Kersting, M. (2003). Assessment Center: Erfolgsmessung und Qualitätskontrolle. In S. Höft & B. Wolf (Hrsg.), *Qualitätsstandards für Personalentwicklung in Wirtschaft und Verwaltung* (S. 72-93). Hamburg: Windmühle.

Stefan Höft, Bernd Wolf

Qualitätsstandards für Personalentwicklung in Wirtschaft und Verwaltung

Wie Konzepte greifen
Mit zahlreichen Umsetzungsbeispielen
aus der Praxis

Herausgeber
Arbeitskreis Assessment Center e.V.

Band 4
Reihe Assessment Center

Assessment Center: Erfolgsmessung und Qualitätskontrolle

Martin Kersting

Der Autor verdeutlicht in seinem Beitrag die Bedeutung und den Nutzen von Erfolgskontrollen für Personalmaßnahmen. Er illustriert dies am Beispiel eines Auswahl-ACs für Führungskräfte der Polizei, das in mehreren deutschen Bundesländern als Entscheidungshilfe für eine Zulassung zur Ausbildung für den höheren Polizeivollzugsdienst durchgeführt wird. Der Autor diskutiert ausführlich die bestehenden Befunde zur Güte des Assessment Centers und geht dabei auf wichtige (statistische) Randbedingungen ein, die bei einer AC-Evaluation berücksichtigt werden müssen. Kritisch beleuchtet er den (Zusatz-) Nutzen von ACs für die Prognose von Ausbildungs- und Berufserfolg gegenüber anderen eignungsdiagnostischen Verfahren.

Umsetzung der Standards

Standard 1:
Zwar wird gesagt, dass das geschilderte AC Bestandteil der Personalentwicklung der Behörde ist, dies wird jedoch nicht weiter ausgeführt bzw. detailliert beschrieben.

Standard 2:
Unter „Unternehmensziel" kann im vorliegenden Fall verstanden werden, dass die zentralen Verantwortungs- und Führungspositionen bei der Polizei der betreffenden Bundesländer nur von Personen eingenommen werden sollen, die über ein nachweislich vorhandenes und zumindest zufriedenstellendes Potenzial an sozialen und intellektuellen Kompetenzen verfügen. Die Umsetzung dieses Ziels wird durch ein Assessment Center für die Auswahl zur Aufstiegsausbildung und in der Folge zum höheren Polizeivollzugsdienst sichergestellt.

Standard 3:
Aus der Zulassungsquote zum höheren Dienst von lediglich zwei Prozent aller Beamten lässt sich unter Berücksichtigung von Sonderfällen und Ausnahmen prinzipiell der absolute Personalbedarf und Mittelbedarf im Sinn der Anzahl durchzuführender ACs ableiten.

Standard 4:
In nahezu mustergültiger Weise wird die Gültigkeit/Tauglichkeit des angewandten ACs zur Vorhersage des Ausbildungserfolgs an der Polizei-Führungsakademie empirisch belegt (kriteriumsbezogene Validität). Dies ist insbesondere auch deshalb bemerkenswert, weil es sich hierbei um eine der immer noch raren Beispiele für eine eigene Bewährungskontrolle handelt. Es wird die

Assessment Center: Erfolgsmessung und Qualitätskontrolle

Martin Kersting

Der Autor verdeutlicht in seinem Beitrag die Bedeutung und den Nutzen von Erfolgskontrollen für Personalmaßnahmen. Er illustriert dies am Beispiel eines Auswahl-ACs für Führungskräfte der Polizei, das in mehreren deutschen Bundesländern als Entscheidungshilfe für eine Zulassung zur Ausbildung für den höheren Polizeivollzugsdienst durchgeführt wird. Der Autor diskutiert ausführlich die bestehenden Befunde zur Güte des Assessment Centers und geht dabei auf wichtige (statistische) Randbedingungen ein, die bei einer AC-Evaluation berücksichtigt werden müssen. Kritisch beleuchtet er den (Zusatz-) Nutzen von ACs für die Prognose von Ausbildungs- und Berufserfolg gegenüber anderen eignungsdiagnostischen Verfahren.

Umsetzung der Standards

Standard 1:
Zwar wird gesagt, dass das geschilderte AC Bestandteil der Personalentwicklung der Behörde ist, dies wird jedoch nicht weiter ausgeführt bzw. detailliert beschrieben.

Standard 2:
Unter „Unternehmensziel" kann im vorliegenden Fall verstanden werden, dass die zentralen Verantwortungs- und Führungspositionen bei der Polizei der betreffenden Bundesländer nur von Personen eingenommen werden sollen, die über ein nachweislich vorhandenes und zumindest zufriedenstellendes Potenzial an sozialen und intellektuellen Kompetenzen verfügen. Die Umsetzung dieses Ziels wird durch ein Assessment Center für die Auswahl zur Aufstiegsausbildung und in der Folge zum höheren Polizeivollzugsdienst sichergestellt.

