

# Zur Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien anhand der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung

Martin Kersting

**Zusammenfassung.** Computergestützte Problemlöseszenarien haben sich in der deutschsprachigen Psychologie etabliert und werden auch in der Eignungsdiagnostik genutzt. Der Artikel fokussiert die bislang unzureichend ausgelotete Validität von Problemlöseszenarien. Dabei steht die Frage im Mittelpunkt, ob mit den Szenarien ein neues Konstrukt „Problemlösefähigkeit“ gemessen wird oder ob die gemeinsame Varianz zwischen den Steuerungsleistungen und den Berufserfolgskriterien durch Intelligenz und Wissen erschlossen werden kann, so daß sich die Annahme einer eigenständigen „Problemlösefähigkeit“ erübrigt. Im empirischen Teil werden die Ergebnisse einer Studie vorgestellt, in der u. a. Intelligenz- und Wissenstests sowie zwei computergestützte Problemlöseszenarien eingesetzt wurden. Zunächst wurde die Konstruktvalidität der Steuerungsleistungen anhand der Daten von 104 Polizisten<sup>1</sup> analysiert. Dabei zeigte sich, daß sich der Problemlöseerfolg substantiell auf Intelligenz und Wissen zurückführen läßt. In einem zweiten Schritt wurde die Vorhersagbarkeit der über Vorgesetztenurteile operationalisierten beruflichen Bewährung von 73 Polizisten durch intellektuelle Fähigkeiten, Wissen und Steuerungsleistungen vergleichend untersucht. Die Berücksichtigung der Problemlösefähigkeit konnte keinen – gegenüber der Intelligenz und dem Wissen – signifikant inkrementellen Beitrag zur Vorhersage der Vorgesetztenurteile leisten. Abschließend wird diskutiert, in welcher Art und Weise computergestützte Problemlöseszenarien diagnostisch genutzt werden können.

Schlüsselwörter: Intelligenz, Wissen, Problemlösen, Validität, Eignungsdiagnostik

On the construct and criterion validity of problem-solving scenarios based on the prediction of supervisor assessment of job performance

**Abstract.** Computer-based problem-solving scenarios have been well established in German-language psychology and are also relied upon in personnel selection. The present article focuses on the so far largely neglected question of how much validity such scenarios might claim. Problem solving scenarios cannot make a well-founded claim for content validity. However, in some studies systematic relations have been demonstrated between problem-solving achievement and criteria of job performance. This article deals with the question whether these scenarios measure a new construct, “problem-solving capacity” or whether the criterion-relevant part of the measurement taps variance that can also be accessed by traditional intelligence and knowledge tests.

In the empirical section the findings of a study are presented where intelligence and knowledge tests as well as two computer-based problem-solving scenarios were used. At first, construct validity of control performance was analyzed based on data from 104 policemen. It turned out that a substantial part of problem-solving achievement could be attributed to intelligence and knowledge. In a second step, the predictability of supervisor assessment of job performance by intellectual capacity, knowledge and problem-solving capacity was compared for a subgroup of 73 persons from the sample. Taking problem-solving capacity into account did not make a significant incremental contribution to the prediction of job performance by intelligence and knowledge.

The final discussion deals with the ways computer-based problem-solving scenarios might contribute to diagnostic purposes.

Key words: intelligence, knowledge, problem solving, validity, personnel selection

Die aus der kognitionspsychologischen Forschung stammenden komplexen computergestützten Problemlöseszenarien werden seit einiger Zeit in der Praxis als diagnosti-

sche Instrumente genutzt und insbesondere für die Eignungs- und Managementdiagnostik empfohlen sowie kommerziell angeboten (siehe Funke, 1995). Entscheidend für den diagnostischen Einsatz computergestützter Problemlöseszenarien ist die Validität. Implikationen der Problemlöseforschung für die diagnostische Praxis wür-

<sup>1</sup> Hier wie im folgenden ist die weibliche Form stets mit gemeint.

den sich vor allem aus dem Nachweis ergeben, daß mit den Szenarien entweder ein vertrautes Konstrukt besser als bisher gemessen wird oder aber daß ein relevantes neues Konstrukt erstmals zuverlässig erfaßt wird. Im folgenden wird die Konstrukt- und Kriteriumsvalidität von Problemlöseszenarien erörtert und anschließend empirisch geprüft.

## Zur Validität von Problemlöseszenarien

Ogleich der Begriff Kontentvalidität nicht explizit verwendet wird, argumentieren Befürworter einer Diagnostik mit Problemlöseszenarien häufig im Sinne der Kontentvalidität, wenn sie den Simulationscharakter der Szenarien betonen und damit eine „job-relatedness“ der Instrumente suggerieren. Kontentvalidität setzt voraus, daß zwei Mengen (z. B. das Original einer Kleinstadt und das am Computer erstellte Modell dieser Kleinstadt oder die Anforderungen in einem bestimmten Beruf und die Anforderungen bei der Szenariensteuerung) in einem bestimmten, repräsentativen Verhältnis zueinander stehen. Voraussetzung hierfür ist, daß die beiden interessierenden Mengen hinreichend definiert sind. Bei den meisten computergestützten Problemlöseszenarien existiert weder ein explizites Modell des vorgeblich simulierten Realitätsbereiches, noch eine Analyse der Aufgabenmerkmale des Szenarios (als eine der Voraussetzungen für die Definition der Steuerungsanforderungen). Aufgrund dieser Mängel können die in der Eignungsdiagnostik zur Zeit verwendeten Problemlöseszenarien keinen begründeten Anspruch auf Kontentvalidität erheben.

Unabhängig von dem Aspekt der Kontentvalidität stellt sich die Frage nach der Kriteriumsvalidität computergestützter Problemlöseszenarien. Der Aufschwung der Diagnostik mit Problemlöseszenarien dürfte sich nicht unwesentlich der Hoffnung verdanken, mit diesen Szenarien könnten „real-life“-Kriterien in bislang unerreichter Güte vorhergesagt werden. Die hohen Erwartungen hinsichtlich der prognostischen Validität der neuen Instrumente wurden geweckt, indem die angeblichen Vorteile der Problemlöseszenarien den vermeintlichen Nachteilen der Intelligenztests gegenübergestellt und die prognostischen Validitäten der Intelligenztests gleichzeitig als „enttäuschend“ charakterisiert wurden. So war Dörner (1986, S. 292) angesichts einer Korrelation zwischen Intelligenztestleistungen und Examensergebnissen in Höhe von  $r = .45$  überrascht, „daß *nur* [Hervorhebung v. Verf.] 20 % der Unterschiede im Examen von den intellektuellen Leistungen der Studenten abhängen“, 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 Validitäten in bezug auf den Ausbildungs- und Berufserfolg aufzuweisen hat und bislang kaum Möglichkeiten gefunden wurden, die mit Intelligenztests geleisteten Vorhersagen inkrementell zu verbessern (Schmidt & Hunter, 1998), muß ein Forschungsvorhaben, dem diese Validitäten nicht annähernd genügen, Größeres im Sinn haben. Gemessen an diesem Superioritätsanspruch fällt

