfahren bei der Bewerberauswahl*. Frankfurt/M.: Lang.

Bartussek, D. (1982). *Modelle der Testfairneß und Selektionsfairneß* (Trierer Psychologische Berichte). Trier: Universität Trier.

Beckmann, J.F. (1994). *Lernen und komplexes Problemlösen. Ein Beitrag zur Konstruktvalidierung von Lerntests*. Bonn: Holos.

Beckmann, J.F. & Funke, J. (1991). Solide Diskussionsbasis? Ein Kommentar zu dem Aufsatz „Komplexes Problemlösen und Verarbeitungskapazität" von Walter Hussy. *Sprache & Kognition, 10*, 221-222.

Beckmann, J.F. & Guthke, J. (1995). Complex problem solving, intelligence, and learning ability. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving. The european perspective* (pp. 177-200). Hillsdale, NJ: Erlbaum.

Berry, D.C. (1993). The control of complex systems. In D.C. Berry & Z. Dienes (Eds.), *Implicit learning. Theoretical and empirical issues* (pp. 19-35). Hillsdale, NJ: Erlbaum.

Berry, D.C. & Broadbent, D.E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology, 36 A*, 209-231.

Berry, D.C. & Broadbent, D.E. (1987). The combination of explicit and implicit learning processes in task control. *Psychological Research, 49*, 7-15.

Berry, D.C. & Broadbent, D.E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology, 79*, 251-272.

Beyer, R., Artz, E. & Guthke, T. (1990). Zur Differenzierung des kognitiven Aufwandes bei der Anregung von Vorwissen. *Zeitschrift für Psychologie, 198*, 9-33.

Birkhan, G. & Reitzig, G. (1989). Das computergestützte Simulationsmodell MANAGE! - Ein Verfahren zur Erfassung der Fähigkeit vernetzten Denkens. In G. Cisek (Hrsg.), *Instrumente der Personalentwicklung auf dem Prüfstand: Entscheidungshilfen für die Führungsspitze* (S. 58-73). Hamburg: Windmühle.

Bortz, J. (1989). *Lehrbuch der Statistik*. Berlin: Springer.

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

Bretz, E. & Oldendörp, H. (1992). Bewährungskontrolle: Vorhersage des Ausbildungserfolges im Angestelltenlehrgang I. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen, *51*, 75-84.

Breuer, K. (1992). *Denk- und Entscheidungsverhalten in komplexen dynamischen Systemen* (Schriftenreihe der Polizei-Führungsakademie) Kuratorium der Polizei- Führungsakademie, *4*, 87-96.

Breuer, K. & Streufert, S. (1995). Computergestützte Eignungsdiagnostik mit komplexen dynamischen Szenarios: Ausräumung von Mißverständnissen. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 34-36.

Broadbent, D.E., FitzGerald, P. & Broadbent, M.H.P. (1986). Implicit and explicit knowledge in the control of complex systems. *British Journal of Psychology, 77*, 33-50.

Brocke, B., Beauducel, A. & Tasche, K. (1998). Der Intelligenz-Struktur-Test: Analysen zur theoretischen Grundlage und technischen Güte. *Diagnostica, 44*, 84-99.

Buchner, A. (1993). *Implizites Lernen*. Weinheim: PVU.

Buchner, A. & Funke, J. (1993). Finite-state automata: Dynamic task environments in problem-solving research. *Quarterly Journal of Experimental Psychology, 46 A*, 83-118.

Buchner, A., Funke, J. & Berry, D.C. (1995). Negative correlations between control performance and verbalizable knowledge: Indicators for implicit learning in process control tasks? *The Quarterly Journal of Experimental Psychology, 48A*, 166-187.

Bucik, V. & Neubauer, A.C. (1996). Bimodality in the Berlin Model of Intelligence Structure (BIS): A replication study. *Personality and Individual Differences, 21*, 987-1005.

Campbell, D.T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin, 54*, 297-312.

Carroll, J.B. (1993). *Human cognitive abilities*. Cambridge: University Press.

Carroll, J.B. & Horn, J.L. (1981). On the scientific basis of ability testing. *American Psychologist, 36*, 1012-1020.

Cattell, R.B. (1957). *Personality and motivation: structure and measurement*. New York: Harcourt, Brace & World.

Cattell, R.B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54*, 1-22.

Cattell, R.B. (1971). *Abilities: Their structure, growth, and action*. Boston: Houghton Mifflin.

Cattell, R.B., Eber, H.W. & Tatsuoka, M.M. (1970). *Handbook for the sixteen Personality Factor Questionnaire (16 PF)*. Champaign, Ill: Institute for Personality and Ability Testing.

Chi, M.T.H. (1984). Bereichsspezifisches Wissen und Metakognition. In F.E. Weinert & R.H. Kluwe (Hrsg.), *Metakognition, Motivation und Lernen* (S. 211-232). Stuttgart: Kohlhammer.

Chi M.T.H., Glaser R. & Rees, E . (1982). Expertise in problem solving. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 1, pp. 7-75). Hillsdale, NJ: Erlbaum.

Cohen, J. & Cohen, P. (1975). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed), *Educational measurement* (pp. 443-507). Washington: American Council on Education.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (Ed.), *Intelligence. Measurement, theory, and public policy* (pp. 147-171). Urbana, IL: University of Illinois Press.

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

Diemand, A., Becker, K. & Schuler, H. (1997). Vorhersage des Berufserfolgs durch standardisierte Verfahren der Potentialanalyse. *Personal, 10*, 524-528.

Dörner, D. (1974). *Die kognitive Organisation beim Problemlösen*. Bern: Huber.

Dörner, D. (1976). *Problemlösen als Informationsverarbeitung*. Stuttgart: Kohlhammer.

Dörner, D. (1981). Über die Schwierigkeiten menschlichen Umgangs mit Komplexität. *Psychologische Rundschau, 32*, 163-179.

Dörner, D. (1984). Denken, Problemlösen und Intelligenz. *Psychologische Rundschau, 35*, 10-20.

Dörner, D. (1985). Verhalten, Denken und Emotionen. In L.H. Eckensberger & E.D. Lantermann (Hrsg.), *Emotion und Reflexivität* (S. 157-181). München: Urban & Schwarzenberg.

Dörner, D. (1986). Diagnostik der operativen Intelligenz. *Diagnostica, 32*, 290-308.

Dörner, D. (1987). On the difficulties people have in dealing with complexity. In J. Rasmussen, K. Duncan and J. Leplat (Eds.), *New technology and human error* (pp. 97-109). New York: Wiley.

Dörner, D. (1988). Wissen und Verhaltensregulation: Versuch einer Integration. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 264-279). Weinheim: Psychologische Verlags Union.

Dörner, D. (1989a). Die kleinen grünen Schildkröten und die Methoden der experimentellen Psychologie. *Sprache & Kognition, 8*, 86-97.

Dörner, D. (1989b). *Die Logik des Mißlingens*. Reinbek: Rowohlt.

Dörner, D. (1989c). Expertise beim Lösen komplexer Probleme oder die Bedeutung von Großmutterregeln. In D. Dörner & W. Michaelis (Hrsg.), *Idola fori et idola theatri* (S. 121-143). Göttingen: Hogrefe.

Dörner, D. (1992). Über die Philosophie der Verwendung von „Mikrowelten" oder „Computerszenarios" in der psychologischen Forschung. In H. Gundlach (Hrsg.), *Psychologische Forschung und Methode: Das Versprechen des Experiments* (S. 53-87). Passau: Passiva Universitäts-Verlag.

Dörner, D. (1993). Denken und Handeln in Unbestimmtheit und Komplexität. *GAIA, 2*, 128-138.

Dörner, D. (1995). Modellbildung und Simulation. In E. Roth (Hrsg.), *Sozial-wissenschaftliche Methoden* (S. 327-340). München: Oldenbourg.

Dörner, D. & Kreuzig, H.W. (1983a). Problemlösefähigkeit und Intelligenz. *Psychologische Rundschau, 34*, 185-192.

Dörner, D., Kreuzig, H.W., Reither, F. & Stäudel, T. (1983b). *Lohhausen. Vom Umgang mit Unbestimmtheit und Komplexität*. Bern: Huber.

Dörner, D. & Pfeifer, E. (1991). Strategisches Denken und Streß. *Zeitschrift für Psychologie, Supplement, 11*, 71-83.

Dörner, D. & Pfeifer, E. (1992). Strategisches Denken, Strategische Fehler, Streß und Intelligenz. *Sprache & Kognition, 11*, 75-90.

Dörner, D. & Preußler, W. (1990). Die Kontrolle eines einfachen ökologischen Systems. *Sprache & Kognition, 9*, 205-217.

Dörner, D. & Reither, F. (1978). Über das Problemlösen in sehr komplexen Realitätsbereichen. *Zeitschrift für Experimentelle und Angewandte Psychologie, 25*, 527-551.

Dörner, D., Schaub, H, Stäudel, T. & Strohschneider, S. (1988). Ein System zur Handlungsregulation oder - Die Interaktion von Emotion, Kognition und Motivation. *Sprache und Kognition, 7*, 217-232.

Duncker, K. (1935). *Zur Psychologie des produktiven Denkens*. Berlin: Springer.

Eyferth, K., Schömann, M. & Widwoski, D. (1986). Der Umgang von Psychologen mit Komplexität. *Sprache & Kognition, 5*, 11-26.

Eysenck, H.J. (1988). The concept of „intelligence": useful or useless? *Intelligence, 12*, 1-16.

Fay, E. & Heilmann, K. (1995). Die Konstruktionsübung „Waage" als Instrument zur Führungskräfte-Diagnostik. In J. Funke & A. Fritz (Hrsg.), *Neue Konzepte und Instrumente zur Planungsdiagnostik* (S. 121-140). Bonn: Deutscher Psychologen Verlag.

Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. *Diagnostica, 30*, 24-46.

Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford: Stanford University Press.

Fillbrandt, H. (1992). Zur Methode der Erforschung von Problemlöseprozessen. *Zeitschrift für Psychologie, 200*, 3-18.

Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests*. Bern: Huber.

Fishbein, M. & Ajzzen, I. (1974). Attitudes towards objects as predictors of single and multiple behavioral criteria. *Psychological Review, 81*, 59-74.

Ford, J.K. & Kraiger, K. (1993). Police officer selection validation project: the multijurisdictional police officer examination. *Journal of Business and Psychology, 7*, 421-429.

Frensch, P.A. & Sternberg, R.J. (1989). Expertise and intelligent thinking: When is it worse to know better? In R.J. Sternberg (Ed.), *Advances on the psychology of human intelligence* (Vol. 5, pp. 157-189). Hillsdale, NJ: Erlbaum.

Fritz, A. & Funke, J. (1988). Komplexes Problemlösen bei Jugendlichen mit Hirnfunktionsstörungen. *Zeitschrift für Psychologie, 196*, 171-187.

Fritz, A. & Funke, J. (1995). Übersicht über vorliegende Verfahren zur Planungsdiagnostik. In J. Funke & A. Fritz (Hrsg.), *Neue Konzepte und Instrumente zur Planungsdiagnostik* (S. 47-78). Bonn: Deutscher Psychologen Verlag.

Fruhner, R. & Schuler, H. (1988). Bewertung eignungsdiagnostischer Verfahren zur Personalauswahl durch potentielle Stellenbewerber. In G. Romkopf, W.D. Fröhlich & I. Lindner (Hrsg.), *Forschung und Praxis im Dialog. Entwicklungen und Perspektiven. Bericht über den 14. Kongreß für Angewandte Psychologie in Mainz 1987.* (Bd. 1, S. 107-111). Bonn: Deutscher Psychologen Verlag.

Fruhner, R., Schuler, H., Funke, U. und Moser, K. (1991). Einige Determinanten der Bewertung von Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie, 35*, 170-178.

Funke, J. (1983). Einige Bemerkungen zu Problemen der Problemlöseforschung oder: Ist Testintelligenz doch ein Prädiktor? *Diagnostica, 14*, 283-302.

Funke, J. (1984). Diagnose der westdeutschen Problemlöseforschung in Form einiger Thesen. *Sprache & Kognition, 3*, 159-172.

Funke, J. (1985a). Problemlösen in komplexen computersimulierten Realitätsbereichen. *Sprache & Kognition, 3*, 113-129.

Funke J. (1985b). Steuerung dynamischer Systeme durch Aufbau und Anwendung subjektiver Kausalmodelle. *Zeitschrift für Psychologie, 193*, 443-465.