Standard 3:
Aus der Zulassungsquote zum höheren Dienst von lediglich zwei Prozent aller Beamten lässt sich unter Berücksichtigung von Sonderfällen und Ausnahmen prinzipiell der absolute Personalbedarf und Mittelbedarf im Sinn der Anzahl durchzuführender ACs ableiten.

Standard 4:
In nahezu mustergültiger Weise wird die Gültigkeit/Tauglichkeit des angewandten ACs zur Vorhersage des Ausbildungserfolgs an der Polizei-Führungsakademie empirisch belegt (kriteriumsbezogene Validität). Dies ist insbesondere auch deshalb bemerkenswert, weil es sich hierbei um eine der immer noch raren Beispiele für eine eigene Bewährungskontrolle handelt. Es wird die tatsächliche Leistungsfähigkeit des eingesetzten ACs ermittelt, anstatt sich mit generellen Hinweisen auf die Validität von AC-Verfahren zu begnügen.

Standard 5:
Konkrete Entwicklungsmaßnahmen werden nicht geschildert.

Standard 6:
Auf die konkrete Umsetzungsverantwortung im Hinblick auf vereinbarte Entwicklungsmaßnahmen wird nicht eingegangen.

Standard 7:
Aufgrund der Tatsache, dass das hier vorgestellte Verfahren durch eine spezialisierte und erfahrene Personalberatungsgesellschaft federführend durchgeführt wird, kann das Vorhandensein der erforderlichen Fach- und Methodenkompetenz wohl zu Recht vorausgesetzt werden.

Standard 8:
Es wird ausgiebig dargelegt und veranschaulicht, dass sämtliche AC-Beobachter („Assessoren") vor ihrer erstmaligen Mitwirkung in einer Beobachterschulung systematisch auf die Vorgehensweise und ihre Rolle im AC vorbereitet werden.

Standard 9:
Es existieren sowohl definierte Kriterien für die Zulassung zum AC als auch verbindliche Erhebungs- und Verrechnungsregeln während des Verfahrens selbst, so dass Transparenz unterstellt werden kann.

Standard 10:
Kersting unterstreicht in seinem Beitrag, wie unverzichtbar Erfolgskontrollen für PE-Maßnahmen sind. Die von ihm vorgestellten Kreuztabellen zeigen, inwieweit sich die durchführende Organisation darauf verlassen kann, dass Bewerber, die den AC-Anforderungen genügen konnten, auch die späteren Ausbildungs- und Berufsanforderungen erfüllen.

Ein Hinweis der Herausgeber in eigener Sache:

Kersting vertritt in seinem Beitrag die These, dass Assessment Center „zur Vorhersage beruflicher Leistungen gegenüber Intelligenztests praktisch unbedeutend und unter diesem Gesichtspunkt gegenüber Intelligenztests nutzlos" seien. Er zieht hierfür eine Vielzahl von wissenschaftlichen Studien heran.

Wir teilen diese Ansicht nicht, da sich u.E. die wissenschaftliche Forschung zum AC zu monokausalistisch auf einfach messbare Kriterien bezieht. Wichtige Praxisanforderungen und -kriterien, wie sie beispielsweise auch der Arbeitskreis Assessment Center formuliert hat, werden dabei übersehen. Wir haben uns aus zwei Gründen dennoch für die Aufnahme des Beitrags entschieden:

1. *Die Streitfrage zur kriteriumsbezogenen Validität des Assessment Centers ist nur ein Nebenaspekt, bezogen auf die eigentliche Fragestellung des Bandes (Umsetzung von PE-Standards in Praxisprojekten). Dass die geschilderte AC-Realisierung den definierten Qualitätskriterien entspricht, dürfte außer Frage stehen.*
2. *Kerstings Beitrag deutet eine größer werdende Kluft zwischen Wissenschaft und Praxis bei der Bewertung von Personalverfahren an. Der Beitrag kann hoffentlich dabei helfen, die Divergenz offen zu legen, und einen Anstoß zur notwendigen Diskussion geben.*

Der Beitrag thematisiert am Beispiel der Auswahl von Führungskräften der Polizei die Frage der Treffsicherheit von Assessment Centern. Vor der Beschreibung des Auswahlinstruments skizziert der Autor kurz einige relevante organisationale Rahmenbedingungen.