die Bilanz nach über zwei Jahrzehnten Forschung mit computergestützten Problemlöseszenarien enttäuschend aus. Bislang existieren jeweils zwei hinreichend dokumentierte Studien zur retrograden (Putz-Osterloh & Köster, 1988; Schoppek, 1991) und zur konkurrenten Kriteriumsvalidierung (Obermann, 1991; Schuler, Funke, Moser & Donat, 1995) sowie lediglich eine Untersuchung zur prädiktiven Kriteriumsvalidierung (Hasselmann, 1993) von Problemlöseszenarien. Da die Studie von Hasselmann (1993) aufgrund der sehr geringen Anzahl von nur 21 untersuchten Personen in ihrer Aussagekraft eingeschränkt ist, gilt die Frage der prognostischen Validität als offen. Für den vorliegenden Artikel sind vor allem die Validierungsstudien von Funke et al. (1995), Hasselmann (1993) und Obermann (1991) bedeutsam, da diese Untersuchungen im eignungsdiagnostischen Kontext angesiedelt sind. Diese Studien wurden an anderer Stelle bereits mehrfach zusammenfassend dargestellt (z. B. Funke, 1995), so daß hier auf eine Wiederholung des Überblicks verzichtet wird. Auch wenn sich in allen drei genannten Studien Hinweise auf die Kriteriumsvalidität der Szenarien abzeichneten, bleiben zahlreiche bedeutsame Aspekte der Validitätsfrage ungeklärt. Neben der bereits erwähnten ungeklärten prognostischen Validität ist bislang insbesondere völlig ungeprüft, ob die neuen Instrumente gegenüber anderen Instrumenten mit überlappendem Geltungsbereich (z. B. Intelligenz- und Wissenstests) *inkrementelle* Validität aufweisen. Erst mit einem solchen direkten Vergleich läßt sich die Nützlichkeit der Szenarien im Sinne von Lienert (1967) prüfen.

Ebenso aufklärungsbedürftig wie die Kriteriumsvalidität ist die Frage der Konstruktvalidität der mit Problemlöseszenarien erhobenen Daten. In der Anfangsphase erschöpften sich viele Studien zum Zusammenhang zwischen Indikatoren des vermeintlich neuen Konstrukts „Problemlösefähigkeit“ und Indikatoren etablierter Konstrukte wie der Intelligenz im Aufzeigen ausbleibender Korrelationen (für einen aktuellen Überblick siehe z. B. Kersting, 1999, S. 117–129; Süß, 1996). Die Nullkorrelationen sollten – als diskriminante Validitäten interpretiert – den Neuheitsanspruch der Szenarien stützen. Nachweise der konvergenten Validität blieben aus. Zur Untersuchung der Konstruktvalidität ist es aber notwendig, das interessierende Konstrukt über mindestens zwei Instrumente zu operationalisieren, da Konstruktannahmen unabhängig von einzelnen Meßinstrumenten getroffen und geprüft werden müssen. Aufgrund des überlappenden Geltungsbereiches sollten Studien zur Konstruktvalidität von Problemlöseszenarien außerdem den Zusammenhang des mit den Szenarien vermeintlich gemessenen neuen Konstrukts »Problemlösefähigkeit« und der mit herkömmlichen Tests erfaßten Intelligenz untersuchen. Dieser Zusammenhang galt lange Zeit als ungeklärt, ein Umstand, der zu einer heftigen offenen Debatte führte. Süß (1996, 1999) hat Widerspruch gegen diese häufig postulierte Position der „offenen Debatte“ eingelegt. Aufgrund theoretischer Erwägungen und empirischer Befunde geht Süß davon aus, daß sich die Annahme einer bereicherspezifischen Problemlösefähigkeit, die über die herkömmlichen Konstrukte Intelligenz und Wissen

hinausgeht, erübrigt. Intelligenztest- und Problemlöseleistungen sind Süß zufolge substantiell miteinander korreliert, anders lautende Befunde sind seiner Ansicht nach im wesentlichen auf die zahlreichen Methodenprobleme der entsprechenden Studien zurückzuführen. Für die diagnostische Praxis bedeutet diese Position, daß die Szenarien lediglich neue Instrumente der Intelligenz- und Wissensdiagnostik darstellen. Süß stützt seine Annahmen auf Studien, die er mit Schülern (Süß, Oberauer & Kersting, 1993) und Studenten (Süß, 1999) durchgeführt hat. Eine Replikation dieser Ergebnisse anhand einer Studie mit berufstätigen lebensälteren Erwachsenen steht noch aus.

Aus diagnostischer Perspektive muß innerhalb der Erörterung der Konstruktvalidität schließlich auch ein besonderes Gewicht auf den Generalitätsaspekt gelegt werden. Im eignungsdiagnostischen Kontext will man z. B. von der diagnostischen Information auf Verhalten außerhalb der aktuellen diagnostischen Situation schließen. Diagnostisch verwertbare Szenarien müssen daher trans-situationale Anforderungen an eine übergeordnete, zeitlich zumindest minimal stabile Leistungsfunktion stellen. Die bei verschiedenen Szenarien erzielten Leistungen müssen folglich substantiell miteinander korrelieren. In den diesbezüglichen Studien zeigte sich hingegen, daß Szenarien überwiegend *systemspezifische*, nicht generalisierbare Leistungen erfassen (Putz-Osterloh, 1987; Schaub, 1990; Strohschneider, 1991).

## Fragestellung

Die hier vorliegende Studie soll zur Konstrukt- und Kriteriumsvalidierung von computergestützten Problemlöseszenarien beitragen. Zunächst wird geprüft, ob die Ergebnisse von Süß (1996) zum Zusammenhang von Intelligenz, Wissen und Problemlösen sich anhand einer Untersuchung mit teilweise anderen Meßinstrumenten und einer anderen Personengruppe replizieren lassen. Die sich anschließende Prüfung der Kriteriumsvalidität der Szenarien umfaßt auch die bislang nicht untersuchte Frage, ob die Szenarien bei der Vorhersage von Vorgesetztenurteilen über die berufliche Bewährung gegenüber den vorhandenen Verfahren der Intelligenz- und Wissensdiagnostik ein Inkrement liefern können. Neue Verfahren müssen sich an der optimalen vorhandenen Verfahrensalternative, im Falle der Eignungsdiagnostik sind dies Intelligenztests (siehe z. B. Schmidt & Hunter, 1998), messen lassen. Der Einsatz von zwei unabhängigen Problemlöseszenarien ermöglicht darüber hinaus die Ableitung von Aussagen zur Generalität von Steuerungsleistungen.

## Methode

### Untersuchungsteilnehmer

Untersucht wurden Führungskräfte der Polizei. Der Berufserfolg in der Polizei gilt mit einer durchschnittlichen

beobachteten Validität von  $r = .09$  ( $SD = .12$ ) als schlecht vorhersagbar (Hirsh, Northrop & Schmidt, 1986), so daß einerseits gerade für diesen Bereich neue diagnostische Instrumente gesucht werden und andererseits für die neuen Instrumente gute Aussichten bestehen, die Validität der herkömmlichen Verfahren zu übertreffen. Problemlöseszenarien bieten sich auch aufgrund der »face validity« an, da die Polizeiführung mit komplexen Problemen konfrontiert ist. Berücksichtigt wurden 104 berufserfahrene (mindestens 10 Jahre Berufserfahrung, Median = 17) überwiegend männliche (92 %) Polizeibeamte im Alter von 28 bis 57 Jahren (Median = 35).