Funke, J. (1986). *Komplexes Problemlösen*. Berlin: Springer.

Funke, J. (1988). Using simulation to study complex problem solving: A review of studies in the FRG. *Simulation & Games, 19*, 277-303.

Funke J. (1990). Systemmerkmale als Determinanten des Umgangs mit dynamischen Systemen. *Sprache & Kognition, 9*, 143-154.

Funke, J. (1991a). Keine Struktur im (selbstverursachten) Chaos? Erwiderung zum Kommentar von Stefan Strohschneider. *Sprache & Kognition, 10*, 114-118.

Funke, J. (1991b). Solving complex problems: Exploration and control of complex systems. In R.J.Sternberg & P.A.Frensch (Eds.), *Complex problem solving: Principles and mechanisms* (pp. 185-222). Hillsdale, NJ: Erlbaum.

Funke, J. (1992). *Wissen über dynamische Systeme: Erwerb, Repräsentation und Anwendung*. Berlin: Springer.

Funke, J. (1993). Computergestützte Arbeitsproben: Begriffsklärung, Beispiele sowie Entwicklungspotentiale. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 119-129.

Funke, J. (1995a). Erforschung komplexen Problemlösens durch computerunterstützte Planspiele: Kritische Anmerkungen zur Forschungsmethodologie. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 205-216). Göttingen:: Verlag für Angewandte Psychologie.

Funke, J. (1995b). Experimental research on complex problem solving. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving. The european perspective* (pp. 243-268). Hillsdale, NJ: Erlbaum.

Funke, J. (1995c). Some pathologies in the study of pathologies. A comment on Anders Jansson (1994). *Sprache & Kognition, 14*, 91-95.

Funke, J. (1998). Computer-based testing and training with scenarios from complex problem-solving research: Advantages and disadvantages. *International Journal of Selection and Assessment, 6*, 90-96.

Funke, J. & Buchner, A. (1992). Finite Automaten als Instrumente für die Analyse von wissensgeleiteten Problemlöseprozessen: Vorstellung eines neuen Untersuchungsparadigmas. *Sprache & Kognition, 11*, 27-37.

Funke, J. & Fritz, A. (Hrsgb.). (1995). *Neue Konzepte und Instrumente zur Planungsdiagnostik*. Bonn: Deutscher Psychologen Verlag.

Funke, J. & Fritz, A. (1995). Über Planen, Problemlösen und Handeln. In J. Funke & A. Fritz (Hrsg.), *Neue Konzepte und Instrumente zur Planungsdiagnostik* (S. 1-45). Bonn: Deutscher Psychologen Verlag.

Funke, J. & Geilhardt, Th. (1996). Diagnostik mit Hilfe von PC-Simulationen. In Arbeitskreis Assessment Center e.V. (Hrsg.), *Assessment Center als Instrument der Personalentwicklung* (S. 201-209). Hamburg: Windmühle.

Funke, J. & Glodowski, A.S. (1990). Planen und Problemlösen: Überlegungen zur neuropsychologischen Diagnostik von Basiskompetenzen beim Planen. *Zeitschrift für Neuropsychologie, 1*, 139-148.

Funke, J. & Müller, H. (1988). Eingreifen und Prognostizieren als Determinanten von Systemidentifikation und Systemsteuerung. *Sprache & Kognition, 7*, 176-186.

Funke, J. & Rasche, B. (1992). Einsatz computersimulierter Szenarien im Rahmen eines Assessment Center. *Zeitschrift für Führung & Organisation, 61*, 110-118.

Funke, U. (1991). Die Validität einer computergestützten Systemsimulation zur Diagnose von Problemlösekompetenz. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 114-122). Stuttgart: Verlag für Angewandte Psychologie.

Funke, U. (1992a). *Diagnostisches interaktives System zur Komplexitätssimulation „DISKo/c". Handbuch*. Filderstadt: Care applications Hofmann KG (Vertrieb).

Funke, U. (1992b). Die Validität einer eignungsdiagnostischen Simulation zum komplexen Problemlösen. In L. Montada (Hrsg.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992* (Bd. 1, S. 495 f.). Göttingen: Hogrefe.

Funke, U. (1993). Computergestützte Eignungsdiagnostik mit komplexen dynamischen Szenarios. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 109-118.

Funke, U. (1995a). Szenarien in der Eignungsdiagnostik und im Personaltraining. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 145-216). Göttingen: Verlag für Angewandte Psychologie.

Funke, U. (1995b). Using complex problem solving tasks in personnel selection and training. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving. The european perspective* (pp. 219-240). Hillsdale, NJ: Erlbaum.

Funke, U. (1995c). Zur Frage der Standards für komplexe dynamische Szenarios in der Eignungsdiagnostik. *Zeitschrift für Arbeits- und Organisationspsychologie, 39*, 36-39.

Funke, U. & Barthel, E. (1990). Nutzen und Kosten von Personalentscheidungen. In W.Sarges (Hrsg.), *Management-Diagnostik* (S. 647-658). Göttingen: Hogrefe.

Funke, U., Krauss, J., Schuler, H. & Stapf, K.H. (1987). Zur Prognostizierbarkeit wissenschaftlich-technischer Leistungen mittels Personvariablen: Eine Metaanalyse der Validität diagnostischer Verfahren im Bereich Forschung und Entwicklung. *Gruppendynamik, 18*, 407-428.

Gardner, M.K. & Sternberg, R.J. (1994). Novelty and intelligence. In R.J. Sternberg & R.K. Wagner (Eds.), *Mind in context* (pp. 38-73). Cambridge: Cambridge University Press.

Gardner, P.H. & Berry, D.C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9*, S55-S79.

Gediga, G., Schöttke, H. & Tücke, M. (1983). Problemlösen in einer komplexen Situation. *Arch. Psychol., 135*, 325-339.

Geilhardt, Th. (1995). Planspiele - Definition und Taxonomie. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 45-55). Göttingen: Verlag für Angewandte Psychologie.

Geilhardt, Th. & Mühlbradt, Th. (1995). Konzepte und Trends im Überblick. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 9-17). Göttingen: Verlag für Angewandte Psychologie.

Gerrards, A. (1988). Vorwissenseinflüsse auf den Erwerb und die Repräsentation von Wissen. *Zeitschrift für Experimentelle und Angewandte Psychologie, 35*, 405-426.

Ghiselli, E.E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology, 26*, 461-477.

Gigerenzer, G. (1981). *Messung und Modellbildung in der Psychologie*. München: Reinhardt.

Gigerenzer, G. (1988). Woher kommen Theorien über kognitive Prozesse? *Psychologische Rundschau, 39*, 91-100.

Gilliland, S.W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*, 694-734.

Graudenz, H. (1982). Bewährungskontrolle II - Vorhersage des Ausbildungserfolges von Beamtenanwärtern des gehobenen Dienstes beim RP Darmstadt. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen , *42*, 32-50.

Grawe, K., Hänni, R., Semmer, N. & Tschan, F. (Hrsgb.). (1990). *Über die richtige Art, Psychologie zu betreiben*. Göttingen: Hogrefe.

Greif, S. (1972). *Gruppenintelligenztests. Untersuchungen am WIT, IST, LPS und AIT*. Frankfurt / M.: Lang.

Groeben, N. & Westmeyer, H. (1975). *Kriterien psychologischer Forschung*. München: Juventa.

Gruber, G. (1986). The police applicant test: a predictive validity study. *Journal of Police Science and Administration, 14*, 121-129.

Grunwald, W. (1995). Aufgaben und Schlüsselqualifikationen von Managern. In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 194-205). Göttingen: Hogrefe.

Guthke, J. (1996). *Intelligenz im Test. Wege der psychologischen Intelligenzdiagnostik*. Göttingen: Vandenhoeck & Ruprecht.

Guthke, J., Böttcher, H.R. & Sprung, L. (1991). *Psychodiagnostik* (Bd. 2). Berlin: Deutscher Verlag der Wissenschaften.

Guthke, J. & Wiedl, K.H. (1996). *Dynamisches Testen*. Göttingen: Hogrefe.

Hager, W. & Hasselhorn, M. (1996). Bedeutet eine Verbesserung der Leistungen in Intelligenztests „zweifelsfrei" auch eine Verbesserung des induktiven Denkens? Ein Kommentar zu Klauers Artikel „Begünstigt induktives Denken das Lösen komplexer Probleme?". *Zeitschrift für Experimentelle Psychologie, 43*, 351-360.

Haider, H. (1991). *Explizites versus implizites Wissen und Lernen* (Unveröffentlichte Dissertation). Hamburg: Universität der Bundeswehr.

Haider, H. (1992). Implizites Wissen und Lernen. Ein Artefakt? *Zeitschrift für Experimentelle und Angewandte Psychologie, 39*, 68-100.

Haider, H. (1993). Was ist implizit am impliziten Wissen und Lernen? *Sprache & Kognition, 12*, 44-52.

Hamborg, K.C. (1996). Zum Einfluß der Komplexität von Software-Systemen auf Fehler bei Computernovizen und Experten. *Zeitschrift für Arbeits- und Organisationspsychologie, 40*, 3-11.

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

Hartung, S. & Schneider, I. (1995). Entwicklung und Anwendung computersimulierter Szenarien. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 219-236). Göttingen: Verlag für Angewandte Psychologie.

Hasselmann, D. (1991). Einsatzmöglichkeiten computersimulierter, komplexer Problemstellungen als neues Mittel der Eignungsdiagnostik? In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 110-113). Stuttgart: Verlag für Angewandte Psychologie.

Hasselmann, D. (1993). *Computersimulierte komplexe Problemstellungen in der Management-Diagnostik*. Hamburg: Windmühle.

Hasselmann, D. (1995). Die Konstruktion computersimulierter Szenarien in der Personalarbeit. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 237-259). Göttingen: Verlag für Angewandte Psychologie.

Hasselmann, D. & Strauß, B. (1993a). *Herausforderung Komplexität. Computersimulierte Problemlöseaufgaben für Management-Diagnostik und -Training. Trainerleitfaden.* Hamburg: Windmühle.

Hasselmann, G., Strauß, B. & Hasselmann, D. (1993b). Entwicklung einer PC-gestützten Unternehmenssimulation als Verfahren der betrieblichen Eignungsdiagnostik. In A. Gebert & U. Winterfeld (Hrsg.), *Arbeits-, Betriebs- und Organisationspsychologie vor Ort* (S. 551-560). Bonn: Deutscher Psychologen Verlag.

Heeg, F.J. & Kleine, G. (1995). Analyse menschlicher Verhaltensweisen und hieraus resultierender Handlungen im Umgang mit rechnergestützten Simulationsmodellen mit Hilfe neuronaler Netze. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 273-287). Göttingen: Verlag für Angewandte Psychologie.

Heidenreich, K. (1995). Grundbegriffe der Meß- und Testtheorie. In Roth, E. (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 342-374). München: Oldenbourg.

Herrmann, T. (1990). Die Experimentiermethodik in der Defensive? *Sprache & Kognition, 9*, 1-11.

Herrmann, T. (1995). Methoden als Problemlösungsmittel. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 20-48). München: Oldenbourg.

Hesse, F.W. (1982). Effekte des semantischen Kontexts auf die Bearbeitung komplexer Probleme. *Zeitschrift für Experimentelle und Angewandte Psychologie, 29*, 62-91.

Hesse, F.W. (1985). Vergleichende Analyse kognitiver Prozesse bei semantisch unterschiedlichen Problemeinbettungen. *Sprache & Kognition, 3*, 139-153.

Hesse, F.W., Spies, K. & Lüer, G. (1983). Einfluß motivationaler Faktoren auf das Problemlöseverhalten im Umgang mit komplexen Problemen. *Zeitschrift für Experimentelle und Angewandte Psychologie, 30*, 400-424.

Hirsh, H.R., Northrop, L.C. & Schmidt, F.L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology, 39*, 399-420.

Hodapp, V. (1984). *Analyse linearer Kausalmodelle.* Bern: Huber.

Hörmann, H.J. & Thomas, M. (1989). Zum Zusammenhang zwischen Intelligenz und komplexem Problemlösen. *Sprache & Kognition, 8*, 23-31.

Horn, J.L. (1980). Integration von Struktur- und Entwicklungskonzepten in der Theorie der flüssigen und kristallisierten Intelligenz. In R.B. Cattell (Hrsg.), *Handbuch der multivariaten experimentellen Psychologie* (S. 592-602). Frankfurt: Fachbuchhandlung für Psychologie.