Organisationaler Hintergrund

Der Polizeivollzugsdienst gliedert sich in die mittlere, gehobene und höhere Laufbahn, wobei nur ein äußerst geringer Anteil (ca. 2%) der Beamten in den höheren Dienst gelangt. Die Beamten des höheren Dienstes werden überwiegend aus Mitarbeitern des gehobenen Dienstes rekrutiert, die Auswahl folgt dem Prinzip der Bestenauslese. Zu Beginn der Laufbahn des höheren Dienstes steht in der Regel eine zweijährige Ausbildung, ein Jahr davon als Studium an der Polizei-Führungsakademie. Mit der Auswahl wird somit zunächst über die Zulassung zur Aufstiegsausbildung entschieden, aber letztendlich ist die Entscheidung auch eine Voraussetzung für die Zulassung zur Polizeiführung. Angehörige des höheren Polizeivollzugsdienstes führen mehrere hundert bis mehrere tausend unterstellte Mitarbeiter und/oder tragen die Verantwortung für die Bewältigung materiell, personell und/oder gesellschaftlich brisanter Lagen. Zusätzliche Brisanz gewinnt die Führungskräfteauswahl durch die Tatsache, dass es sich um eine Entscheidung über organisationsinterne Mitarbeiter handelt. Bei der Gestaltung interner Personalentscheidungsverfahren ist u.a. dafür Sorge zu tragen, dass dauerhaft möglichst viele potenziell geeignete Mitarbeiter bereit sind, sich dem Verfahren zu unterziehen. Das Vorgehen sollte diejenigen, die nicht berücksichtigt werden können, so wenig wie möglich demotivieren, also keine „Verliererproblematik" erzeugen. Das Verfahren der Personalentscheidung ist Ausdruck der Führungskultur einer Organisation. Die Relevanz der in Frage stehenden Positionen und die zuletzt genannten Akzeptanzgesichtspunkte (siehe Kersting, 1998; Schuler, 1990) erfordern eine sorgfältige Planung, Durchführung und Evaluation der Verfahren.

Implementierung

Das hier thematisierte Assessment Center für die Polizei verschiedener deutscher Bundesländer ist integrativer Bestandteil der jeweiligen Personalentwicklungskonzepte. Die Zulassung zum Verfahren ist formal geregelt. Die Personalvertretung hat den Auswahl- und Zulassungsrichtlinien zugestimmt. Wesentliche Elemente der Implementierung, z.B. die Bedarfsanalyse, die Organisationsanalyse, die Zulassung zum Verfahren sowie die Förderungsmaßnahmen, werden durch die Polizei selbst realisiert. Lediglich bei der Gestaltung, Durchführung, Auswertung und Interpretation des Assessment Centers bedienen sich einige Länder der Unterstützung eines Beratungsunternehmens.

Der vorliegende Bericht konzentriert sich auf das Assessment Center. Dieses ist ein Bestandteil der Personalentwicklung, auf die übrigen Instrumente der Personalentwicklung wird hier nicht eingegangen.

Gestaltung des Assessment Centers

Konzeption: Ausgangspunkt Anforderungsanalyse

Konzeptioneller Ausgangspunkt von Personalentscheidungen ist das aus der Arbeits- und Anforderungsanalyse abgeleitete Anforderungsprofil. Zur Realisierung von Methodenvielfalt wurden bei der Erarbeitung des hier zugrunde gelegten Anforderungsprofils für den höheren Polizeivollzugsdienst drei verschiedene Ansätze genutzt, nämlich die arbeitsplatzanalytisch-empirische, die personbezogen-empirische und die erfahrungsgeleitet-intuitive Methode (siehe Schuler, 1988, S. 202 f.). Ergebnisse zur arbeitsplatzanalytischen Methode lagen aus der Studie von Althoff (1974) vor, bei der die deutsche Bearbeitung des „Position Analysis Questionnaire" (Frieling und Hoyos, 1978) zur Untersuchung von Polizeiarbeitsplätzen eingesetzt wurde. Dieser verhaltensorientierte Ansatz wurde durch den Rückgriff auf die Ergebnisse mehrerer polizeispezifischer Bewährungskontrollen (Althoff, 1977; Greif, 1972; Jäger und Althoff, 1994) um einen personenzentrierten Ansatz ergänzt. Auf der Grundlage der Studien wurden die Merkmale erfolgreicher Polizisten ermittelt. Diese Merkmale wurden dann als Anforderungsmerkmale zukünftiger Polizisten betrachtet. Die Nutzung der Ergebnisse der beiden genannten Ansätze war dadurch eingeschränkt, dass die genannten Studien sich überwiegend auf den gehobenen Polizeivollzugsdienst beziehen. Allerdings ist es plausibel, dass der höhere Polizeivollzugsdienst sich hinsichtlich der Anforderungen vom gehobenen Polizeivollzugsdienst eher graduell als grundsätzlich unterscheidet. Mit der erfahrungsgeleitet-intuitiven Methode konnte diese Annahme überprüft werden. Das mit Hilfe der beiden zunächst genannten Methoden von psychologischen Experten erstellte Anforderungsprofil für den höheren Polizeivollzugsdienst wurde in Workshops polizeiinternen Experten zur Begutachtung und Modifikation vorgelegt. Bei der Festlegung des Anforderungsprofils war vor allem zu berücksichtigen, dass die im höheren Polizeivollzugsdienst anfallenden Arbeitsaufgaben einerseits sehr variabel (z.B. Stabs-, Führungs- und Lehrfunktionen) und andererseits über die Zeit hinweg sehr änderungsintensiv sind, so dass sich die Anforderungsanalyse auf globale Eig-

nungsmerkmale beschränken musste. Das resultierende Anforderungsprofil sah insgesamt die folgenden neun Dimensionen vor:

1. Kommunikationsfähigkeit
2. Soziale Kompetenz
3. Sicherheit/Belastbarkeit
4. Zielorientierung und Sachbezogenheit
5. Aktivität und Dynamik
6. Einfallsreichtum/Flexibilität
7. Führungsverhalten
8. Motivation
9. intellektuelle Fähigkeiten und Kenntnisse.