### Material und Vorgehensweise der Prädiktorerhebung

61 % der Teilnehmer bearbeiteten den Test zum Berliner Intelligenzstrukturmodell („BIS-4“; Jäger, Süß & Beau-ducel, 1997), die übrigen Teilnehmer bearbeiteten einen unveröffentlichten Intelligenztest der Deutschen Gesellschaft für Personalwesen (DGP-Test). Obwohl der DGP-Test um zusätzliche elf Subtests aus dem BIS-Test ergänzt wurde, waren in dem DGP-Test mit Ausnahme der Verarbeitungskapazität die übrigen drei BIS-Operationsklassen durch vergleichsweise wenige Aufgaben repräsentiert. Aus diesem Grunde beschränkt sich die Auswertung – vergleichbar dem Vorgehen der Auswertung der „Kurzform“ im „BIS-4“-Test – auf zwei Skalen, eine für die „Allgemeine Intelligenz (AI)“ und eine für die in beiden Tests umfangreich repräsentierte „Verarbeitungskapazität (K)“. Die Variation des Testverfahrens war durch administrative Vorgaben seitens der Polizei bedingt. Eine Erörterung kann hier aus Platzgründen nicht gegeben werden; diese findet sich, mit einer Darstellung der hohen Ähnlichkeit der beiden eingesetzten Tests sowie empirischen Daten zur Äquivalenz, bei Kersting (1999).

Vor oder im Anschluß an die Intelligenztests (die Abfolge wurde systematisch variiert) steuerten alle Teilnehmer sowohl die Berliner Version des Szenarios „Schneiderwerkstatt“ (SWS) als auch das Szenario „DISKo“ (Chipfabrik) von U. Funke (1992). Die Leistung in beiden Szenarien blieb statistisch unbeeinflusst von der systematisch variierten Darbietungsabfolge der Szenarien. Im Anschluß an die Steuerung der „SWS“ bearbeiteten die Probanden den kontentvaliden „WIS-2“-Test (siehe Kersting, 1999) zur Erfassung des „SWS“-systemspezifischen Sachwissens. Die im folgenden berichteten Ergebnisse beruhen auf der Skala „Zusammenhänge verbal“. Bei „DISKo“ wird programmintern ein Parameter zur Ausprägung des systemspezifischen Sachwissen bestimmt.

Für vier Teilnehmer konnten die Daten der „DISKo“-Steuerung nicht genutzt werden, eine Person versäumte den Intelligenztest.

### Problemlösegütemaße

Beide Szenarien waren mit dem Ziel zu steuern, das Gesamtvermögen am Ende der Bearbeitungszeit zu maxi-

mieren. Die Steuerungsleistung ist somit instruktionsgemäß am „Kapitalendwert“ zu messen. Bei dem Szenario „DISKO“ wird zusätzlich programmintern ein Parameter zur Beurteilung der Verhaltensweisen und Strategien (im folgenden „Verhaltensmaß“ genannt) berechnet (siehe Funke, 1992). Die Variablen für die Steuerungsleistungen in den beiden Szenarien und das „DISKO“-Verhaltensmaß waren nicht normalverteilt und wurden normalisiert. Alle vorgenommenen Normalisierungen erfolgten über das Verfahren von Blom (1958), bei dem auf der Basis der Rohdaten Ränge ( $r$ ) berechnet werden. Die Ränge werden dann mit  $(r-3/8) / (N+1/4)$  transformiert, wobei  $N$  die Anzahl der Fälle darstellt. Diese Werte werden als kumulative Prozentwerte aufgefaßt, und die korrespondierenden  $z$ -Werte aus der Standardnormalverteilung werden dann als normalisierte Werte ausgegeben.

Seit den von Süß et al. (1993) für eine ältere Version der „SWS“ durchgeführten Aufgabenanalysen ist bekannt, daß das Problemlösegütemaß (PLG) „Kapitalendwert“ bei einer zu hohen Aufgabenschwierigkeit die Steuerungsleistung der Probanden nicht korrekt abbildet. Es wurde daher zunächst geprüft, ob die beiden Szenarien einen angemessenen Schwierigkeitsgrad aufwiesen. Dabei zeigte sich, daß es nur 17 % der Teilnehmer im Szenario „DISKO“ gelang, den „Kapitalendwert“ zu steigern. 38 % der Teilnehmer erwirtschafteten in keinem einzigen „DISKO“-Bearbeitungstakt Gewinn, nur 19 % gelang dies für mehr als die Hälfte der Bearbeitungstakte. 83 % der Probanden haben durch ihre Eingriffe im Szenario „DISKO“ ein schlechteres Ergebnis erzielt, als wenn sie für 12 Bearbeitungstakte die Voreinstellung des Programms ohne irgendeine eigene Entscheidung bestätigt hätten. Für die „SWS“ lauten die entsprechenden Werte 62 %, 7 %, 64 % und 19 %. Für die untersuchte Gruppe war die Aufgabe, das Szenario „DISKO“ zu steuern, zu schwer. Nach Anwendung der von Süß et al. (1993) entwickelten Aufgabenanalyse zeigte sich, daß bei „DISKO“ – ebenso wie bei der älteren „SWS“ Version – das PLG „Kapitalendwert“ unter den Umständen einer zu hohen Schwierigkeit die Steuerungsleistung der Teilnehmer nicht korrekt abbildet. Dies ist darauf zurückzuführen, daß die Zielvariable „Kapitalendwert“ als *Produkt* der beiden Teilziele „Anzahl verkaufter Produktionseinheiten“ und der „Gewinnspanne pro Produktionseinheit“ bestimmt wird. Ist die Steuerungsaufgabe zu schwer, so nimmt die „Gewinnspanne“ bei fast allen Teilnehmern einen *negativen Wert* an. Die negative Gewinnspanne führt dazu, daß die Probanden mit jeder verkauften Produktionseinheit *Verluste* erwirtschaften. Das negative Vorzeichen der Gewinnspanne konterkariert den ursprünglichen Qualitätsnachweis eines guten Problemlösers – nämlich seine guten Verkaufszahlen. Die beiden Teilziele der Steuerung sind unter diesen Umständen inkompatibel. In Folge dessen differenziert das PLG „Kapitalendwert“ nicht mehr zwischen guten und schlechten Problemlösern. Da die Probanden keine Möglichkeit haben, sich über die Ausprägung der „Gewinnspanne“ zu informieren, kann die Inkompatibilität der beiden Teilziele von ihnen nicht erschlossen werden. Analog zu dem Vorgehen bei der früheren Version der „SWS“ wurde nun für „DISKO“ ein

neues PLG gebildet, indem die beiden (vorab  $z$ -transformierten) Teilgütemaße „Anzahl an verkauften Produktionseinheiten“ und „Gewinnspanne pro Produktionseinheit“ addiert (statt multipliziert) wurden. Die so gebildete Variable wurde normalisiert. Für die „SWS“ konnte der empirische Nachweis erbracht werden, daß das herkömmliche PLG mit dem so gebildeten neuen PLG unter der Bedingung einer angemessenen Schwierigkeit äquivalent ist (Süß, 1996). Zu Vergleichszwecken wurde das neue Gütemaß auch für die „SWS“ gebildet. Das neue Gütemaß für „DISKO“ wurde über die ersten acht Bearbeitungstakte berechnet. Bei Kersting (1999) ist diese Beschränkung auf die ersten acht Takte begründet; dort findet sich auch eine ausführliche Beschreibung der Aufgabenanalyse sowie der Bildung des neuen PLGs. In den folgenden Analysen wird für „DISKO“ sowohl das herkömmliche, instruktionsgemäße PLG als auch das neu gebildete Maß analysiert. Für die Frage der diagnostischen Tauglichkeit der Szenarien sollte aber ausschließlich das ursprüngliche PLG herangezogen werden, da in der Praxis das neu gebildete PLG nicht zur Verfügung steht.