Hornke, L.F. & Habon, M. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben. *Zeitschrift für Differentielle und Diagnostische Psychologie, 5*, 203-212.

Hossiep, R. (1995). *Berufseignungsdiagnostische Entscheidungen. Zur Bewährung eignungsdiagnostischer Ansätze.* Göttingen: Hogrefe.

Hossiep, R. (1997). *Konsequenzen aus neueren Erkenntnissen der Potentialbeurteilung.* Vortrag anläßlich der Sommerakademie zum Thema „Potentialbeurteilung in Unternehmen" in Landshut.

Hübner, R. (1987). Eine naheliegende Fehleinschätzung des Zielabstandes bei der zeitoptimalen Regelung dynamischer Systeme. *Zeitschrift für Experimentelle und Angewandte Psychologie, 34*, 38-53.

Hübner, R. (1988). Die kognitive Regelung dynamischer Systeme und der Einfluß analoger versus digitaler Informationsdarbietung. *Zeitschrift für Psychologie, 196*, 161-170.

Hübner, R. (1989a). Repräsenation dynamischer Strukturen durch lineare Systeme. *Zeitschrift für Experimentelle und Angewandte Psychologie, 36*, 51-71.

Hübner, R. (1989b). Methoden zur Analyse und Konstruktion von Aufgaben zur kognitiven Steuerung dynamischer Systeme. *Zeitschrift für Experimentelle und Angewandte Psychologie , 36*, 221-238.

Hunt, E. (1983). On the nature of intelligence. *Science, 219*, 141-146.

Hunter, J.E. & Hunter, R.F. (1984). Validity and utility of alternative predictors of job performance. *Psychological Bulletin, 96*, 72-98.

Hussy, W. (1984). *Denkpsychologie* (Bd. 1 und 2). Stuttgart: Kohlhammer.

Hussy, W. (1985). Komplexes Problemlösen - Eine Sackgasse? *Zeitschrift für Experimentelle und Angewandte Psychologie, 32*, 55-74.

Hussy, W. (1989). Intelligenz und komplexes Problemlösen. *Diagnostica, 35*, 1-16.

Hussy, W. (1991a). Eine experimentelle Studie zum Intelligenzkonzept „Verarbeitungskapazität". *Diagnostica, 37*, 314-333.

Hussy, W. (1991b). Komplexes Problemlösen und Verarbeitungskapazität. *Sprache & Kognition, 10*, 208-220.

Hussy, W. & Granzow, S. (1987). Komplexes Problemlösen, Gedächtnis und Verarbeitungsstil. *Zeitschrift für Experimentelle und Angewandte Psychologie, 34*, 212-227.

Jäger, A.O. (1967). *Dimensionen der Intelligenz*. Göttingen: Hogrefe.

Jäger, A.O. (1970). Personalauslese. In K.Gottschaldt, Ph. Lersch. F.Sander & H.Thomae (Hrsg.), *Handbuch der Psychologie* (Bd. 9, Betriebspsychologie, S. 613-667). Göttingen: Hogrefe.

Jäger, A.O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica, 28*, 195-226.

Jäger, A.O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*, 21-35.

Jäger, A.O. (1986). Validität von Intelligenztests. *Diagnostica, 32*, 272-289.

Jäger, A.O. (1991). Beziehungen zwischen komplexem Problemlösen und Intelligenz - eine Einleitung. *Diagnostica, 37*, 287-290.

Jäger, A.O. & Althoff, K. (1994). *Der WILDE-Intelligenz-Test (WIT)* (revidierte Auflage). Göttingen: Hogrefe.

Jäger, A.O., Süß, H.M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. Form 4. Handanweisung*. Göttingen: Hogrefe.

Jäger, A.O. & Tesch-Römer, C. (1988). Replikation des Berliner Intelligenzstrukturmodells (BIS) in den „Kit of Reference Test for Cognitive Factors" nach French, Ekstrom & Price (1963). *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 77-96.

Jannson, A. (1994). Pathologies in dynamic decision making: consequences or precursors of failure? *Sprache & Kognition, 13*, 160-173.

Jöreskog, K.G. & Sörbom, D. (1993). *LISREL 8. User's reference guide*. Chicago, IL: Scientific Software.

Kaminski, G. (1988). Ökologische Perspektiven in psychologischer Diagnostik? *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 155-168.

Kastner, M. (1978). Zur Intelligenzmessung im Rahmen der Pädagogischen Diagnostik. In K.J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik* (Bd. 2, S. 333-339). Düsseldorf: Schwann.

Kastner, M. (1995). Systemisches Denken. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 27-43). Göttingen: Verlag für Angewandte Psychologie.

Kerlinger, F.N. & Pedhazur, E.J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.

Kersting, M. (1991). *Wissensdiagnostik beim Problemlösen. Entwicklung und erste Bewährungsprobe eines kontentvalide konstruierten problemspezifischen Wissenstests* (Unveröffentlichte Diplomarbeit). Berlin: FU Berlin.

Kersting, M. (1995). Der Einsatz „westdeutscher" Tests zur Personalauswahl in den neuen Bundesländern und die Fairneßfrage. Auswirkungen der Testleistungsdisparität zwischen Ost und West auf die Auswahlentscheidung. *Report Psychologie, 20*, 32-41.

Kersting, M. (1996). Ost-West-Leistungsunterschiede in Berufseignungstests in Abhängigkeit von der kulturspezifischen Wirkung einiger Aufgabenmerkmale. *Zeitschrift für Arbeits- und Organisationspsychologie, 40*, 106-117.

Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien als Personalauswahlverfahren. *Zeitschrift für Arbeits- und Organisationspsychologie, 42*, 61-75.

Kersting, M. & Beauducel, A. (1997). Der neue DGP-Leistungstest auf der Basis des Berliner Intelligenzstrukturmodells: Informationen zu ausgewählten Testgütekriterien und zur Normierung. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen, *55*, 93-103.

Kersting, M. & Süß, H.M. (1995). Kontentvalide Wissensdiagnostik und Problemlösen: Zur Entwicklung, testtheoretischen Begründung und empirischen Bewährung eines problemspezifischen Diagnoseverfahrens. *Zeitschrift für Pädagogische Psychologie, 9*, 83-94.

Klauer, K.C. (1993). *Belastung und Entlastung beim Problemlösen. Eine Theorie des deklarativen Vereinfachens*. Göttingen: Hogrefe.

Klauer, K.C. (1995). Grundlagen der Problemlöseforschung. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 17-42). Göttingen: Verlag für Angewandte Psychologie.

Klauer, K.J. (1983). Kriteriumsorientierte Tests. In H. Feger & J. Bredenkamp (Hrsg.), *Enzyklopädie der Psychologie* (Bd. 3. Messen und Testen, S. 693-726). Göttingen: Hogrefe.

Klauer, K.J. (1984a). Kontentvalidität. *Diagnostica, 30*, 1-23.

Klauer, K.J. (1984b). Über Parallelität, Reliabilität und Validität kontentvalider Paralleltests. *Diagnostica, 30*, 67-80.

Klauer, K.J. (1996a). Begünstigt induktives Denken das Lösen komplexer Probleme? *Zeitschrift für Experimentelle Psychologie, 43*, 85-113.

Klauer, K.J. (1996b). Immer neue Wiederholungen machen ein Argument nicht gewichtiger. Eine Replik auf den Kommentar von Hager und Hasselhorn. *Zeitschrift für Experimentelle Psychologie, 43*, 361-366.

Kleinevoss, R. (1983). Untersuchungen zur Vorhersage des Ausbildungserfolges von Anwärtern des gehobenen Dienstes einer Bundesbehörde. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen , *43*, 41-72.

Kleinmann, M. & Strauß, B. (1995). Softwareergonomische Voraussetzungen computersimulierter Szenarien. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 127-141). Göttingen: Verlag für Angewandte Psychologie.

Klix, F. (1971). *Information und Verhalten*. Berlin: Deutscher Verlag der Wissenschaften.

Klix, F. & Lander, H.J. (1967). Die Strukturanalyse von Denkprozessen als Mittel der Intelligenzdiagnostik. In F. Klix, F. Gutjahr & J. Mehl (Hrsg.), *Intelligenzdiagnostik* (S. 245-271). Berlin: Deutscher Verlag der Wissenschaften.

Kluwe, R.H. (1988). Methoden der Psychologie zur Gewinnung von Daten über menschliches Wissen. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie*. Weinheim: Psychologische Verlags Union.

Kluwe, R.H. (1990a). Kontrolle und Steuerung komplexer Systeme durch Menschen: Anmerkungen zum Stand der kognitionspsychologischen Forschung. In K. Grawe, R. Hänni, N. Semmer & F. Tschan (Hrsg.), *Über die richtige Art, Psychologie zu betreiben* (S. 239-254). Göttingen: Hogrefe.

Kluwe, R.H. (1990b). Problemlösen, Entscheiden und Denkfehler. In Graf C. Hoyos & B. Zimolong (Hrsg.), *Enzyklopädie der Psychologie, Serie III, Ingenieurpsychologie* (S. 121-147). Göttingen: Hogrefe.

Kluwe, R.H. (1991). Zum Problem der Wissensvoraussetzungen für Prozeß- und Systemsteuerung. *Zeitschrift für Psychologie, Suppl., 11*, 311-324.

Kluwe, R.H. (1995). Computergestützte Systemsimulationen. In W.Sarges (Hrsg.), *Management-Diagnostik* (2. Aufl., S. 572-578). Göttingen: Hogrefe.

Kluwe, R. H. & Haider, H. (1990). Modelle zur internen Repräsentation komplexer technischer Systeme. *Sprache & Kognition, 9*, 173-192.

Kluwe, R.H., Misiak, C. & Haider, H. (1989). Erste Ergebnisse zu einem Modell der Steuerung eines komplexen Systems. In D. Dörner & W. Michaelis (Hrsg.), *Idola fori et idola theatri* (S. 101-119). Göttingen: Hogrefe.

Kluwe, R.H., Misiak, C. & Haider, H. (1990). Learning by doing in the control of a complex system. In H. Mandl, E. de Corte, N. Bennett & H.F. Friedrich (Eds.), *Learning and instruction* (pp. 197-218). Oxford: Pergamon Press.

Kluwe, R.H., Misiak, C. & Haider, H. (1991a). The control of complex systems and performance in intelligence tests. In H. Rowe (Ed.), *Intelligence: reconceptualization and measurement* (pp. 227-244). Hillsdale, NJ: Erlbaum.

Kluwe, R.H., Misiak, C. & Haider, H. (1991b). Modelling the process of complex system control. In P.M. Milling & E.O.K. Zahn (Eds.), *Computer-based management of complex systems. Proceedings of the 1989 international conference of the system dynamics society* (pp. 335-342). Berlin: Springer.

Kluwe, R.H., Schilde, A., Fischer, C. & Oellerer, N. (1991c). Problemlöseleistungen beim Umgang mit komplexen Systemen und Intelligenz. *Diagnostica, 37*, 291-313.

Köchling, A.C. & Körner, St. (1996). Personalauswahl aus der Sicht der Betroffenen: Zur bewerberorientierten Gestaltung von Beurteilungssituationen. *Zeitschrift für Arbeits- und Organisationspsychologie, 40*, 22-36.

Köhler, W. (1921). *Intelligenzprüfungen an Menschenaffen. (Zweite Auflage)*. Berlin: Springer.

Köller, O., Dauenheimer, D.G. & Strauß, B. (1993). Unterschiede zwischen Einzelpersonen und Dyaden beim Lösen komplexer Probleme in Abhängigkeit von der Ausgangsfähigkeit. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40*, 194-221.

Köller, O., & Strauß, B. (1994). Was mißt der Kompetenzfragebogen? Eine Reanalyse der Kurzform des Kompetenzfragebogens von Stäudel. *Diagnostica, 40*, 42-60.

Köller, O., Strauß, B. & Sievers, K. (1995). Zum Zusammenhang von (selbst eingeschätzter) Kompetenz und Problemlöseleistungen in komplexen Situationen. *Sprache & Kognition, 14*, 210-220.

Kolb, St., Petzing, F. & Stumpf, S. (1992). Komplexes Problemlösen: Bestimmung der Problemlösegüte von Probanden mittels Verfahren des Operation Research - ein interdisziplinärer Ansatz. *Sprache & Kognition, 11*, 115-128.