Der Nachweis der notwendigen Fachkenntnisse wird im Rahmen der Zulassung zum Auswahlverfahren erbracht, so dass die Fachkenntnisse nicht im Assessment Center geprüft werden müssen. Für die Anforderungsdimension wurden verhaltensnahe Operationalisierungen erarbeitet und Beobachtungschecklisten erstellt. Verhaltensanker für die „soziale Kompetenz" in Gruppensituationen sind beispielsweise die Beobachtungen: „geht von sich aus auf andere zu", „bezieht die Gruppe ein", „nimmt Anregungen anderer auf".

Planung: Auswahl bzw. Konstruktion der Aufgaben, Festlegung der Vorgehensweise

Im Rahmen der Planung wurden Aufgaben zur Erfassung der im Anforderungsprofil aufgelisteten Eignungsmerkmale ausgewählt oder konstruiert. Um die Zuverlässigkeit der Messung zu gewährleisten, sollte dabei jedes Eignungsmerkmal nach Möglichkeit mehrfach und mit unterschiedlichen Herangehensweisen erfasst werden. Hinsichtlich der Diagnose der intellektuellen Leistungen und der Kenntnisse fiel die Wahl auf einen schriftlichen psychometrischen Test, die übrigen Eignungsmerkmale sollten mit situativen Verfahren erfasst werden. Die Verfahren werden weiter unten beschrieben.

Vorinformation der Teilnehmer

Alle Teilnehmer erhalten vor dem Verfahren eine 39 DIN-A5 Seiten umfassende Broschüre mit allgemeinen Informationen über Personalentscheidungsverfahren sowie einen beispielhaften Ablaufplan des schriftlichen und mündlichen Verfahrens. Der Großteil der Broschüre ist Beispielaufgaben zum schriftlichen Test vorbehalten und regt die Teilnehmer zur Übung an. Die Broschüre informiert abschließend darüber, welche Rechte Teilnehmer an Personalentscheidungsverfahren haben und wie sie ihre Rechte wahren können. Zusätzlich werden die Teilnehmer vorab über organisatorische Aspekte des Verfahrens (Dauer, Ort, Unterbringungsmöglichkeiten, mitwirkende Personen usw.) informiert. Die Bedeutung des Assessment Centers ergibt sich für die Bewerber aus den Auswahlrichtlinien der Polizei.

Vorbereitung der Assessoren

Jedes Assessment Center-Kommissionsmitglied der Polizei wird hinsichtlich der Aspekte „Grundlagen der Beobachtung und Beurteilung" sowie „Beurteilungsstandards" und „Sensibilisierung für Beurteilungsfehler" trainiert. Das vorab versendete Anforderungsprofil und die Vorgehensweise werden erläutert. Die Polizei-Assessoren weisen ein hohes Maß an Erfahrung im Personalwesen auf und nehmen nicht nur einmal, sondern über mehrere Jahre hinweg wiederholt die Tätigkeit eines Kommissionsmitgliedes im Assessment Center wahr.

Eingesetzte Verfahren

Schriftlicher Test

Zur Diagnose von drei verschiedenen Dimensionen intellektueller Fähigkeiten sowie der allgemeinen Kenntnisse wird ein psychometrischer Eignungstest eingesetzt. Mit mehreren Aufgabengruppen wird die Fähigkeit zum gedanklichen Umgang mit (1) verbalem und (2) numerischem Material geprüft. Mit zwei verschiedenen, nach Art einer Arbeitsprobe konstruierten, weiteren Aufgabengruppen wird (3) die Effizienz bei der Bearbeitung von Routinetätigkeiten diagnostiziert. Darüber hinaus kommen (4) Aufgabengruppen mit Kenntnisfragen zu verschiedenen Wissensdomänen (z.B. Wirtschaft) zum Einsatz. Statistische Angaben zur Qualität einiger der genannten Aufgabengruppen sind bei Kersting (1999, S. 182) berichtet, Hinweise zur faktoriellen Struktur ergeben sich aus den bei Kersting (1994, S. 53) dargestellten Analysen. Bei dem Test handelt es sich um ein von der Beratungsorganisation entwickeltes Verfahren, welches nicht im Handel erhältlich ist, sondern exklusiv für ihre Kunden zur Verfügung gestellt wird, um dadurch der Verbreitung der Testaufgaben vorzubeugen. Einzelne Aufgaben des Tests sind später in den „BIS-r-DGP" Test (Kersting & Beauducel, 2001) aufgenommen worden.