## Kriterienerhebung

### Der Fragebogen zur Beurteilung durch die Vorgesetzten

Das Hauptaugenmerk der vorliegenden Studie galt der vergleichenden prädiktiven Kriteriumsvalidierung von Intelligenztests und Problemlöseszenarien. Diesbezüglich kommt der inhaltlichen Angemessenheit und der psychometrischen Qualität des gewählten Kriteriums eine herausragende Bedeutung bei. Um die Wahl des Kriteriums nachvollziehen zu können, muß man mit den Instrumentarien der Personalsteuerung *bei Beamten* vertraut sein. Während es sich bei anderen Berufsgruppen u. U. empfiehlt, die berufliche Bewährung u. a. über sogenannte „harte“ Kriterien zu operationalisieren (z. B. das Gehalt oder die hierarchische Stellung), stellen solche Kriterien für Beamte vergleichsweise schwache Indikatoren beruflicher Bewährung dar. Die Höhe der Besoldung richtet sich bei Beamten nicht nach der individuellen Leistung, sondern nach der Dienstpostenbewertung. Die Beförderung hängt u. a. von formalen Rahmenbedingungen (z. B. Anzahl freier Planstellen) ab und wird de facto vom Dienstalter wesentlich mitbestimmt und vom Endamt der jeweiligen Laufbahn begrenzt. Demgegenüber kann die *dienstliche Beurteilung* ein vergleichsweise unmittelbarer und bedeutsamer Indikator der beruflichen Bewährung von Beamten sein. 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* Dienstpostens erfüllt hat. Um einen dienstpostenübergreifenden Vergleich von Beamten unterschiedlicher Aufgabenfeldern zu ermöglichen, erstreckt sich die dienstliche Beurteilung vielmehr explizit auch auf die Beurteilung im Dienst gezeigter sogenannter „allgemeiner Fähigkeiten“ – wie z. B. der „Denk- und Urteilsfähig-

keit“. Um die für den Berufserfolg von Beamten bedeutsamen dienstlichen Beurteilungen nachzuvollziehen, wurde zur Validierung ebenfalls eine Vorgesetztenbefragung durchgeführt. Allerdings wurden verschiedene Maßnahmen ergriffen, um bei der Befragung eine höhere psychometrische Qualität zu erzielen als bei üblichen dienstlichen Beurteilungen. So wurden z. B. ausschließlich Skalen (anstelle von Single-Items) berücksichtigt, und es wurde auf die Korrespondenz zwischen Prädiktoren und Kriterien geachtet.

In dem zur Kriteriengewinnung neu konstruierten Fragebogen wurden die Vorgesetzten zunächst gebeten, auf einer vierstufigen Skala anzugeben, ob der Beurteilte hinsichtlich seiner im Berufsalltag gezeigten (1.) Intelligenz, (2.) Problemlösefähigkeit und (3.) Kooperationsfähigkeit (a) zu den unteren 25 %, zum (b) schlechteren oder (c) besseren „Mittelfeld“ (jeweils 25 %) oder zu den (d) oberen 25 % der Vergleichsgruppe (Laufbahnkollegen) gehört. Ein dem Beurteilungsbogen beigefügtes Informationsblatt mit ausführlichen Definitionen der drei interessierenden Fähigkeiten diente der Sicherung eines einheitlichen Verständnisses. Nachdem mit den ersten drei Fragen die Ausprägungen der interessierenden Fähigkeiten ganz allgemein beurteilt werden sollten, wurden mit den folgenden Items konkrete Leistungen beschrieben, die als Indikatoren der im Berufsalltag gezeigten „Intelligenz“, „Problemlösefähigkeit“ und „Kooperationsfähigkeit“ gewertet werden können. Die Vorgesetzten wurden bei jedem Item gefragt, wie die konkret beschriebene Leistung bei der zu beurteilenden Person im Vergleich mit Laufbahn-Kollegen ausgeprägt ist (sechsstufige Skala von [1] „sehr schwache“ bis [6] „sehr hohe“ Ausprägung). Drei dieser Items dienten der Beurteilung der im Berufsalltag gezeigten Kooperationsfähigkeit (Skala 1, Cronbach's  $\alpha = .90$ ), vier Items der Beurteilung der im Berufsalltag gezeigten Intelligenz (Skala 2,  $\alpha = .81$ ) und fünf Items der Beurteilung der im Berufsalltag gezeigten Problemlösefähigkeit (Skala 3,  $\alpha = .85$ ). Die von den Vorgesetzten vorgenommene globale Einschätzung der im Berufsalltag gezeigten Intelligenz der Polizisten korrelierte zu  $r = .64$  mit der aus den diesbezüglichen spezifischen Items gebildeten Skala 2. In ähnlicher Höhe korrelierten die allgemeinen und spezifischen Beurteilungen der im Berufsalltag gezeigten Problemlösefähigkeit ( $r = .55$ ) und Kooperationsfähigkeit ( $r = .61$ ) miteinander. Die globale Beurteilung wurde daher für alle drei interessierenden Bereiche mit der jeweiligen spezifischen Beurteilung (Skalen 1, 2 und 3) zusammengefaßt (einfache Aggregationsbildung mit vorangehender z-Transformation), wodurch eine Interpretation von Daten auf der Basis von einzelnen Items vermieden wurde. Lediglich das aggregierte Maß der Urteile über die „Kooperationsfähigkeit“ entsprach nicht der Normalverteilung und wurde normalisiert. Auf der Aggregatsebene waren die Urteile über die im Berufsalltag gezeigte Intelligenz der Polizisten zu  $r = .71$  mit den Urteilen über ihre im Berufsalltag gezeigte Problemlösefähigkeit korreliert, wodurch die Bildung einer weiteren Kriteriumsvariablen indiziert war (einfaches Aggregat der zwei z-transformierten Variablen).

Die Vorgesetzten wurden zusätzlich um eine Einschätzung der Bedeutung der erfragten Merkmale für die erfolgreiche berufliche Bewährung der beurteilten Polizisten gebeten. Die bei Kersting (1999, S. 238f.) dokumentierten Ergebnisse dieser Befragung bestätigen die hohe berufliche Relevanz der erfaßten Kriterien.

Der vollständige Fragebogen der Vorgesetztenbeurteilung und alle Informationen zur Skalenbildung können unter der Web-Adresse <http://www.psych.rwth-aachen.de/~kersting> eingesehen werden.