Kotovsky, K., Hayes, J.R. & Simon, H.A. (1985). Why are some problems hard? Evidence from Tower of Hanoi. *Cognitive Psychology, 17*, 248-294.

Kotovsky, K. & Simon, H.A. (1990). What makes some problems really hard. Explorations in the problem space of difficulty. *Cognitive Psychology, 22*, 143-183.

Krahn, H. (1990). Mädchen und Computer. In Deutsches Institut für Fernstudien (Hrsg.), *Lehren und Lernen mit dem Computer* (S. 176-188). Tübingen: Deutsches Institut für Fernstudien.

Krause, W. (1982a). Problemlösen - Stand und Perspektiven. Teil I. *Zeitschrift für Psychologie, 190*, 17-36.

Krause, W. (1982b). Problemlösen - Stand und Perspektiven. Teil II. *Zeitschrift für Psychologie, 190*, 141-169.

Krauth, J. & Lienert, G.A. (1973). *Die Konfigurationsfrequenzanalyse (KFA)*. Freiburg: Alber.

Kreuzig, H.W. (1981). Über den Zugang zu komplexen Problemlösen mittels prozeßorientierter kognitiver Persönlichkeitsmerkmale. *Zeitschrift für Experimentelle und Angewandte Psychologie, 28*, 294-308.

Kreuzig, H.W. (1995a). Personalentwicklung. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 87-103). Göttingen: Verlag für Angewandte Psychologie.

Kreuzig, H.W. (1995b). Die Computer-Simulation MANAGE! In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 387-400). Göttingen: Verlag für Angewandte Psychologie.

Kreuzig, H.W. & Schlotthauer, J.A. (1991). Ein Computer-Simulations-Verfahren in der Praxis: Offene Fragen - empirische Antworten. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 106-109). Stuttgart: Verlag für Angewandte Psychologie.

Kubinger, K.D. (1993). Testtheoretische Probleme der Computerdiagnostik. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 130-137.

Kühle, H.J. & Badke, P. (1986). Die Entwicklung von Lösungsvorstellungen in komplexen Problemsituationen und die Gedächtnisstruktur. *Sprache & Kognition, 5*, 95-105.

Kuhn, Th.S. (1976). *Die Struktur wissenschaftlicher Revolutionen*. Frankfurt /M.: Suhrkamp.

Kulik, J.A., Bangert-Drowns, R.L. & Kulik, C.L. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179-188.

Lang, M. (1992). *Computer in Schule und Lehrerbildung*. Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Larkin, J. H. & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science, 11*, 65-99.

Leutner, D. (1989). Implementation und experimentelle Evaluation von Lernhilfen im computersimulierten Planspiel „Hunger in Nordafrika". In Schrettenbrunner, H. (Hrsg.), *Software für den Geographieunterricht* (S. 81-109). Lüneburg: Selbstverlag des Hochschulverbandes für Geographie und ihre Didaktik e.V.

Leutner, D. (1990). Simulation und Modellbildung. In Deutsches Institut für Fernstudien (Hrsg.), *Lehren und Lernen mit dem Computer* (S. 22-52). Tübingen: Deutsches Institut für Fernstudien.

Leutner, D. (1995). Computerunterstützte Planspiele als Instrument der Personalentwicklung. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 105-116). Göttingen: Verlag für Angewandte Psychologie.

Lewin, K. (1963). *Grundzüge der topologischen Psychologie*. Bern: Huber.

Lienert, G.A. (1958). Ein Form-Lege-Test als Prüfmittel der praktischen Intelligenz. *Zeitschrift für Experimentelle und Angewandte Psychologie, 5*, 82-107.

Lienert, G.A. (1967). *Testaufbau und Testanalyse*. Weinheim: Beltz.

Lipsey, M.W. & Wilson, D.B. (1993). The efficacy of psychological, educational and behavioral treatment. Confirmation from Meta-Analysis. *American Psychologist, 48*, 1181-1209.

Locher, J. (1997). *Transfereffekte bei der Bearbeitung computersimulierter Problemszenarien* (Unveröffentl. Dissertation). Paderborn: Universität-Gesamthochschule Paderborn.

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review, 72*, 143-155.

Lüer, G. & Spada, H. (1990). Denken und Problemlösen. In H. Spada (Hrsg.), *Allgemeine Psychologie* (S. 189-280). Bern: Huber.

Mabe III, P.A. & West, S.G. (1982). Validity of self-evaluation of ability: a review and meta-analysis. *Journal of Applied Psychology, 67*, 280-296.

Mané, A. & Donchin, E. (1989). The space fortress game. *Acta Psychologica, 71*, 17-22.

Marjoribanks, K. (1994). Perceptions of parents' involvement in learning and adolescents' aspirations. *Psychological Reports, 75*, 192-194.

Marshall, E.C., Duncan, K.D. & Baker, S.M. (1981). The role of withheld information in the training of process plant fault diagnosis. *Ergonomics, 24*, 711-724.

Meehl, P.E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*, 806-834.

Melter, A. (1995). Computerunterstützte Planspiele in den Streitkräften. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 289-303). Göttingen: Verlag für Angewandte Psychologie.

Messick S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Erlbaum.

Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K.J.Groffmann & L. Michel (Hrsg.), *Enzyklopädie der Psychologie. Themenbereich B, Methodologie und Methoden. Serie II, Psychologische Diagnostik. Band 1, Grundlagen psychologischer Diagnostik* (S. 1-129). Göttingen: Hogrefe.

Milling, P.M. (1996). Modeling innovation processes for decision support and management simulation. *System Dynamics Review, 12*, 211-234.

Morris, N.M. & Rouse, W.B. (1985). The effect of type of knowledge upon human problem solving in a process control task. *IEEE Transactions on System, Man and Cybernetics, 15*, 698-707.

Moser, K. (1987). Inhaltsvalidität als Kriterium psychologischer Tests. *Diagnostica, 33*, 110-122.

Müller, E. (1991). *Risikoverhalten in komplexen Problemsituationen* (Unveröffentlichte Dissertation). Berlin: FU Berlin.

Müller, H. (1993). *Komplexes Problemlösen: Reliabilität und Wissen*. Bonn: Holos.

Neisser, U. (1976). General, academic and artificial intelligence. In L.B. Resnick (Ed.), *The nature of intelligence* (pp. 135-144). Hillsdale, NJ: Erlbaum.

Neubauer, R. (1995). Führungskräfteauswahl in der Praxis. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 155-171). Göttingen: Verlag für Angewandte Psychologie.

Newell, A. & Simon H.A. (1972). *Human problem solving*. New Jersey: Prentice-Hall.

Niederdrenk-Felgner, C. (1993). *Mädchen und Computer. Modelle für eine mädchengerechtere Unterrichtsgestaltung*. Tübingen: Deutsches Institut für Fernstudien.

Norris, D.R. & Snyder, C.A. (1982). External validation of simulation games. *Simulation & Games*, *13*, 73-85.

Oberauer, K. (1993a). Die Koordination kognitiver Operationen. Eine Studie zum Zusammenhang von „working memory" und Intelligenz. *Zeitschrift f. Psychologie*, *201*, 57-84.

Oberauer, K. (1993b). Prozedurales und deklaratives Wissen und das Paradigma der Informationsverarbeitung. *Sprache & Kognition*, *12*, 30-43.

Obermann, C. (1991). *Airport Problemlösesimulation V.2.2 Handbuch*. Göttingen: Hogrefe.

Obermann, C. (1992). *Assessment Center*. Wiesbaden: Gabler.

Obermann, C. (1995). Computergestützte Planspiele in der Mitarbeiterauswahl - Anwendungsbeispiel Airport. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 401-409). Göttingen: Verlag für Angewandte Psychologie.

Oerter, R. (1977). *Psychologie des Denkens* (5. Aufl.). Donauwörth: Auer.

Opwis, K. & Plötzner, R. (1996). *Kognitive Psychologie mit dem Computer*. Heidelberg: Spektrum.

Page, B. (1983). Der Gültigkeitsnachweis von komplexen Simulationsmodellen. *Angewandte Informatik*, *25*, 149-157.

Pawlik, K. (1976). Ökologische Validität: Ein Beispiel aus der Kulturvergleichsforschung. In G. Kaminski (Hrsg.), *Umweltpsychologie* (S. 59-72). Stuttgart: Klett.

Pawlik, K. (1988). Psychodiagnostik zwischen Allgemeiner und Differentieller Psychologie. *Zeitschrift für Differentielle und Diagnostische Psychologie*, *9*, 147-153.

Pawlik, K. (1997). Gutachten zum Thema: „Unterstützung psychologischer Eignungsdiagnostik durch den Computer. Bewertung neuester Entwicklungen im Psychologischen Dienst der Bundeswehr. In K. Puzicha (Hrsg.), *Arbeitsberichte des Psychologischen Dienstes der Bundeswehr* (Bd. 1, S. 145-192). Bonn: Bundesministerium der Verteidigung.

Plötzner, R. & Spada, H. (1992). Analysis-based learning on multiple levels of mental domain representation. In E. De Corte, M.C. Linn, H. Mandl & L. Verschaffel (Eds.), *Computer-based learning environments and problem solving* (S. 103-127). Berlin: Springer.

Plötzner, R., Spada, H., Stumpf, M. & Opwis, K. (1990). Learning qualitative and quantitative reasoning in a microworld for elastic impacts. *European Joarnal of Psychology and Education*, *4*, 501-516.

Preußler, W. (1996). Zur Rolle expliziten und impliziten Wissens bei der Steuerung dynamischer Systeme. *Zeitschrift für Experimentelle Psychologie*, *43*, 399-434.

Putz-Osterloh, W. (1981). *Problemlöseprozesse und Intelligenzleistungen*. Bern: Huber.

Putz-Osterloh, W. (1983). Über Determinanten komplexer Problemlöseleistungen und Möglichkeiten zu ihrer Erfassung. *Sprache & Kognition* , *2*, 100-116.

Putz-Osterloh, W. (1985). Selbstreflexion, Testintelligenz und interindividuelle Unterschiede bei der Bewältigung komplexer Probleme. *Sprache & Kognition*, *4*, 203-216.

Putz-Osterloh, W. (1987). Gibt es Experten für komplexe Probleme? *Zeitschrift für Psychologie*, *193*, 63-84.

Putz-Osterloh, W. (1988). Wissen und Problemlösen. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 247-263). Weinheim: Psychologische Verlags Union.

Putz-Osterloh, W. (1989). Problemlöseforschung und Intelligenzdiagnostik: Ein Anwendungsbeispiel. In D. Dörner & W. Michaelis (Hrsg.), *Idola fori et idola theatri* (S. 87-100). Göttingen: Hogrefe.

Putz-Osterloh, W. (1990). Problemlösen. In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 193-199). Göttingen: Hogrefe.

Putz-Osterloh, W. (1991a). Computergestützte Eignungsdiagnostik: Warum Strategien informativer als Leistungen sein können. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 97-102). Stuttgart: Verlag für Angew. Psychologie.

Putz-Osterloh, W. (1991b). Wissenserwerb und Wissensanwendung bei der Steuerung eines realitätsnahen und realitätsfernen Systems. *Zeitschrift für Psychologie, Supplement, 11*, 341-351.

Putz-Osterloh, W. (1993a). Complex problem solving as a diagnostic tool. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 289-301). Hillsdale, NJ: Erlbaum.

Putz-Osterloh, W. (1993b). Unterschiede im Erwerb und in der Reichweite des Wissens bei der Steuerung eines dynamischen Systems. *Zeitschrift für Experimentelle und Angewandte Psychologie, 40*, 386-410.

Putz-Osterloh, W. (1995). Komplexes Problemlösen. In M. Amelang (Hrsg.), *Enzyklopädie der Psychologie* (Bd. 2, Serie 8, S. 403-434). Göttingen: Hogrefe.

Putz-Osterloh W. & Bott, B. (1990). Sind objektive Systemmerkmale auch subjektiv als Anforderungen wirksam? *Zeitschrift für Experimentelle und Angewandte Psychologie, 37*, 281-303.

Putz-Osterloh, W., Bott, B. & Houben, I. (1988). Beeinflußt Wissen über ein realitätsnahes System dessen Steuerung? *Sprache & Kognition*, 240-251.

Putz-Osterloh, W. & Haupts, I. (1989). Zur Reliabilität und Validität computergestützter Diagnostik komplexer Organisations- und Entscheidungsstrategien. *Untersuchungen des psychologischen Dienstes der Bundeswehr, 24*, 5-48.