Situative Verfahren

Zur Operationaliserung wurde den zunächst auf der Ebene von Fähigkeitsbegriffen beschriebenen Dimensionen des Anforderungsprofils eine Liste von möglichen konkreten Einzelbeobachtungen zugeordnet. Anschließend wurden Übungen konstruiert, mit denen das interessierende Verhalten provoziert werden kann. Letztendlich wurden die im Folgenden skizzierten vier Assessment Center-Übungen zur Erfassung der Eignungsmerkmale vorgesehen.

- Einzelvorstellung (EV)/eignungsdiagnostisches Interview
  In der Einzelvorstellung sprechen die Assessoren jeweils fünfundzwanzig Minuten zielgerichtet mit jedem Bewerber einzeln über seinen bisherigen beruflichen Werdegang und seine Aufstiegsmotivation. Thematisiert werden insbesondere die bisherigen Führungserfahrungen. Das Interview erfolgt in halbstandardisierter Form und nutzt sowohl situative als auch biographisch orientierte Fragen.

- Kurzreferat (KR)
  Jeder Bewerber bereitet sich individuell auf ein Referat zu einem vorgegebenen Thema vor. Als Ausgangslage erhalten alle die gleichen Informationsmaterialien. Nach einer für alle Teilnehmer auf exakt vierzig Minuten festgelegten Vorbereitungszeit präsentieren die Bewerber der Kommission einzeln einen fünfminütigen Vortrag.

- Gruppendiskussion (GD)
  Eine Gruppe von vier bis maximal sieben Bewerbern diskutiert nacheinander zwei vorgegebene Themen, von denen eines allgemeiner Art und eines berufseinschlägig ist. Pro Diskussionsrunde stehen ca. fünfzehn Minuten zur Verfügung. Anschließend wird der Gruppe eine Aufgabe gestellt, für die sie innerhalb von zwanzig Minuten gemeinsam einen konkreten Lösungsvorschlag erarbeiten soll. So ist z.B. ein Trainingsprogramm zur Vorbereitung eines genau definierten Polizeieinsatzes zu konzipieren.

- Presseschau / Problemlöseszenario (PR)
  Dem Großteil der in der weiter unten dargestellten Analyse berücksichtigten Personen wurde die Aufgabe gestellt, in einer Gruppe von vier bis maximal sieben Bewerbern auf der Grundlage zur Verfügung gestellter aktueller Tageszeitungen eine Presseschau vorzubereiten. Die Bewerber erhielten zunächst ein Zeitbudget für ein individuelles Quellenstudium, für die anschließende gemeinsame Gestaltung der Presseschau standen 50 Minuten zur Verfügung. Im Laufe des Berichtszeitraums wurde die Übung „Presseschau" durch die Übung „Problemlöseszenario" ersetzt. Hierbei übernimmt eine Bewerbergruppe (vier bis maximal sieben Personen) für zwölf Entscheidungstakte die Führung einer kleinen, am Computer simulierten Fabrik. Für die gemeinschaftliche Steuerung des Problemlöseszenarios stehen den Bewerbern 40 Minuten zur Verfügung.

Mit Ausnahme der Motivation können alle Eignungsmerkmale in mehreren Übungen beobachtet werden. Die intellektuellen Fähigkeiten werden mit schriftlichen Leistungstests überprüft, an die Stelle mehrerer Übungen treten hier mehrere Aufgabentypen. Aus der in Tabelle 5 dargestellten Anforderungs-Übungs-Matrix ist ersichtlich, welches Eignungsmerkmal in welcher Übung beobachtet und beurteilt wird. (Sollte ein Eignungsmerkmal in einer Übung nicht beurteilt werden, ist die entsprechende Zelle in der Darstellung grau hinterlegt).

Beurteilungsprozess und Beurteilungsskala

Sowohl die Ergebnisse des schriftlichen Leistungstests als auch die letztendlichen Beurteilungen des Verhaltens im mündlichen Teil des Assessment Centers werden auf einer vorab definierten Skala beurteilt und festgehalten. Die Skala reicht von dem negativen Wert „1" bis zum positiven Wert „5". Zwischenstufen sind die Werte „1 plus", „2 minus", „2", „2 plus", „3 minus", „3"

	Einzelvorstellung **EV**	Gruppendiskussion **GD**	Presseschau/ Szenario **PR**	Kurzreferat **KR**	**Dimensionsurteil** (Mittelwert)
Kommunikationsfähigkeit					
Soziale Kompetenz					
Sicherheit / Belastbarkeit					
Zielorientierung u. Sachbe.					
Aktivität und Dynamik					
Einfallsreichtum/ Flexibilität					
Führungsverhalten					
Motivation					
Verhaltensurteil (Mittelwert der acht Dimensionsurteile)					
Übungsurteil (Mittelwert)					
Gesamtempfehlung:	(Testurteil + Verhaltensurteil) / 2				

Tab. 5: Anforderungs-Übungs-Matrix des Assessment Centers

usw. bis „5 minus", so dass sich insgesamt dreizehn Skalenstufen ergeben. (Bei einem kleinen Teil der im Folgenden analysierten Gruppe kam eine geringfügig modifizierte Skala zum Einsatz. Diese Ergebnisse wurden für die Analyse in die dargestellte Skala transformiert.)