## Kriterienerhebung: Zeitpunkt und Teilnehmer

Die postalisch durchgeführte Erfassung der Kriteriendaten erfolgte neunzehneinhalb Monate (Median = 20 Monate; SD = 4.5 Monate) nach der Prädiktorerhebung. Eine Subgruppe der Teilnehmer im Umfang von 26 Personen hatte zum Erhebungszeitpunkt ihre aktive Berufstätigkeit für ein Studium an der Polizeiführungsakademie unterbrochen. Dementsprechend konnte für diese Gruppe keine Befragung zur aktuellen beruflichen Bewährung vorgenommen werden. Für 73 Teilnehmer konnten die Kriteriendaten erhoben werden, dies entspricht bei 78 angefragten Ergebnissen einer Rücklaufquote von 93.6 %. Die Beurteilung wurde überwiegend (87.7 %) von dem unmittelbaren Linienvorgesetzten der Untersuchungsteilnehmer abgegeben, in den übrigen Fällen handelte es sich um Beurteilungen durch Kollegen der beurteilten Personen. Da die Polizisten unterschiedliche Vorgesetzte haben, wurden die Beurteilungen zwangsläufig von verschiedenen Beurteilern ausgesprochen. Dies entspricht der Berufspraxis, wo ebenfalls dienstliche Beurteilungen verschiedener Vorgesetzter miteinander verglichen werden müssen. Die dem Fragebogen beigefügten ausführlichen Beschreibungen der interessierenden Fähigkeiten und die Items mit Darstellungen konkreter Leistungen sollten hier der Wahrung eines einheitlichen Vergleichsmaßstabes dienen. Die Beurteiler wurden nicht über die Befunde der Prädiktorerhebung informiert.

## Ergebnisse

### Konstruktvalidität der Steuerungsleistung

Tabelle 1 verzeichnet die Interkorrelationen aller zur Messung der Intelligenz, der Steuerungsleistung und des systemspezifischen Wissens eingesetzten Variablen. Bei den für die Problemlösefähigkeit und für das systemspezifische Wissen über die beiden Szenarien hinweg gebildeten Aggregaten handelt es sich um die einfache Summe der vorab z-transformierten Einzelvariablen.

Die Interkorrelationen der Indikatoren des Steuerungserfolgs in den beiden Szenarien variieren deutlich in Abhängigkeit von dem jeweils zugrunde gelegten Gütemaß. Während die bei den beiden Szenarien erzielten Kapitalendwerte zu  $r = .05$  miteinander dissoziiert sind, korreliert das neu berechnete PLG für „DISKO“ signifi-

Tabelle 1. Korrelationen der Steuerungsleistungen mit Wissen und Intelligenz

		Intelligenz		Sachwissen über ...			Steuerungsleistung					
		AI	K	SWS	DISKo	Aggr.	SWS		DISKo			Aggr.
							Kap.	PLG	KAP	PLG8	Verh.	
Intelligenz	AI		N=103	N=103	N= 99	N= 99	N=103	N=103	N= 99	N= 99	N= 99	N= 99
	K	.83**		N=103	N= 99	N= 99	N=103	N=103	N= 99	N= 99	N= 99	N= 99
Wissen über...	SWS	.42**	.50**		N=100	N=100	N=104	N=104	N=100	N=100	N=100	N=100
	DISKo	.00	.03	.15		N=100	N=100	N=100	N=100	N=100	N=100	N=100
	Aggregat	.27*	.37**	.76**	.71**		N=100	N=100	N=100	N=100	N=100	N=100
Steuerung SWS	Kapital	.20*	.21*	.26*	.21*	.35**		N=104	N=100	N=100	N=100	N=100
	PLG	.23*	.22*	.22*	.22*	.29**	.59**		N=100	N=100	N=100	N=100
Steuerung DISKo	Kapital	.00	.02	.19*	.23*	.26*	.05	-.02		N=100	N=100	N=100
	PLG-8	.30**	.31**	.33**	.31**	.39**	.28**	.39	.09		N=100	N=100
	Verhalten	.32**	.33**	.21*	.11pw	.18pw	.26*	.26	.14	.24*		N=100
Steuerung	Aggr.	.33**	.34**	.34**	.32**	.41**	.51**	.83**	.04	.83**		

Anmerkungen: \*  $p < .05$ , \*\*  $p < .01$ , <sup>pw</sup> part-whole korrigierte Korrelationen; SWS: „Schneiderwerkstatt“; AI: Intelligenztestskaala „Allgemeine Intelligenz“; K: Intelligenztestskaala „Verarbeitungskapazität“; Kapital: Kapitalendwert der Systemsteuerung; PLG: Problemlösegütemaß; PLG-8: berechnet über die ersten acht Bearbeitungsstakte; Verhalten: programminterne Beurteilung der Verhaltensweisen und Strategien; Aggr.: Aggregat.

kant zu  $r = .28$  mit dem Kapitalendwert bei der „SWS“ und zu  $r = .39$  mit dem für die „SWS“ berechneten neuen PLG. Dieser statistische Zusammenhang zwischen den Steuerungsleistungen bei beiden Szenarien ist allerdings – wie Kersting (1999) zeigen konnte – wesentlich auf deren gemeinsamen Intelligenz- und Wissensanteil zurückzuführen. Der Nachweis einer generellen, szenarienübergreifenden Problemlösefähigkeit steht weiterhin aus.

Für das Szenario „DISKo“ variiert auch die Höhe der Korrelation zwischen Intelligenz und Wissen einerseits sowie den Steuerungsleistungen andererseits in Abhängigkeit von dem jeweiligen PLG. Nur mit dem neuen PLG, nicht aber mit dem herkömmlichen Indikator „Kapitalendwert“ kann dieser Zusammenhang aufgezeigt werden.

Der Zusammenhang zwischen Intelligenz, Wissen und der Steuerungsleistung läßt sich am besten abschätzen, wenn man als Indikator der Steuerungsleistung das über beide Szenarien hinweg aggregierte – und somit reliabilitätsgesteigerte – neue Gütemaß betrachtet. Auf dieser Ebene zeigen sich substantielle Korrelationen zwischen den Intelligenz- und Wissensindikatoren einerseits und dem Problemlöseerfolg andererseits.

Über die bivariaten Analysen hinaus wurde mit einer hierarchischen Regressionsanalyse geprüft, ob durch eine Prädiktorkombination der Variablen „systemspezifisches Wissen“ und „Intelligenz“ die Vorhersage der Steuerungsleistung verbessert werden kann. Als Kriterium galt das über beide Szenarien aggregierte neue PLG. Zunächst wurde der Wissensindikator für die „SWS“, dann der Wissensindikator für „DISKo“ und schließlich die Skala „Allgemeine Intelligenz“ in die Regressionsgleichung eingeführt. Jede zusätzlich berücksichtigte Variable führte zu einer statistisch bedeutsamen Steigerung des

$R^2$ -Wertes. Nachdem alle Prädiktoren in die Regressionsgleichung aufgenommen worden waren, beträgt  $R = .49$  ( $F(3,95) = 9.97, p < .001$ ).