Putz-Osterloh, W. & Haupts, I. (1990). Diagnostik komplexer Organisations- und Entscheidungsstrategien in dynamischen Situationen. *Untersuchungen des psychologischen Dienstes der Bundeswehr, 25*, 107-167.

Putz-Osterloh, W. & Köster, K. (1988). Diagnostik komplexer Entscheidungsstrategien bei einem computersimulierten Planspiel. *Untersuchungen des psychologischen Dienstes der Bundeswehr, 23*, 223-255.

Putz-Osterloh, W. & Lemme, M. (1987). Knowledge and it's intelligent application to problem solving. *The German Journal of Psychology, 11*, 286-303.

Putz-Osterloh, W. & Lüer, G. (1981). Über die Vorhersagbarkeit komplexer Problemlöseleistungen durch Ergebnisse in einem Intelligenztest. *Zeitschrift für Experimentelle und Angewandte Psychologie, 28*, 309-334.

Putz-Osterloh, W. & Schroiff, M. (1987). Komplexe Verhaltensmaße zur Erfassung von Hochbegabung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 8*, 207-216.

Raaheim, K. (1974). *Problem solving and intelligence*. Oslo: Universitetsforlaget.

Reichert, U. & Dörner, D. (1988). Heurismen beim Umgang mit einem „einfachen" dynamischen System. *Sprache & Kognition, 7*, 12-24.

Reichert, U. & Stäudel, T. (1991). Computergestützte Diagnostik der Fähigkeiten für den Umgang mit komplexen und vernetzten Systemen. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 102-105). Stuttgart: Verlag für Angewandte Psychologie.

Reither, F. (1981). Thinking and acting in complex situations. A study of experts' behavior. *Simulation & Games, 12*, 125-140.

Renkl, A., Gruber, H., Mandl, H. & Hinkofer, L. (1994). Hilft Wissen bei der Identifikation und Kontrolle eines komplexen ökonomischen Systems? *Unterrichtswissenschaften, 22*, 195-202.

Resnick, L.B. (1976). Task analysis in instructional design: some cases from mathematics. In Klahr, D. (Ed.), *Cognition and instruction* (pp. 51-80). Hillsdale, NJ: Erlbaum.

Resnick, L.B. & Ford, W.W. (1981). *The psychology of mathematics for instruction*. Hillsdale, NJ: Erlbaum.

Rhenius, D. (1994). Selbstsicherheit und die Fähigkeit, Probleme zu lösen. In D. Bartussek & M. Amelang (Hrsg.), *Fortschritte der Differentiellen Psychologie und Psychologischen Diagnostik* (S. 67-75). Göttingen: Hogrefe.

Ringelband, O.J., Misiak, C. & Kluwe, R.H. (1990). Mental models and strategies in the control of a complex system. In D. Ackermann & M.J. Tauber (Eds.), *Mental models and human-computer interaction 1* (pp. 151-164). Amsterdam: North-Holland.

Roelofsma, P.H.M.P. (1995). Kognitive Heuristiken beim statischen und dynamischen Problemlösen. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 217-236). Göttingen: Verlag für Angewandte Psychologie.

Rohn, W. E. (1995a). Ursprung und Entwicklung des Planspiels. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 57-67). Göttingen: Verlag für Angewandte Psychologie.

Rohn, W. E. (1995b). Einsatzgebiete und Formen des Planspiels. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 69-77). Göttingen: Verlag für Angewandte Psychologie.

Rosen, L.D. & Maguire, P. (1990). Myths and realities of computerphobia: a Meta-Analysis. *Anxiety Research, 3*, 175-191.

Rynes, S.L. (1993). When recruitment fails to attract: Individual expectations meet organizational realities in recruitment. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 27-40). Hillsdale, NJ: Erlbaum.

Sarges, W. (1994). Eignungsdiagnostische Überlegungen für den Management-Bereich. In D. Bartussek & M. Amelang (Hrsg.), *Fortschritte der differentiellen Psychologie und psychologischen Diagnostik* (S. 415-434). Göttingen: Hogrefe.

Schaub, H. (1990). Die Situationsspezifität des Problemlöseverhaltens. *Zeitschrift für Psychologie, 198*, 83-96.

Schaub, H. (1993). *Modellierung der Handlungsorganisation*. Bern: Huber.

Schaub, H. & Strohschneider, S. (1992). Die Auswirkungen unterschiedlicher Problemlöseerfahrung auf den Umgang mit einem unbekannten komplexen Problem. *Zeitschrift für Arbeits- und Organisationspsychologie, 36*, 117-126.

Scheele, B. & Groeben, N. (1984). *Die Heidelberger Struktur-Lege-Technik: Eine Dialog-Konsens Methode zur Erfassung subjektiver Theorien mittlerer Reichweite*. Weinheim: Beltz.

Schmidt, F.L. (1988). Validity generalization and the future of criterion-related validity. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 173-189). Hillsdale, NJ: Erlbaum.

Schmidt, F.L. (1992). What do data really mean? Research findings, meta-analysis and cumulative knowledge in psychology. *American Psychologist, 47*, 1173-1181.

Schmidt, F.L., Ones, D.S. & Hunter, J.E. (1992). Personnel selection. *Annual Review of Psychology, 43*, 627-670.

Schmidt, J.U. (1986). Analysen zum Berliner Intelligenzstrukturmodell und der Eignungstestbatterie der DGP. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen , *46*, 2-24.

Schmidt, J.U. (1993). Thurstones Primary Mental Abilities und das Berliner Intelligenzstrukturmodell - Eine empirische Gegenüberstellung. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 87-100.

Schmidt-Atzert, L., Hommers, W. & Heß, M. (1995). Der I-S-T 70. Eine Analyse und Neubewertung. *Diagnostica, 41*, 108-130.

Schmitt, N., Gooding, R.Z., Noe, R.A. & Kirsch, M. (1984). Metaanalysis of validity studies published between 1964 and 1982 and the investigation of study characteristics. *Personnel Psychology, 37*, 407-422.

Schmuck, P. (1992). Zum Zusammenhang zwischen der Effizienz exekutiver Kontrolle und dem mehrfachen Lösen eines komplexen Problems. *Sprache & Kognition, 11*, 193-207.

Schmuck, P. & Strohschneider, St. (1995). Exekutive Kontrolle und Verhaltensstabilität beim Bearbeiten eines komplexen Problems: Eine Replikation. *Diagnostica, 41*, 150-171.

Schneider, W. (1986). Strukturgleichungsmodelle der zweiten Generation: Eine Einführung. In C. Möbus & W. Schneider (Hrsg.), *Strukturmodelle für Längsschnittdaten und Zeitreihen: LISREL, Pfad- und Varianzanalysen* (S. 13-26). Bern: Huber.

Schönpflug, W. (1989). Anxiety, worry, prospective orientation, and prevention. In C.D. Spielberger, I.G. Sarason & J. Strelau (Eds.), *Stress and anxiety* (Vol. 12, pp. 245-258). Washington, DC: Hemisphere.

Schönpflug, W. (1993). Feldforschung, Simulation, Experiment: Methodenvariation als Mittel der Theorieentwicklung. In W. Bungard & T. Herrmann (Hrsg.), *Arbeits- und Organisationspsychologie im Spannungsfeld zwischen Grundlagenorientierung und Anwendung* (S. 207-222). Bern: Huber.

Schoppek, W. (1991). Spiel und Wirklichkeit - Reliabilität und Validität von Verhaltensmustern in komplexen Situationen. *Sprache & Kognition, 10*, 15-27.

Schoppek, W. (1996). *Kompetenz, Kontrollmeinung und komplexe Probleme. Zur Vorhersage individueller Unterschiede bei der Systemsteuerung*. Bonn: Holos.

Schorr, A. (1995). Stand und Perspektiven diagnostischer Verfahren in der Praxis. Ergebnisse einer repräsentativen Befragung westdeutscher Psychologen. *Diagnostica, 41*, 3-20.

Schott, F. (1984). Regelgeleitete Itemkonstruktion. Ein Verfahren zur Definition von Itemuniversa und deren kontentvalider Abbildung in Itemmengen für Tests und Treatments. *Diagnostica, 30*, 47-66.

Schrader, F.-W. & Helmke, A. (1990). Lassen sich Lehrer bei der Leistungsbeurteilung von sachfremden Gesichtspunkten leiten? Eine Untersuchung zu Determinanten diagnostischer Lehrerurteile. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 22*, 312-324.

Schreiber, L. (1995). Der Einsatz eines computersimulierten Szenarios im Assessment-Center bei MERCK. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 273-284). Göttingen: Verlag für Angewandte Psychologie.

Schuler, H. (1990). Personalauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen. *Zeitschrift für Arbeits- und Organisationspsychologie, 34*, 184-191.

Schuler, H. (1993). Social validity of selection situations: a concept and some empirical results. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 11-26). Hillsdale, NJ: Erlbaum.

Schuler, H. (1996). *Psychologische Personalauswahl. Einführung in die Berufseignungsdiagnostik*. Göttingen: Verlag für Angewandte Psychologie.

Schuler, H., Frier, D. & Kaufmann, M. (1991). Validität, Praktikabilität und Akzeptanz eignungsdiagnostischer Verfahren in der Einschätzung der Verwender. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 26-32). Stuttgart: Verlag für Angewandte Psychologie.

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

Schuler, H. & Stehle, W. (1983). Neuere Entwicklungen des Assessment-Center-Ansatzes - beurteilt unter dem Aspekt der sozialen Validität. *Zeitschrift für Arbeits- und Organisationspsychologie, 27*, 33-44.

Schuler, H. & Stehle, W. (1985). Soziale Validität eignungsdiagnostischer Verfahren: Anforderungen für die Zukunft. In H. Schuler & W. Stehle (Hrsg.), *Organisationspsychologie und Unternehmenspraxis. Perspektiven der Kooperation* (S. 133-138). Stuttgart: Verlag für Angewandte Psychologie.

Seggebruch, G. (1982). Bewährungskontrolle I - Anwärter des mittleren Dienstes einer Kommunalverwaltung. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen , *42*, 6-31.

Sonnenberg, H.G. (1993). Computergestützte psychologische Diagnoseverfahren bei der Auswahl von Führungskräften. *Zeitschrift für Arbeits- und Organisationspsychologie, 37*, 146-149.

Spada, H. & Reimann, P. (1988). Wissensdiagnostik auf kognitionswissenschaftlicher Basis. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9*, 183-192.

Spies, K. & Hesse, F.W. (1987). Problemlösen. In G. Lüer (Hrsg.), *Allgemeine experimentelle Psychologie* (S. 371-425). Stuttgart: G. Fischer.

Stäudel, T. (1987). *Problemlösen, Emotionen und Kompetenz*. Regensburg: Roderer.

Stäudel, T. (1988). Der Kompetenzfragebogen. Überprüfung eines Verfahrens zur Erfassung der Selbsteinschätzung der heuristischen Kompetenz, belastenden Emotionen und Verhaltenstendenzen beim Lösen komplexer Probleme. *Diagnostica, 34*, 136-147.

Stapf, K.H. (1995). Laboruntersuchungen. In E. Roth (Hrsg.), *Sozialwissenschaftliche Methoden* (S. 228-244). München: Oldenbourg.

Stern, W. (1911). *Die differentielle Psychologie in ihren methodischen Grundlagen*. Leipzig: Ambrosius Barth.

Sternberg, R.J. (1982). Reasoning, problem solving and intelligence. In R.J. Sternberg (Ed.), *Handbook of human intelligence* (pp. 225-307). Cambridge: Cambridge University Press.

Sternberg, R.J. (1984). Toward a triarchic theory of human intelligence. *The Behavioral and Brain Sciences, 7*, 269-287.

Sternberg, R.J. (1989). Intelligence, wisdom, and creativity: their natures and interrelationsships. In R.L. Linn (Ed.), *Intelligence. Measurement, theory, and public policy* (pp. 119-146). Urbana Il: University of Illinois Press.

Sternberg, R.J. (1990). Intellectual styles. In Clark, K.E. & Clark, M.B. (Eds.), *Measures of leadership* (pp. 481-492). West Orange, NJ: Leadership Library.

Sternberg, R.J. (1995). Expertise in complex problem solving: a comparison of alternative conceptions. In P.A. Frensch & J. Funke (Eds.), *Complex problem solving. The european perspective* (pp. 295-321). Hillsdale, NJ: Erlbaum.