Zur Bedeutung der einzelnen Skalenwerte: Eine gute Prognose liegt für Bewerber mit einem Empfehlungsgrad von „5", „den Anforderungen voll entsprechend", vor. Die Empfehlungsgradstufe „4" („den Anforderungen weitgehend entsprechend") signalisiert leichte Einschränkungen in Bezug auf den zu erwartenden Ausbildungs- und Berufserfolg. Eine hohes Risiko ergibt sich bei der Empfehlungsgradstufe „3", „den Anforderungen nur teilweise entsprechend". Bei mit „3" beurteilten Bewerbern bestehen Bedenken gegenüber einer Zulassung. Die „Dreier-Kandidaten" sind nicht durchgängig leistungsschwach (das sind die Bewerber mit den Empfehlungsgraden „2" und „1"), weisen aber neben Stärken auch Schwächen auf. Eine Prognose ist hier schwierig, es sind „Risikokandidaten".

Schriftlicher Test

Die als „Empfehlungsgrad Test" bezeichnete zusammenfassende Wertung aller schriftlichen Einzelleistungen wird von Diplom-Psychologen vor dem Hintergrund eines Anforderungsprofils vorgenommen (so genannte „klinische Urteilsbildung"). Zusätzlich wird in der hier vorgestellten Bewährungskontrolle noch ein post festum statistisch gebildeter „Testmittelwert" berücksichtigt.

Dabei handelt es sich um den Mittelwert der in den oben genannten vier geprüften Dimensionen erzielten Testergebnisse. Der Testmittelwert geht auf die Leistungen in den Einzeltests zurück, die auf der z-Skala skaliert sind.

Situative Verfahren
Das Verhalten der Bewerber wird von einer Kommission beobachtet und beurteilt. Die Beurteilungskommission setzt sich aus zwei Vertretern der einstellenden Organisation sowie zwei externen Diplom-Psychologen zusammen. Die Kommissionsmitglieder beteiligen sich in keiner Übung an der Diskussion der Teilnehmer oder an der Aufgabenlösung, lediglich während der Einzelvorstellung greifen die Kommissionsmitglieder mit Fragen aktiv in das Geschehen ein.

Nach jeder Übung wird das Verhalten der Kandidaten in den für diese Übung beobachtungsrelevanten Dimensionen (siehe Tab. 1, S. 24) durch die Kommissionsmitglieder bewertet. Dabei liegt für jede Übung ein spezifischer Beurteilungsbogen mit vorgegebenen Beurteilungskategorien vor. Die Teilnehmer werden zunächst von jedem Kommissionsmitglied unabhängig beobachtet und durch den Eintrag eines Skalenwerts pro Verhaltensdimension beurteilt. Bei ihrem Urteil sollen sich die Kommissionsmitglieder auf konkrete Verhaltensbeobachtungen stützen, Beobachtung und Beurteilung sollen möglichst voneinander getrennt werden. Im Anschluss an jede Übung haben die Kommissionsmitglieder dann Zeit, sich über ihre Beobachtungen zum Ablauf des Geschehens und zum Verhalten der Teilnehmer auszutauschen. Die Kommissionsmitglieder diskutieren ihre aus den Beobachtungen gezogenen Schlussfolgerungen (die Einzelbeurteilungen pro Verhaltensdimension) und legen für jede zu beurteilende Verhaltensdimension ein gemeinsames Urteil fest. Dabei wird besonderes Augenmerk darauf gelegt, dass nur an konkreten Verhaltensweisen festgemachte Urteile abschließend berücksichtigt werden. In der Dokumentation wird nicht nur das Konsensurteil, sondern auch das Einzelurteil der Assessoren (vor Diskussion) festgehalten.

Zusätzlich zu den Einzelurteilen (pro Dimension und Übung) wird zum Ende des Assessment Centers ein abschließendes Dimensionsurteil gebildet. Dieses ist der Mittelwert aller Beurteilungen für ein und dieselbe Verhaltensdimension (z.B. soziale Kompetenz). Als Übungsurteil wird der Mittelwert aller Beurteilungen für ein und dieselbe Übung (z.B. Gruppendiskussion) festgelegt. Der mündliche Teil des Assessment Centers endet mit einem Verhaltensurteil („Empfehlungsgrad Verhalten", Mittelwert der acht Dimensionsurteile).