## Prädiktive Kriteriumsvalidierung

Tabelle 2 führt für die Prädiktoren die Korrelationen mit den zeitlich versetzt erhobenen Kriterien auf. Für die Fälle, in denen sich überzufällige Zusammenhänge ergeben, sind zusätzlich Partialkorrelationen verzeichnet. Trotz der Ähnlichkeit ( $r = .71$ ) der Urteile über die im Berufsalltag gezeigte Intelligenz einerseits und über die Problemlösefähigkeit andererseits variiert die Validität der Prädiktoren in Abhängigkeit von dem jeweils vorhergesagten Kriterium. Jedes Verfahren hat nominell seine Vorhersagestärke für den theoretisch ähnlichsten Kriterienbereich.

Bei der „SWS“ können mit dem – den Zielvorgaben der Steuerung entsprechenden – Standardmaß der Problemlösegüte, nämlich dem „Kapitalendwert“, die intendierten Kriterien prognostiziert werden. Beim Szenario „DISKo“ erwies sich dieses herkömmliche PLG hingegen als prognostisch unbrauchbar. Dieser Befund reiht sich in die Kette der übrigen Belege für die Tatsache ein, daß der „DISKo-Kapitalendwert“ für die hier analysierte Gruppe kein sinnvoll interpretierbares PLG ist. Im „normalen“ Anwendungsfall wären unter Umständen aufgrund dieses sinnlosen „Problemlösegütemaßes“ schwerwiegende Personalentscheidungen getroffen worden. Erst die Berücksichtigung des neu bestimmten PLGs zeigt die Vorhersagekraft der „DISKo“-Steuerungsleistungen.

Für alle Prädiktoren kann anhand des Vorgesetztenurteils über die im Beruf gezeigte Kooperationsfähigkeit der Polizisten der Nachweis der diskriminanten Validität erbracht werden.

Tabelle 2. Korrelation der Prädiktoren mit dem Vorgesetztenurteil über die im Berufsalltag gezeigte Intelligenz („Int.“) und Problemlösefähigkeit („Prob.“) der Polizisten sowie mit dem Aggregat („Aggr.“) dieser beiden Urteile und (zur diskriminanten Validierung) mit dem Urteil über die im Berufsalltag gezeigte Kooperationsfähigkeit („Koop.“)

Prädiktoren	Kriterien: Vorgesetztenurteil über ...				N
	Int.	Prob.	Aggr.	Koop.	
Skala Allg. Intell. (AI)	.43**	.28*	.39**	.07	72
<i>Nach stat. Kontrolle von</i>					
SWS-Kapitalendwert (SWS)	.39**	.21 <sup>1</sup>	.33**		
Sachwissen über SWS <sup>□</sup>	.38**	.23 <sup>1</sup>	.33**		
SWS & Sachwissen üb. SWS <sup>□</sup>	.36**	.19	.30*		
SWS Kapitalendwert	.31**	.37**	.37**	.19	73
<i>Nach stat. Kontrolle von</i>					
Allg. Intelligenz (AI)	.21 <sup>1</sup>	.31**	.29*		
Sachwissen über SWS <sup>□</sup>	.23 <sup>1</sup>	.30**	.29*		
AI & Sachwissen über SWS <sup>□</sup>	.14	.26*	.22 <sup>1</sup>		
DISKo Kapitalendwert	-.01	-.17	-.10	-.09	69
DISKo neues PLG / 8 Takte	.16	.33**	.26*	.08	69
<i>Nach stat. Kontrolle von</i>					
Allg. Intelligenz (AI)	.02	.26*	.16		
Sachwissen über DISKo	.14	.34**	.26*		
Sachwissen über SWS <sup>□</sup>	.09	.28*	.20 <sup>1</sup>		
AI & Sachwissen über SWS <sup>□</sup>	-.03	.23 <sup>1</sup>	.11		
DISKo Verhaltensbeurteilung	.26*	.17	.23*	.11	69
<i>Nach stat. Kontrolle von</i>					
Allg. Intelligenz (AI)	.02	.02	.02		
Sachwissen über SWS <sup>□</sup>	.32**	.28**	.32**	.02	73
<i>Nach stat. Kontrolle von</i>					
Allg. Intelligenz (AI)	.26*	.23 <sup>1</sup>	.27*		
SWS-Kapitalendwert	.24*	.17	.22 <sup>1</sup>		
AI und SWS-Kapitalendwert	.21 <sup>1</sup>	.15	.19		
Sachwissen über DISKo	.10	.03	.07	.02	69

Anmerkungen: <sup>□</sup> Sachwissen über „SWS“ = Skala I des „WIS-2-Tests“; \* p < .05; \*\* p < .01; <sup>1</sup> p < .10.

In den bei Kersting (1999, S. 239 ff.) dargestellten Relevanzeinschätzungen haben die Vorgesetzten zum Ausdruck gebracht, daß für die Bewährung im Polizeiberuf sowohl die Intelligenz als auch die Problemlösefähigkeit bedeutsam ist. Die folgenden Analysen beschränken sich daher auf das *aggregierte* Maß.

Mit Hilfe der multiplen Regressionsanalyse wurde geprüft, ob durch die Kombination der Prädiktoren die

Vorhersagegenauigkeit inkrementell gesteigert werden kann. Bei den Analysen stellte sich heraus, daß die Vorhersagekraft der Steuerungsleistung durch die Berücksichtigung der „SWS“ ausreichend repräsentiert ist und die zusätzliche Einbeziehung der mit dem Szenario „DISKo“ erzielten Indikatoren die Validität dieser Prädiktorgruppe nicht steigert. Auf der Ebene des Sachwissens konnte sich die Analyse auf das Wissen zur „SWS“ beschränken, der zusätzlich berücksichtigte Indikator für

das „DISKO“-spezifische Wissen steigert die Vorhersagekraft dieser Prädiktorgruppe nicht weiter. Die Intelligenztestleistungen wurden in Form der Skala „Allgemeine Intelligenz“ in die Prädiktion einbezogen. Diese Skala korreliert bei der analysierten Gruppe im Umfang von 72 Personen zu  $r = .26$  ( $p < .05$ ) mit der „SWS“-Steuerungsleistung (Kapitalendwert) und zu  $r = .25$  ( $p < .05$ ) mit der berücksichtigten Wissensskala. Die beiden zuletzt genannten Prädiktoren sind zu  $r = .36$  ( $p < .01$ ) miteinander assoziiert. Zum Ende des letzten Schritts der Regressionsanalyse, nach Aufnahme aller Prädiktoren in die Regressionsgleichung, beträgt die multiple Korrelation  $R = .50$  ( $F(3,68) = 7.53$ ,  $p < .001$ ). Im ersten Schritt wurde aufgrund der höchsten bivariaten Korrelation die Variable „Allgemeine Intelligenz“ aufgenommen. Mit dem Einbezug des Wissensindikators als zweitem Prädiktor steigt das  $R$  von  $.39$  auf  $.46$  an ( $F_{change}(2,69) = 5.25$ ,  $p < .05$ ). Dies bedeutet eine zusätzliche Varianzaufklärung von fast 6 %. Die durch die Aufnahme des dritten Prädiktors – der Steuerungsleistung bei der „SWS“ – erzielte Steigerung der Vorhersagegenauigkeit beträgt 3.8 % und ist lediglich auf einem Signifikanzniveau von 10 % statistisch bedeutsam ( $F_{change}(3,68) = 3.54$ ,  $p < .10$ ). Zur Prüfung der Stichprobenunabhängigkeit der multiplen Korrelation wurde eine Kreuzvalidierung durchgeführt, indem die an der Gesamtgruppe ermittelte Gleichung auf die Daten von zwei stratifizierten Zufallsteilstichproben angewandt wurde. Die an der Gesamtgruppe gewonnenen Regressionsgewichte führen in den beiden Teilstichproben zu einer multiplen Korrelation von  $R = .53$  und  $R = .49$ . Die Berücksichtigung eines gegenüber der Intelligenz zusätzlichen Prädiktors liefert somit einen inkrementellen Beitrag zur Steigerung der Vorhersageleistung. Statistisch erwiesen sich die Steuerungsleistung und das Wissen als gleichermaßen geeignet zur Prädiktionssteigerung. Die im zweiten Schritt zusätzlich zur Intelligenz in die Gleichung aufgenommene Variable führte zu einer statistisch bedeutsamen Steigerung des  $R^2$ -Wertes, die im dritten Schritt berücksichtigte Variable hingegen bei der gegebenen statistischen *power* nicht.