Sternberg, R.J. & Frensch, P.A. (1990). Intelligence and cognition. In M.W. Eysenck (Ed.), *Cognitive Psychology: An International Review* (pp. 57-103). Chicester: Wiley.

Sternberg, R.J. & Kaufman, J.C. (1996). Innovation and intelligence testing: the curious case of the dog that didn't bark. *European Journal of Psychological Assessment, 12*, 175-182.

Sternberg, R.J. & Wagner, R.K. (1993). The g-ocentric- view of intelligence and job performance is wrong. *Current Directions in Psychological Science, 2*, 1-4.

Stöber, J. (1996). Anxiety and the regulation of complex problem situations: Playing it safe? In W. Battmann & S. Dutke (Eds.), *Processes of the molar regulation of behavior* (pp. 105-118). Lengerich: Pabst.

Strauß, B. (1993). *Konfundierungen beim Komplexen Problemlösen*. Bonn: Holos.

Strauß, B. (1995). Zur Operationalisierung der Komplexität in dynamischen Systemen. *Zeitschrift für Psychologie, 203*, 73-99.

Strauß, B. & Kleinmann, M. (Hrsgb.). (1995a). *Computersimulierte Szenarien in der Personalarbeit*. Göttingen: Verlag für Angewandte Psychologie.

Strauß, B. & Kleinmann, M. (1995b). Die formale Beschreibung computersimulierter Szenarien. In B. Strauß & M. Kleinmann (Hrsg.), *Computersimulierte Szenarien in der Personalarbeit* (S. 105-125). Göttingen: Verlag für Angewandte Psychologie.

Strauß, B. & Kleinmann, M. (1996). Computersimulierte Szenarien im Assessment Center. In W. Sarges (Hrsg.), *Weiterentwicklungen der Assessment Center Methode* (S. 69-86). Göttingen: Verlag für Angewandte Psychologie.

Strauß, B. & Kleinmann, M. (1997). Computersimulierte Szenarien in der Personalarbeit. In H. Mandl (Hrsg.), *Bericht über den 40. Kongreß der Deutschen Gesellschaft für Psychologie in München 1996* (S. 457-462). Göttingen: Hogrefe.

Strauß, B. & Kleinmann, M. (1998). Validity and implementation of computer-simulated scenarios in personnel assessment. *International Journal of Selection and Assessment, 6*, 97-105.

Streufert, S., Breuer, K. & Michalik, B. (1995). *Strategische Management Simulationen*. Göttingen: Hogrefe Apparatezentrum.

Streufert, S., Pogash, R. & Piasecki, M. (1988). Simulated-based assessment of managerial competence: Reliability and validity. *Personnel Psychology, 41*, 537-557.

Streufert, S. & Swezey, R.W. (1986). *Complexity, managers, and organizations*. Orlando, FL: Academic Press.

Strohschneider, St. (1986). Zur Stabilität und Validität von Handeln in komplexen Realitätsbereichen. *Sprache & Kognition, 5*, 42-48.

Strohschneider, St. (1990). *Wissenserwerb und Handlungsregulation*. Wiesbaden: Deutscher Universitäts-Verlag.

Strohschneider, St. (1991a). Problemlösen und Intelligenz: Über die Effekte der Konkretisierung komplexer Probleme. *Diagnostica, 37*, 353-371.

Strohschneider, St. (1991b). Kein System von Systemen! Kommentar zu dem Aufsatz 'Systemmerkmale als Determinanten des Umgangs mit dynamischen Systemen' von Joachim Funke. *Sprache & Kognition, 10*, 109-113.

Strohschneider, St. (1994). Strategien beim Umgang mit einem komplexen Problem: Ein deutsch - deutscher Vergleich. *Zeitschrift für Arbeits- und Organisationspsychologie, 38*, 34-40.

Strohschneider, St. (Hrsgb.) (1996a). *Denken in Deutschland*. Bern: Huber.

Strohschneider, St. (1996b). Problemlösen und Kultur. In St. Strohschneider (Hrsg.), *Denken in Deutschland* (S. 17-47). Bern: Huber.

Strohschneider, St. (1996c). Strategien beim Umgang mit einem zieloffenen komplexen Problem. In St. Strohschneider (Hrsg.), *Denken in Deutschland* (S.71-95). Bern: Huber.

Strohschneider, St. & Schaub, H. (1991). Können Manager wirklich so gut managen? Über die Effekte unterschiedlichen heuristischen Wissens beim Umgang mit komplexen Problemen. *Zeitschrift für Psychologie, Suppl., 11*, 325-339.

Strohschneider, St. & Schaub, H. (1995). Problemlösen. In Th. Geilhardt & Th. Mühlbradt (Hrsg.), *Planspiele im Personal- und Organisationsmanagement* (S. 187-203). Göttingen: Verlag für Angewandte Psychologie.

Strohschneider, St. & von der Weth, R. (Hrsg.). (1993). *Ja, mach nur einen Plan. Planen und Fehlschläge-Ursachen, Beispiele, Lösungen*. Bern: Huber.

Süß, H.M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttingen: Hogrefe.

Süß, H.M., Beauducel, A., Kersting, M. & Oberauer, K. (1992). Wissen und Problemlösen. In L. Montada (Hrsg.), *Bericht über den 38. Kongreß der Deutschen Gesellschaft für Psychologie in Trier 1992* (Bd. 1, S. 347 f.). Göttingen: Hogrefe.

Süß, H.M., Kersting, M. & Oberauer, K. (1991). Intelligenz und Wissen als Prädiktoren für Leistungen bei computersimulierten komplexen Problemen. *Diagnostica, 37*, 334-352.

Süß, H.M., Kersting, M. & Oberauer, K. (1993a). Zur Vorhersage von Steuerungsleistungen an computersimulierten Systemen durch Wissen und Intelligenz. *Zeitschrift für Differentielle und Diagnostische Psychologie, 14*, 189-203.

Süß, H.M., Oberauer, K. & Kersting, M. (1993b). Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme. *Sprache & Kognition, 12*, 83-97.

Tent, L. (1992). Über den Ursprung des neuerlichen Unbehagens an der experimentellen Psychologie oder: die Skrupel des Prometheus? In H. Gundlach (Hrsg.), *Psychologische Forschung und Methode: Das Versprechen des Experiments* (S. 205-226). Passau: Passavia Universitätsverlag.

Tergan, S.O. (1988). Qualitative Wissensdiagnose. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie*. Weinheim: Psychologische Verlags Union.

Tergan, S.O. (1989a). Psychologische Grundlagen der Erfassung individueller Wissensrepräsentationen Teil I: Grundlagen der Wissensmodellierung. *Sprache und Kognition, 8*, 152-165.

Tergan, S.O. (1989b). Psychologische Grundlagen der Erfassung individueller Wissensrepräsentationen Teil II: Methodologische Aspekte. *Sprache und Kognition, 8*, 193-202.

Testkuratorium der Föderation deutscher Psychologenverbände. (1986). Mitteilung: Beschreibung der einzelnen Kriterien für die Testbeurteilung. *Diagnostica, 32*, 358-360.

Tewes, U. (Hrsgb.). (1991). *Hamburg-Wechsler Intelligenztest für Erwachsene. Revision 1991. Handbuch und Testanweisung*. Bern: Huber.

Tiedemann, J. (1995). Kognitive Stile. In M. Amelang (Hrsg.), *Enzyklopädie der Psychologie* (Bd. 2, Serie 8, S. 508-533). Göttingen: Hogrefe.

Todt, E. (1992). Interesse männlich - Interesse weiblich. In Jugendwerk der Deutschen Shell (Hrsg.), *Jugend '92* (Bd. 2, Im Spiegel der Wissenschaften, S. 301-317). Opladen: Leske & Budrich.

Trost, G. (1993). Attitudes and reactions of West German students with respect to scholastic aptitude test in selection and counseling programs. In B. Nevo & R.S. Jäger (Hrsg.), *Educational and psychological testing: The test taker's outlook* (pp. 177-201). Göttingen: Hogrefe.

Vernon, P.E. (1979). *Intelligence: Heredity and environment*. San Francisco: Freeman and Company.

Wagner, R.K. & Sternberg, R.J. (1985). Practical intelligence in real world pursuits: the role of tacit knowledge. *Journal of Personality and Social Psychology, 49*, 436-458.

Weber, W. & Werner, R. (1983). Auswahlverfahren und Prüfungsleistungen bei Anwärtern des mittleren und des gehobenen nicht-technischen Dienstes. *DGP-Informationen*. Hannover: Deutsche Gesellschaft für Personalwesen , *43*, 1-40.

Wertheimer, M. (1945). *Produktives Denken*. Frankfurt a.M.: W. Kramer.

Westmeyer, H. (1976). Grundlagenprobleme psychologischer Diagnostik. In K. Pawlik (Hrsg.), *Diagnose der Diagnostik* (S. 71-101). Stuttgart: Klett.

Westmeyer, H. (1993). Psychologie als Grundlagenwissenschaft und als angewandte Diziplin: Eine strukturalistische Analyse der technologischen Sichtweise. In W. Bungard & T. Herrmann (Hrsg.), *Arbeits- und Organisationspsychologie im Spannungsfeld zwischen Grundlagenorientierung und Anwendung* (S. 49-63). Bern: Huber.

Wigdor, A.K. & Sackett, P.R. (1993). Employment testing and public policy: The case of the general aptitude test battery. In H. Schuler, J.L. Farr & M. Smith (Eds.), *Personnel selection and assessment. Individual and organizational perspectives* (pp. 183-204). Hillsdale, NJ: Erlbaum.

Wiggins, J.S. (1973). *Personality and prediction: Principles of personality assessment*. Reading, MA: Addison-Wesley.

Wittmann, W.W. (1988). Multivariate reliability theory: Principles of symmetry and successful validation strategies. In J.R. Nesselroade & R.B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (Vol. 2nd ed., S. 505-560). New York: Plenum.

Wittmann, W.W. & Matt, G.E. (1986). Aggregation und Symmetrie. Grundlagen einer multivariaten Reliabilitäts- und Validitätstheorie, dargestellt am Beispiel der differentiellen Validität des Berliner Intelligenzstrukturmodells. *Diagnostica, 32*, 309-329.

Wittmann, W.W., Süß, H.M. & Oberauer, K. (1996). *Determinanten komplexen Problemlösens* (Berichte des Lehrstuhls Psychologie II der Universität Mannheim). Mannheim: Universität Mannheim, *Heft 9*.

Wolf, B. (1990). *Zusammenhänge zwischen Ergebnissen von eignungsdiagnostischen Verfahren und Ausbildungserfolg - dargestellt am Beispiel der Auswahl von Bewerbern für den gehobenen Polizeivollzugsdienst* (unveröffentlichte Diplomarbeit). Göttingen: Georg-August Universität zu Göttingen.

Wolfe, J. & Roberts, C.R. (1986). The external validity of a business management game. A five-year longitudinal study. *Simulation & Games, 17*, 45-59.

Wottawa, H. (1987a). Hypotheses agglutination (HYPAG): A method for configuration-based analyses of multivariate data. *Methodika, 1*, 68-92.

Wottawa, H. (1987b). Konfigurale Auswertungsmethoden in der Psychotherapieforschung. *Zeitschrift für Klinische Psychologie, Psychopathologie und Psychotherapie, 35*, 100-113.

Wottawa, H. & Amelang, M. (1980). Einige Probleme der „Testfairness" und ihre Implikationen für Hochschulzulassungsverfahren. *Diagnostica, 26*, 199-221.

20. Anhang

Forschungsprojekt **Problemlösen**

Rückantwort

Herrn
Dipl.-Psych.
M. Kersting
Str. 8
1 Berlin

Es geht um die Einschätzung von

Frau/Herrn_____

Alle Angaben zum Forschungsprojekt werden nur in anonymisierter Form ausgewertet. Der Name der hier einzustufenden Person wird -ebenso wie Ihr Name- nicht registriert, bzw. gelöscht, sobald die Zuordnung zu den bisherigen Daten erfolgt ist. Lediglich allgemeine demographische Aussagen (z.B. "92 % der Stichprobe waren männlich") werden getroffen. Für diese Zwecke wird eine Angabe über Ihr Geschlecht, über Ihr Alter und Ihren Dienstgrad benötigt. Bitte kreuzen Sie das auf Sie zutreffende Kästchen an und tragen Sie Ihr Alter und Ihren Dienstgrad (Abkürzung reicht) ein.