Abschließendes Verfahrensurteil („Gesamtempfehlungsgrad")
Grundlage der Personalentscheidung und der im vorliegenden Beitrag vorgenommenen Qualitätskontrolle ist das abschließende Verfahrensurteil („Gesamtempfehlungsgrad"). Dieser Wert wird auf der Basis der Testleistungen einerseits und des Urteils über den mündlichen Teils des Assessment Centers andererseits gebildet. Im vorliegenden Bericht ist das abschließende

Verfahrensurteil der Mittelwert aus den beiden Empfehlungsgraden der Test- und Verhaltensbeurteilung. Der Mittelwert wird durch Auf- oder Abrundung auf die nächstliegende Stufe der weiter oben dargestellten Skala transformiert. Liegt der Mittelwert exakt zwischen zwei Skalenstufen, so gibt die Verhaltensbeurteilung den Ausschlag.

Ablauf des Assessment Centers

Das Assessment Center ist zweitägig angelegt, die Bewerber erscheinen in Gruppen von maximal zehn Personen. Zu Beginn des Verfahrens werden die Teilnehmer über den Ablauf des Verfahrens genauer informiert, die beteiligten Personen stellen sich den Bewerbern vor, und es besteht Gelegenheit, Fragen und organisatorische Dinge zu klären. Jeder Teilnehmer erhält einen individuellen Ablaufplan mit präzisen Angaben über Beginn und Dauer der einzelnen Übungen. Die schriftlichen Tests und der mündliche Teil des Verfahrens laufen teilweise parallel. Auf diese Art und Weise werden belastende Wartezeiten für die Teilnehmer minimiert und die Informationsausschöpfung wird maximiert. Das Verfahren sieht ausreichende Pausen vor und gewährleistet den Teilnehmern zu jeder Zeit Transparenz über den Ablauf und über die gestellten Anforderungen. Am zweiten Tag führt jeder Teilnehmer ein Feedback-Gespräch mit einem Diplom-Psychologen. In diesem Gespräch werden Selbst und Fremdbild verglichen, realisierbare Verhaltensänderungen und persönliche Entwicklungsmöglichkeiten angesprochen und offen gebliebene Fragen geklärt. Außerdem wird jeder Teilnehmer angeregt, das Verfahren und das Verhalten der Kommissionsmitglieder aus seiner Sicht zu werten und Vorschläge für Verbesserungen zu unterbreiten.

Bewertung und Umsetzung der Ergebnisse

Das Assessment Center endet am zweiten Tag mit der gemeinsamen Würdigung des erlebten und protokollierten Verhaltens der einzelnen Bewerber durch die Kommissionsmitglieder. Die einzelnen Urteile werden zu den abschließenden Übungs- und Dimensionsurteilen verrechnet, aufgrund der Dimensionsurteile wird der „Empfehlungsgrad Verhalten" bestimmt (siehe oben). Die Diplom-Psychologen erläutern die Testergebnisse und geben den „Empfehlungsgrad Test" bekannt. Aus den Empfehlungsgraden für den mündlichen und den schriftlichen Teil ergibt sich der abschließende „Gesamtempfehlungsgrad". Die wesentlichen Beobachtungen und Beurteilungen werden außerdem von den Diplom-Psychologen in einer sehr kurzen schriftlichen Stellungnahme zusammengefasst. Vor dem Hintergrund der Befunde denkt die Kommission gemeinsam über Angebote zur Gestaltung und Förderung der beruflichen Entwicklung der einzelnen Kandidaten nach.

Die letzte Stufe des Verfahrens besteht in einem organisationsinternen Zulassungsgespräch unter Beteiligung der höchsten Ebenen der Polizeiführung.

Dieses Gespräch steht nur noch Bewerbern offen, die im Assessment Center eine definierte Mindestwertungsstufe erzielt haben. Hier findet in der Regel keine weitere Selektion mehr statt.

Evaluation des Verfahrens

Die Frage der Qualität der Personalentscheidung hängt maßgeblich von der Validität (Gültigkeit) ab. Um Informationen zur Validität einzelner Instrumente der Personalentscheidung zu erhalten, kann man einerseits auf Erkenntnisse vorhandener Studien zurückgreifen und andererseits eigene Bewährungskontrollen anstellen.