Die relative Bedeutung der in der Regressionsanalyse verwendeten Prädiktoren wurde schließlich mit Hilfe der Kommunalitätenanalyse (Kerlinger & Pedhazur, 1973) untersucht, die gegenüber der hierarchischen Regressionsanalyse den Vorteil einer von theoretischen Setzungen unabhängigen Betrachtung der überlappenden Varianzanteile bietet. Die insgesamt aufgeklärte Varianz läßt sich mit Hilfe der Kommunalitätenanalyse separieren in *spezifische Varianz*, die nur durch jeweils einen der drei Prädiktoren „Intelligenz“, „Wissen“ und „Steuerungsleistung“ vorhergesagt wird sowie in *konfundierte Varianz*, die durch zwei oder drei der Prädiktoren erklärt wird. Insgesamt können 24.9 % der Kriteriumsvarianz vorhergesagt werden. Relativ gesehen stellt die Intelligenz mit 7.3 % den größten Anteil an spezifischer Varianz und ist somit ein unverzichtbarer Prädiktor. Auf die Steuerungsleistung entfällt 3.9 %, auf das Wissen 3.0 % spezifische Varianz. An konfundierter Varianz ergaben sich 2.5 % für die Kombination aus Intelligenz und Steuerungsleistung, 2.0 % für die Kombination Intelligenz und Wissen und

3.1 % für die Kombination aus Steuerungsleistung und Wissen sowie 3.2 % für die Kombination aller drei Prädiktoren.

## Diskussion und Ausblick

Der diagnostische Einsatz von computergestützten Problemlöseszenarien setzt voraus, daß sich aus den Steuerungsleistungen valide Informationen über die Diagnostikanden ableiten lassen. Die Validierung der Problemlöseszenarien wurde ungebührlich vernachlässigt und ist bis heute defizitär. Problemlöseszenarien können keine Kontentvalidität beanspruchen, Validitätshinweise finden sich allerdings im Bereich der Beziehungen der Steuerungsleistungen zu Außenkriterien. Auch in der hier vorgestellten Studie konnte der Nachweis der Kriteriumsvalidität geführt werden, allerdings ließ sich der Großteil der kriteriumsrelevanten Varianz der Steuerungsleistung mit den vorhandenen Intelligenz- und Wissensindikatoren empirisch abbilden. Dieser Befund stellt (1) die Nützlichkeit und (2) die Konstruktvalidität computergestützter Problemlöseszenarien in Frage. Ad 1: Entgegen den von Dörner (z. B. 1986, S. 292) und anderen geweckten Erwartungen sind Problemlöseszenarien im Vergleich zu Intelligenztests vergleichsweise wenig valide. Nach Maßgabe der inkrementellen Kriteriumsvalidität stellt die Steuerungsleistung in Problemlöseszenarien auch keine sinnvolle Ergänzung einer Kombination aus Intelligenz- und szenarienspezifischen Wissenstests dar. Die neuen Instrumente ändern somit nichts an dem bekannten Befund, daß Intelligenztests über die höchste Validität bezüglich der Vorhersage berufsbezogenen Lernens und beruflicher Leistung verfügen und daß mit Ergänzungen anderer Verfahren kaum Validitätszuwächse zu erzielen sind (Schmidt & Hunter, 1998).

Ad 2: Da die mit den Problemlöseszenarien geleisteten Messungen in ihrem systematischen Anteil an der Kriteriumsvarianz kaum etwas indizieren, was nicht bereits durch Intelligenz- und Wissenstests indiziert wird, ist eine Interpretation der Steuerungsleistung als Indikator für (Problemlöse-) Fähigkeit, für ein neues Konstrukt also, empirisch kaum zu legitimieren (siehe auch Süß, 1996). Die Szenariensteuerung stellt Anforderungen an eine *Kombination* aus Intelligenz und Wissen.

Die Tatsache, daß nicht nur durch den Wissenstest, sondern auch durch die Steuerungsleistung im Problemlöseszenario die Prognose inkrementell gesteigert werden konnte, kann als statistische Rechtfertigung für eine Kombination aus Intelligenztests und Problemlöseszenarien gewertet werden. Diese Kombinationsmöglichkeit ist problematisch, weil dabei die beiden sinnvoll differenzierbaren Konstrukte „Intelligenz“ und „Wissen“ in der Steuerungsleistung zu einem nicht mehr unterscheidbaren Konglomerat verschmolzen werden. Darüber hinaus ist bei computergestützten Problemlöseszenarien nach wie vor die Frage der Voraussetzungenfreiheit (Einfluß der Computererfahrung und der Einstellung zur Computernutzung) und die damit verknüpfte Frage der geschlechtsspe-