Ich bin [weiblich] [männlich] *und* _____ *Jahre alt. Vom Dienstgrad her bin ich* _____.

Bitte lesen Sie im folgenden Fragebogen anstelle der Pünktchen "..." immer *"Frau/Herr .....".*

Wie lange kennen Sie *"Frau/Herrn ..."* beruflich? - Antwort: *Ich kenne "Frau/Herrn ..." seit* _____ *Jahr(en).*
Unabhängig davon, wie lange Sie *"Frau/Herrn ..."* schon kennen: Die folgenden Fragen sollen Sie bitte auf das berufliche Verhalten beziehen, das *"Frau/Herr ..."* in den ***letzten*** Wochen und Monaten gezeigt hat. Sollte Ihr letzter beruflicher Kontakt mit *"Frau/Herrn ..."* länger als ein halbes Jahr zurückliegen, so geben Sie bitte alle Unterlagen kurzfristig an die/den aktuelle(n) Vorgesetzte(n) weiter. Bitte tragen Sie hier das aktuelle Datum ein: *Heute ist der* _____.

In welchem dienstlichen Verhältnis stehen Sie zu *"Frau/Herrn ..."*? Kreuzen Sie bitte das zutreffende Kästchen an:
a) [Ich bin ihr(e)/sein(e) Vorgesetzte(r)] b) [Sie/er ist mein(e) Vorgesetzte(r)] c) [Keinerlei Vorgesetztenverhältnis]

Bitte benennen Sie mit einem Oberbegriff (z.B. "Fachlehrer(in)", "Sachbearbeiter(in)" usw.) die Tätigkeit, die *"Frau/Herr..."* in den letzten Monaten überwiegend ausführte.
Antwort: *"Frau/Herr ..." war in den letzten Monaten überwiegend als* _____ *tätig.*

Geben Sie bitte im folgenden Kästchen die Zahl der Mitarbeiter(innen) an, gegenüber denen *"Frau/Herr ..."* in den letzten Monaten weisungsbefugt war, bzw. gegenüber denen sie/er Vorgesetztenfunktionen wahrgenommen hat.
a) regulär: [ ] b) in Ausnahmefällen (z.B. vertretungshalber) [ ]

Nun zum Fragebogen: Zunächst werden Sie gebeten einzuschätzen, wie bedeutend bestimmte Fähigkeiten Ihrer Ansicht nach für die erfolgreiche Bewährung im Arbeitsgebiet von *"Frau/Herrn ..."* überhaupt sind. Die Antwortkategorien gehen von *"unwichtig"* über *"eher unwichtig"* und *"eher wichtig"* bis zu *"wichtig"*. Es geht um die Fähigkeiten "Intelligenz", "Problemlösefähigkeit" und "Kooperationsfähigkeit". Auf dem Instruktionsblatt ist erläutert, was mit den einzelnen Begriffen genau gemeint ist. Bitte lesen Sie diese Beschreibung der Fähigkeiten auf dem Instruktionsblatt. Beziehen Sie Ihre Antworten auf die Fähigkeiten, so wie sie im Instruktionsblatt dargestellt sind und auf das Arbeitsgebiet von *"Frau/Herrn ..."* und ergänzen Sie dann durch Ankreuzen folgende Aussage:

Für die erfolgreiche Bewährung im Arbeitsgebiet von *"Frau/Herrn ..."* ist die ...

	a) unwichtig	b) eher unwichtig	c) eher wichtig	d) wichtig
1.1 Intelligenz	[a]	[b]	[c]	[d]
1.2 Problemlösefähigkeit	[a]	[b]	[c]	[d]
1.3 Kooperationsfähigkeit	[a]	[b]	[c]	[d]

Forschungsprojekt **Problemlösen**

Seite 1 ---Fortsetzung: Rückseite---

Abb. 15: Fragebogen zur Vorgesetztenbeurteilung

Lesen Sie bitte noch einmal die Beschreibung der Fähigkeiten auf dem Instruktionsblatt. Vergleichen Sie bitte die Ausprägung jeder Fähigkeit bei *"Frau/Herrn ..."* mit der Ausprägung, die die jeweilige Fähigkeit bei anderen Kolleg(inn)en des gehobenen Dienstes hat. Kreuzen Sie dann an, ob *"Frau/Herr ..."* hinsichtlich der beschriebenen Fähigkeiten eher zu den oberen 25%, zum besseren oder schlechteren "Mittelfeld" (jeweils 25%) oder zu den unteren 25% der Vergleichsgruppe gehört.

	a) untere 25%	b) unteres Mittelfeld	c) oberes Mittelfeld	d) obere 25%
2.1 Intelligenz	a	b	c	d
2.2 Problemlösefähigkeit	a	b	c	d
2.3 Kooperationsfähigkeit	a	b	c	d

Um die vorherige Frage beantworten zu können, mußten Sie *"Frau/Herrn ..."* in eine Rangreihe zu anderen Personen bringen. Jeder Mensch hat aber auch *individuelle* Stärken und Schwächen. Bei den folgenden Fragen geht es darum, nicht Personen, sondern die genannten -und auf dem Instruktionsblatt beschriebenen- Fähigkeiten in eine individuelle Rangreihe für *"Frau/Herrn ..."* zu bringen. Es geht also nicht um einen Vergleich mit anderen Personen, sondern nur um die individuellen Stärken und Schwächen von *"Frau/Herrn ..."*. Die Fähigkeiten sind paarweise geordnet. Dabei dürfen Sie bei jedem gegebenen Paar **nur eine** der beiden Fähigkeiten wählen. Ein solcher Wahlzwang verärgert manchmal, weil man keine Gelegenheit erhält, etwas über die Höhe der Zustimmung oder den Abstand zwischen den Alternativen zu äußern. Lassen Sie sich bitte dennoch für drei Fragen auf dieses Verfahren ein und kreuzen Sie nur ein Kästchen pro Paar an. Kreuzen Sie für jede Paar an, welche der beiden genannten Fähigkeiten bei *"Frau/Herrn ..."* höher ausgeprägt ist:

3.1 Im Vergleich zwischen der Intelligenz und der Kooperationsfähigkeit von *"Frau/Herrn ..."* ist die [Intelligenz] [Kooperationsfähigkeit] höher ausgeprägt.

3.2 Im Vergleich zwischen der Problemlösefähigkeit und der Intelligenz von *"Frau/Herrn ..."* ist die [Problemlösefähigkeit] [Intelligenz] höher ausgeprägt.

3.3 Im Vergleich zwischen der Kooperationsfähigkeit und der Problemlösefähigkeit von *"Frau/Herrn ..."* ist die [Kooperationsfähigkeit] [Problemlösefähigkeit] höher ausgeprägt.

Nachdem in den bisherigen Fragen die Fähigkeiten ganz allgemein beurteilt werden sollten, geht es in den folgenden Fragen um konkrete Verhaltensweisen oder Leistungen, die zu den Oberbegriffen "Intelligenz", "Problemlösefähigkeit" und "Kontaktfähigkeit" gehören.

Die Befragung gliedert sich dabei stets in zwei Teile. Zunächst geht es darum, wie bedeutend die genannte Leistung Ihrer Ansicht nach für die erfolgreiche Bewährung im Arbeitsgebiet von *"Frau/Herrn ..."* überhaupt ist. Bitte kreuzen Sie für jede der folgenden Beschreibungen auf der linken Seite an, welche Bedeutung Ihrer Ansicht nach der beschriebenen Leistung für eine erfolgreiche Bewährung im Arbeitsgebiet von *"Frau/Herrn ..."* zukommt. Die Antwortkategorien gehen von *"geringe"* (Bedeutung) über *"eher geringe"*, und *"eher hohe"* bis zu *"hohe"* (Bedeutung).

Im zweiten Teil der Frage geht es dann darum, wie die beschriebene Leistung bei *"Frau/Herrn ..."* in den letzten Monaten ausgeprägt war. Orientieren Sie sich dabei an den jeweils aufgeführten Verhaltensbeschreibungen, die Beispiele für die Leistung in der gefragten Dimension darstellen. Vergleichen Sie zur Beantwortung dieses Teils der Frage *"Frau/Herrn..."* wieder mit anderen Kolleg(inn)en des gehobenen Dienstes, und überlegen Sie welchen Rangplatz *"Frau/Herr ..."* hinsichtlich ihrer/seiner Leistungen in den letzten Monaten Ihrer Ansicht nach eingenommen hat. Auf die Frage nach der Ausprägung der Leistung gibt es jeweils sechs Antwortkategorien, nämlich

"+++" =sehr hohe Ausprägung, gehört hinsichtlich d. beschriebenen Leistung z. weit überdurchschnittl. Spitzengruppe
"++" =hohe Ausprägung, gehört hinsichtlich der beschriebenen Leistung zur leistungsstarken Gruppe
"+" =eher hohe Ausprägung, gehört gehört hinsichtlich der beschriebenen Leistung zum besseren Durchschnitt
"-" =eher schwache Ausprägung, gehört gehört hinsichtlich d. beschriebenen Leistung zum schlechteren Durchschnitt
"--" =schwache Ausprägung, gehört hinsichtlich der beschriebenen Leistung zur leistungsschwachen Gruppe
"---" =sehr schwache Ausprägung, gehört hinsichtlich d. beschriebenen Leistung z. weit unterdurchschnittlichen Gruppe

Sie sollen bei jeder Frage die Bedeutung **und** die Ausprägung der Leistung einschätzen.

Fortsetzung Abb. 15: Fragebogen zur Vorgesetztenbeurteilung

Einschätzung für Frau/Herrn .....

4.1 ***Einfallsreichtum***

Die Leistung, viele und verschiedene brauchbare Ideen bei der Lösung von Problemen zu produzieren

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.2 ***Analyse von Informationen und Daten***

Die Leistung, Informationen in ihre *Komponenten* zu zerlegen, wobei *zugrundeliegende* Prinzipien oder Tatsachen identifiziert werden

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.3 ***Teamarbeit***

Die Leistung, die kooperative und sachbezogene Zusammenarbeit unter den Kolleg(inn)en aktiv zu fördern, wobei die jeweiligen Stärken der einzelnen zielgerichtet zu einem leistungsstarken Team ergänzt werden

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.4 ***Selektive Informationsaufnahme***

Die Leistung, bei der Bearbeitung von Informationen Wesentliches von Unwesentlichem zu unterscheiden

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.5 ***Strategisches Denken***

Die Leistung, bei Problembearbeitungen strategisch, systematisch und umsichtig vorzugehen, wobei einzelne Maßnahmen und ihre Abfolge bewußt geplant werden und dabei nicht nur die Haupt-, sondern auch die Nebenwirkungen der Maßnahmen Berücksichtigung finden

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.6 ***Kombination von Informationen***

Die Leistung, Informationen aus verschiedenen Quellen logisch und folgerichtig zusammenzustellen und zu kombinieren, wobei Sachverhalte erkannt und aufeinanderbezogene Informationen zu einem überschaubaren und sachgerechten Konzept zusammengefaßt werden

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.7 ***Umgang mit Mehrdeutigkeit und Auslegungsspielraum***

Die Leistung, auch unter erschwerten Bedingungen sachdienliche Entscheidungen zu treffen, nämlich auch dann, wenn die Vorgaben uneindeutig sind (oder mehrere Vorgaben sich widersprechen) und wenn bei der Beurteilung von Tatbeständen, Sachverhalten und Arbeitsergebnissen ein beträchtlicher Auslegungs-, Beurteilungs- oder Ermessensspielraum besteht

hat im Aufgabengebiet von *"Fraw/Herrn ..."* eine	ist bei *"Fraw/Herrn ..."* folgendermaßen ausgeprägt:
a) geringe b) eher geringe c) eher hohe d) hohe	
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

Forschungsprojekt ProblemlöseN

Seite 3 ---Fortsetzung: Rückseite---

Fortsetzung Abb. 15: Fragebogen zur Vorgesetztenbeurteilung

4.8 ***Kontaktaufnahme***

Die Leistung, mit anderen (auch fremden) Kolleg(inn)en Kontakt aufzunehmen, sich auf ihre Bedürfnisse und Gefühle einzulassen, sich in neue Arbeitsgruppen zu integrieren

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.9 ***Stresstolerantes Selbstmanagement***

Die Leistung, auch dann zielgerichtet zu handeln, wenn Zeitdruck und frustrierende Zwischenergebnisse die Arbeit beeinträchtigen

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.10 ***Effizienz bei der Erledigung von Routinetätigkeiten***

Die Leistung, einfache Routinetätigkeiten mit hohem Arbeitstempo und gleichzeitig mit hoher Sorgfalt zu erledigen

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.11 ***Deduktives Denken***

Die Leistung, das Besondere, den Einzelfall logisch aus dem Allgemeinen abzuleiten

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.12 ***Sensibilität und Flexibilität gegenüber Veränderungen***

Die Leistung, das eigene Vorgehen immer wieder an der Wirklichkeit zu überprüfen und die eigenen Entscheidungen immer wieder flexibel an die Rückmeldung anzupassen

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.13 ***Konfliktvermittlung***

Die Leistung, bei Konflikten unter Kolleg(inn)en zu vermitteln, zu schlichten und einen Konsens herbeizuführen

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.14 ***Überblick***

Die Leistung, sich in unübersichtlichen Situationen und Sachbeständen durch die Suche nach und die Aufnahme von Informationen einen Überblick über die Lage zu verschaffen, ohne sich dabei in Detailanalysen zu verlieren

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

4.15 ***Gedächtnis***

Die Leistung, sich sprachliche, numerische oder figurale Informationen aktiv einzuprägen und kurzfristig zu reproduzieren oder wiederzuerkennen

hat im Aufgabengebiet von *"Frau/Herrn ..."* eine a) geringe b) eher geringe c) eher hohe d) hohe	ist bei *"Frau/Herrn ..."* folgendermaßen ausgeprägt:
[a] [b] **Bedeutung** [c] [d]	[---] [--] [-] [+] [++] [+++]

Forschungsprojekt ProblemlöseN

Seite 4---Ende!--- Vielen Dank für Ihre Mitarbeit!