zifischen Fairneß ungeklärt. Schließlich lassen sich bei der Szenariensteuerung vergleichsweise leicht und drastisch Fähigkeiten vortäuschen (zu den zuletzt genannten Aspekten siehe Kersting, 1999, S. 164 f.). Computergestützte Problemlöseszenarien bieten damit keinen Zugang zu einem neuen Konstrukt oder zu einer valideren Eignungsdiagnostik. Gleichwohl bestätigt die Studie die Forderung nach einer Erweiterung der herkömmlichen Intelligenzdiagnostik. Die Steuerungsleistungen bei komplexen Szenarien gehen nämlich insofern über die Intelligenz hinaus, als bei der Steuerung nicht nur Intelligenz, sondern zusätzlich auch systemspezifisches Wissen erforderlich ist. Durch die Hinzunahme eines systemspezifischen Wissenstests konnte die durch Intelligenztests geleistete Prognose inkrementell gesteigert werden. Für die diagnostische Praxis könnte es sich daher lohnen, die Intelligenzdiagnostik durch eine Wissensdiagnostik zu ergänzen. Problemlöseszenarien könnten bei der Wissensdiagnostik hilfreich sein, falls bei der Bearbeitung der Szenarien – begünstigt durch das verfahrenskennzeichnende Merkmal des Feedbacks – Wissenserwerb und Wissensanwendung stattfindet und falls es gelingt, diese Aspekte zuverlässig zu messen. Die Erweiterung der Intelligenzdiagnostik um eine auf Problemlöseszenarien aufbauende Wissens(erwerbs)diagnostik erscheint vielversprechend. Eine Ursache für die derzeitige diagnostische Vernachlässigung von Wissen dürfte darin liegen, daß Wissen nicht den Grad der Voraussetzungenfreiheit erfüllt, den man in vielen diagnostischen Anwendungsfeldern (z. B. bei der Auswahl von *Berufseinsteigern*) erwartet. Interessant ist eigentlich nicht das Wissen selbst, welches rasch veraltet, sondern die Fähigkeit zum Wissenserwerb und zur Wissensnutzung. Die Szenariensteuerung könnte als Anwendungsfeld intellektueller Fähigkeiten über Prozesse der systematischen Hypothesentestung Gelegenheit zur Aneignung und zum Ausbau von Wissen bieten. Da alle Probanden Gelegenheit zur Systemsteuerung und somit Zugang zur Erfahrungsgrundlage erhalten, ist die anschließende Diagnose des systemspezifischen Wissens vergleichsweise voraussetzungsärmer als der Einsatz herkömmlicher Kenntnistests. Das von Kröner (2001) entwickelte Instrument „MultiFlux“ entspricht z. B. diesem Ansatz, indem hier die Szenarientechnik explizit für die Diagnostik des Wissenserwerbs und der intelligenten Wissensanwendung genutzt wird. Die Perspektive der Nutzung von Problemlöseszenarien zur Provokation von testbaren Wissenserwerbsprozessen geht als Ausblick explizit über den empirischen Beitrag der vorliegenden Arbeit hinaus. Bei der vorgestellten Studie ist es im nachhinein nicht möglich zu entscheiden, ob es sich bei dem prognostisch validen Wissensanteil um systemspezifisches Vorwissen oder – diagnostisch reizvoller – um erworbenes Wissen handelte.

Für jegliche Form des diagnostischen Einsatzes von computergestützten Problemlöseszenarien wäre es förderlich, wenn zukünftig der sorgfältigen Evaluation Vorrang vor der Konstruktion immer neuer Problemlöseszenarien und Szenarienvarianten eingeräumt würde. Funke (1998, S. 95) hat für die Zukunft computergestützter Szenarien die Integration von Video-Clips sowie Elementen der

virtuellen Realität in Aussicht gestellt. Für die *Forschung* sind diese Ansätze zu begrüßen. Zu befürchten ist aber, daß auch diese Neuerungen unmittelbar und umstandslos Einzug in die diagnostische Praxis halten werden. Somit würde auch weiterhin das Bestreben um die Adaption an kurzlebige äußerlich-technische Aktualitäten die permanente Veränderung bestehender und die Konstruktion immer neuer Problemlöseszenarien motivieren. Auf diese Art und Weise sind bereits etliche Instrument-Unikate erzeugt worden, aber kaum replizierte Befunde. Diese Halbwertszeit der Instrumente vereitelt längerfristige Evaluationen, ein kumulativer Erkenntnisgewinn findet nicht statt. Der Test (und beim diagnostischen Einsatz von Problemlöseszenarien werden 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). In diesem Sinne kommt es in der Diagnostik auch darauf an, mit einem Instrument ein gewisses Maß an Erfahrung zu sammeln, bevor man es für den diagnostischen Ernstfall heranzieht und bevor man Modifikationen an dem Instrument vornimmt. Neue diagnostische Entwicklungen sollten auf das zielen, worauf es ankommt, nicht auf das, was ankommt.

## Literatur

- Blom, G. (1958). *Statistical estimates and transformed beta variables*. New York: John Wiley & Sons.
- Dörner, D. (1986). Diagnostik der operativen Intelligenz. *Diagnostica*, 32, 290-308.
- Dörner, D. & Kreuzig, H. W. (1983). Problemlösefähigkeit und Intelligenz. *Psychologische Rundschau*, 34, 185-192.
- Dörner, D. & Reither, F. (1978). Über das Problemlösen in sehr komplexen Realitätsbereichen. *Zeitschrift für Experimentelle und Angewandte Psychologie*, 25, 527-551.
- 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, U. (1992). *Diagnostisches interaktives System zur Komplexitätssimulation „DISKO“*. Filderstadt: Care applications.
- Funke, U. (1995). 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.
- Hasselmann, D. (1993). *Computersimulierte komplexe Problemstellungen in der Management-Diagnostik*. Hamburg: Windmühle.
- Hirsh, H. R., Northrop, L. C. & Schmidt, F. L. (1986). Validity generalization results for law enforcement occupations. *Personnel Psychology*, 39, 399-420.
- Jäger, A.O., Süß, H. M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test*. Göttingen: Hogrefe.

- Kerlinger, F. N. & Pedhazur, E. J. (1973). *Multiple regression in behavioral research*. New York: Holt, Rinehart and Winston.
- Kersting, M. (1999). *Diagnostik und Personalauswahl mit computergestützten Problemlösenszenarien? Zur Kriteriumsvalidität von Problemlösenszenarien und Intelligenztests*. Göttingen: Hogrefe.
- Kröner, S. (2001). *Intelligenzdiagnostik per Computersimulation*. Münster: Waxmann.
- Lienert, G. A. (1967). *Testaufbau und Testanalyse*. Weinheim: Beltz.
- Obermann, C. (1991). *Airport. Problemlösesimulation V.2.2. Handbuch*. Göttingen: Hogrefe.
- Putz-Osterloh, W. (1987). Gibt es Experten für komplexe Probleme? *Zeitschrift für Psychologie*, 193, 63–84.
- Putz-Osterloh, W. & Köster, K. (1988). Diagnostik komplexer Entscheidungsstrategien bei einem computersimulierten Planspiel. *Untersuchungen des psychologischen Dienstes der Bundeswehr*, 23, 223–255.
- Schaub, H. (1990). Die Situationsspezifität des Problemlöseverhaltens. *Zeitschrift für Psychologie*, 198, 83–96.
- Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262–274.
- Schoppek, W. (1991). Spiel und Wirklichkeit – Reliabilität und Validität von Verhaltensmustern in komplexen Situationen. *Sprache & Kognition*, 10, 15–27.
- Schuler, H., Funke, U., Moser, K. & Donat, M. (1995). *Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren*. Göttingen: Hogrefe.
- Stern, W. (1911). *Die differentielle Psychologie in ihren methodischen Grundlagen*. Leipzig: Ambrosius Barth.
- Strohschneider, S. (1991). Problemlösen und Intelligenz: Über die Effekte der Konkretisierung komplexer Probleme. *Diagnostica*, 37, 353–371.
- Süß, H. M. (1996). *Intelligenz, Wissen und Problemlösen*. Göttingen: Hogrefe.
- Süß, H. M. (1999). Intelligenz und komplexes Problemlösen: Perspektiven für eine Kooperation zwischen differentiellpsychometrischer und kognitionspsychologischer Forschung. *Psychologische Rundschau*, 50, 220–228.
- Süß, H. M., Oberauer, K. & Kersting, M. (1993). Intellektuelle Fähigkeiten und die Steuerung komplexer Systeme. *Sprache & Kognition*, 12, 83–97.

Dr. Martin Kersting

Institut für Psychologie  
der RWTH Aachen  
Jägerstr. 17/19  
D-52056 Aachen  
E-Mail: Kersting@psych.rwth-aachen.de