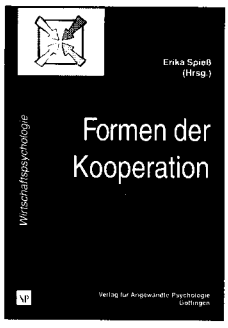
Fortsetzung Abb. 15: Fragebogen zur Vorgesetztenbeurteilung

Wirtschaftspsychologie

Erika Spieß (Hrsg.)

Formen der Kooperation

Bedingungen und Perspektiven
(Wirtschaftspsychologie)
1998, 316 Seiten, geb., DM 69, / sFr. 60,- öS 504, • ISBN 3-8017-1018-1

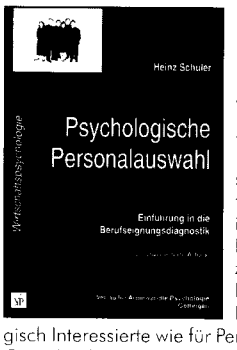


Bei der bislang vorherrschenden strategischen Form der Kooperation in Wirtschaftsorganisationen wird zunehmend der zusätzliche Bedarf nach einer empathischen Kooperation deutlich. Besondere Bedeutung erlangt diese Form der Kooperation in internationalen Kontexten und mit Angehörigen fremder Kulturen. Neben verschiedenen Formen der Kooperation beschäftigt sich dieses Buch mit den Bedingungen für Kooperation. So wird z.B. die Rolle von Organisationsformen und kulturellen Einflüssen diskutiert.

Heinz Schuler

Psychologische Personalauswahl

Einführung in die Berufseignungsdiagnostik
(Wirtschaftspsychologie)
2., unveränd. Aufl. 1998, 246 Seiten, geb., DM 59,- sFr. 56,- • öS 431,- • ISBN 3-8017-0865-9

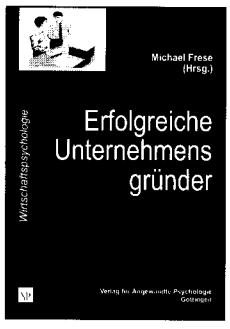


Wie hängen menschliche Merkmale mit beruflichem Erfolg zusammen, und wie kann man beide messen? Welches sind die wichtigsten Methoden der Personalauswahl, wo können sie eingesetzt werden, wie funktionieren sie, wie weit ist Verlaß auf sie, ist ihr Einsatz verantwortbar, akzeptabel und Rechtens? Die Antworten auf diese Fragen sind für psychologisch Interessierte wie für Personalverantwortliche von Gewicht, denn es werden diejenigen Verfahrensweisen und ihre Grundlagen dargestellt, die dem aktuellen wissenschaftlichen Stand entsprechen und gleichzeitig von Nutzen sind, Personalentscheidungen in der Praxis zu verbessern.

Michael Frese (Hrsg.)

Erfolgreiche Unternehmensgründer

Psychologische Analysen und praktische Anleitungen für Unternehmer in Ost- und Westdeutschland
(Wirtschaftspsychologie)
1998, VIII/230 Seiten, geb., DM 69,- / sFr. 60,- öS 504,- • ISBN 3-8017-1113-7



Das Buch zeigt auf, wie bedeutend psychologische Faktoren für eine erfolgreiche Unternehmensführung sind, unter welchen Gegebenheiten diese Faktoren besonders relevant sind und welche Konsequenzen sich daraus für die Praxis ergeben. Dazu wird u.a. die Bedeutung von Gründungsmotiven und Unterschieden zwischen Ost- und Westdeutschland für den unternehmerischen Erfolg behandelt sowie dem Zusammenhang zwischen Persönlichkeit und erfolgreichen Handlungsstrategien nachgegangen.

Lutz von Rosenstiel / Friedemann W. Nerdinger
Erika Spieß (Hrsg.)

Von der Hochschule in den Beruf

Wechsel der Welten in Ost und West
(Wirtschaftspsychologie)
1998, 222 Seiten, geb., DM 59,- / sFr. 51, öS 431,- • ISBN 3-8017-1060-2

Das Buch zeigt auf, was Hochschulabsolventen unternehmen, um eine adäquate Arbeitsstelle zu finden, wie sich die Integration in Organisationen vollzieht und welche Rolle hierbei ihre beruflichen Ansprüche und Ziele spielen. Es werden Aspekte der Arbeitslosigkeit, der beruflichen Selbständigkeit, der Einarbeitung neuer Mitarbeiter und der Weiterbildungsbereitschaft behandelt. In differenzierten Analysen wird auf erste Berufsverläufe von Frauen und Männern sowie Hochschulabsolventen aus den alten und neuen Bundesländern eingegangen.

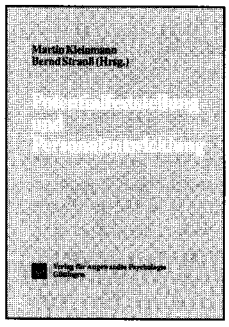
Verlag für Angewandte Psychologie
Rohnsweg 25, 37085 Göttingen • Tel. 0551/49609-0 • http://www.hogrefe.de

Personalentwicklung

Martin Kleinmann / Bernd Strauß (Hrsg.)

Potentialfeststellung und Personalentwicklung

(Psychologie für das Personalmanagement)
1998, 248 Seiten, DM 69,– / sFr. 60,– / öS 504,
ISBN 3-8017-0905-1

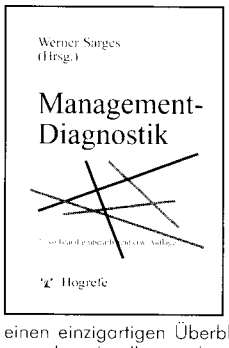


Das Buch vermittelt aktuelle wissenschaftliche Erkenntnisse und praktische Erfahrungen mit Potentialanalyse- und Personalentwicklungsinstrumenten. Im einzelnen wird u.a. dargestellt, welche Potentialfeststellungsverfahren und Personalentwicklungsinstrumente im modernen Personalmanagement eingesetzt werden, auf welchen theoretischen Grundlagen der Einsatz der Verfahren basiert und welche praktischen Erfahrungen mit diesen Instrumenten vorliegen.

Werner Sarges (Hrsg.)

Management-Diagnostik

2., vollständig überarb. und erw. Aufl. 1995,
XII/980 S., geb. DM 198,–/sFr. 196,–/öS 1.445,
ISBN 3-8017-0740-7



Die Rekrutierung und Entwicklung von Führungs- und Nachwuchsführungskräften wird in den nächsten Jahren immer dringlicher werden. Eine verbesserte Eignungsdiagnostik zur Potentialfeststellung, Auswahl und Plazierung von Führungskräften kann das Problem erheblich mildern. Die vorliegende Überarbeitung des erfolgreichen Handbuches liefert einen einzigartigen Überblick zur Management-Diagnostik und stellt ein sehr umfangreiches, so nirgends gebündeltes und hochaktuelles Expertenwissen dar.

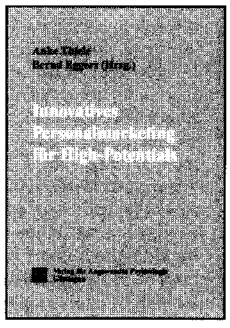
Hogrefe - Verlag
Rohnsweg 25, 37085 Göttingen • http://www.hogrefe.de

Buchtips

Anke Thiele / Bernd Eggers (Hrsg.)

Innovatives Personalmarketing für High-Potentials

(Psychologie für das Personalmanagement)
1999, 211 Seiten, DM 59,– / sFr. 51,–
öS 431,– • ISBN 3-8017-1103-X

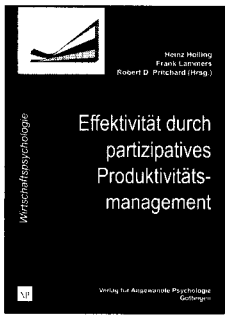


Strategien und Instrumente, Chancen und Risiken sowie Entwicklungstendenzen in der Rekrutierung und Förderung von High-Potentials stehen im Mittelpunkt dieses Bandes. Anhand von Praxisbeispielen renommierter Unternehmen werden effiziente und erfolgversprechende Wege bei der Rekrutierung und Förderung von Top-Nachwuchskräften aufgezeigt. Professionalität, richtiges Timing, gezielte Zielgruppenansprache sowie integre Kommunikation sind nur einige der im Buch dargestellten Schlüsselfaktoren eines erfolgreichen Personalmarketings.

Heinz Holling / Frank Lammers
Robert D. Pritchard (Hrsg.)

Effektivität durch Partizipatives Produktivitätsmanagement

Überblick, neue theoretische Entwicklungen und europäische Fallbeispiele (Wirtschaftspsychologie)
1999, X/186 S., DM 69,– / sFr. 60,–
öS 504,– • ISBN 3-8017-0842-X



Maßnahmen zur Entwicklung von Organisationen sind in der heutigen Zeit ein wichtiger Schlüssel zum Erfolg, da eine zunehmende Flexibilität und Veränderungsbereitschaft von Organisationen gefordert wird. Das Buch beschreibt nicht nur die Vorteile und die Effektivität, sondern auch die Probleme und Schwierigkeiten des »Partizipativen Produktivitätsmanagements« – eines neuen, leistungsfähigen Instruments zur Organisationsentwicklung.

Verlag für Angewandte Psychologie
Rohnsweg 25 • 37085 Göttingen • http://www.hogrefe.de

Zahlreiche Personalverantwortliche äußern den Wunsch nach neuen bzw. verbesserten diagnostischen Verfahren, insbesondere zur Erfassung der Berufseignung bei Führungskräften. Gerade hinsichtlich dieser Anwendungsfelder wurde für computergestützte Problemlöseszenarien ein konzeptioneller diagnostischer Fortschritt postuliert. Die Schwächen der als »veraltet« und »realistätsfremd« empfundenen Intelligenztests sollten mit den neuen Instrumenten überwunden werden.

Dieses Buch behandelt systematisch die relevanten Aspekte, die bei einer möglichen professionellen Diagnostik mit Hilfe computergestützter Problemlöseszenarien beachtet werden müssen und geht dabei insbesondere auf die Personalauswahl und die sogenannte »Managementdiagnostik« ein. Die für Problemlöseszenarien proklamierten diagnostischen Vorzüge (Simulationsargument, Akzeptanz, Erweiterung der herkömmlichen Intelligenzdiagnostik) werden kritisch diskutiert.

Im zweiten Teil der Arbeit werden die Ergebnisse einer Studie dargestellt, in der erstmals die Vorhersage beruflicher Leistungen durch Problemlöseszenarien einerseits und Intelligenztests andererseits miteinander verglichen wurde.

ISBN 3-8017-1259-1