Nutzung vorhandener Erkenntnisse zur Abschätzung der Validität

Die Treffsicherheit von Personalentscheidungsverfahren wird seit fast einem Jahrhundert immer wieder systematisch untersucht. Seit Mitte der siebziger Jahre ist es mit Hilfe einer bestimmten statistischen Verfahrensgruppe – der so genannten „Metaanalyse" – möglich, Ergebnisse aus verschiedenen Studien auf quantitative Art und Weise zusammenzufassen. Bei bestimmten metaanalytischen Methoden können einzelne Fehlerquellen, wie z.B. geringe Stichprobengrößen, Messfehler, Unreliabilität und Streuungseinschränkungen bei Prädiktoren und Kriterien, kontrolliert und Moderatorvariablen der Validität identifiziert werden. Metaanalytische Befunde stützen sich auf eine Datenbasis, die in zahlreichen unabhängigen Studien an zehntausenden Personen gesammelt wurde. Eine aktuelle Übersicht über metaanalytisch gewonnene Erkenntnisse zur Vorhersagekraft verschiedener Instrumente der Personalentscheidung präsentieren Schmidt und Hunter (1998). Dieser Bericht liegt auch in deutscher Übersetzung vor (siehe Kleinmann & Strauß, 1998). Stand der Dinge ist demnach, dass professionelle Verfahren der Personalentscheidung – u.a. Intelligenztests und Assessment Center – eine gute Vorhersage beruflicher Leistungen ermöglichen. In den beiden folgenden Abschnitten werden einige ausgewählte allgemeine Erkenntnisse zur Kriteriumsvalidität von Intelligenztests einerseits und Assessment Centern andererseits skizziert. Da im weiteren Verlauf des Artikels eine Bewährungskontrolle für den Bereich „Polizei" dargestellt wird, soll bei den folgenden Ausführungen die Aussagekraft der Verfahren im Allgemeinen und die Vorhersage polizeispezifischer Leistungen im Besonderen Berücksichtigung finden.

Zur Kriteriumsvalidität von Intelligenztests

Nach der Sichtung metaanalytischer Studien resümieren Schmidt und Hunter (1998), dass aufgrund von Intelligenztests gewonnene Aussagen mit die höchste Validität bei der Vorhersage zukünftiger Leistungen erzielen, und zwar sowohl bei der Vorhersage von Ausbildungs- als auch bei der Vorhersage von Berufsleistungen. Hunter und Hunter (1984) berichten entsprechende gemittelte Koeffizienten von r=.54 für den Ausbildungs- und r=.45 für den

Berufserfolg. Nominell geringfügig bessere Validitätswerte werden demzufolge nur beim Einsatz von Arbeitsproben erzielt, die allerdings deutlich teurer sind und deren Anwendbarkeit häufig auf Bewerber mit Berufskenntnissen beschränkt bleibt. Diese Werte ergaben sich als Mittelung über alle analysierten Ausbildungsgänge und Berufsbilder. Polizeispezifische Befunde sind seltener, offensichtlich besteht hier berufsspezifischer Nachholbedarf (siehe Campbell, zitiert nach Hirsh, Northrop & Schmidt, 1986, S. 400). Einen Großteil der vorliegenden Ergebnisse der (überwiegend nordamerikanischen) Bewährungskontrollen aus dem Bereich „law enforcement" fassten Hirsch et al. (1986) metaanalytisch zusammen. Für unterschiedliche kognitive Testverfahren ergab sich über verschiedene Studien mit insgesamt 12.897 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=14.991).

Zur Erklärung der vergleichsweise geringen Vorhersagbarkeit des polizeilichen Berufserfolgs (siehe bereits Ghiselli 1973, S. 471f.) diskutieren Hirsch et al. (1986) 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. Ein Nachweis der Prognostizierbarkeit des polizeilichen Berufserfolgs durch Intelligenztests sowie durch Problemlöseszenarien wurde in jüngerer Zeit im deutschsprachigen Raum von Kersting (1999, 2001) erbracht.

Zur Kriteriumsvalidität von Assessment Centern

Hinsichtlich der Treffsicherheit der aufgrund von Assessment Centern gewonnenen Eignungsaussagen kann u.a. die Arbeit von Thornton, Gaugler, Rosental und Bentson (1987) zitiert werden, derzufolge die entsprechende korrigierte mittlere Kriteriumsvalidität bei r=.37 liegt. In dieser Metaanalyse wurden fünfzig einzelne Bewährungskontrollen zu Assessment Centern berücksichtigt. Einen polizeispezifischen Wert berichtet z.B. Chan (1996). In seiner Studie mit 46 Polizisten der Singapore Police Force konnte mit einem Assessment Center die zukünftige berufliche Karriere (r=.59), nicht aber das Vorgesetztenurteil (r=.06) vorhergesagt werden.

Es gilt insgesamt als relativ gut gesichert, dass sorgfältig konstruierte Assessment Center eine Vorhersage beruflicher Leistungen erlauben. Qualitätsmerkmale von Assessment Centern sind der Analyse von Thornton et al. (1992) zufolge eine große Anzahl verschiedener Übungen sowie der Einsatz professioneller Beurteiler (Thornton et al. nennen hier u.a. Psychologen), die mit internen Kennern der Organisation zusammenarbeiten sollten. In der Studie

von Sagie und Magnezy (1997) wirkte sich die Beteiligung von Psychologen positiv auf die Konstruktvalidität aus.

Wohlkonstruierte und professionell durchgeführte Assessment Center können unter rein technischen Gesichtspunkten der Vorhersageeffizienz als funktionstüchtige Instrumente der Personalentscheidung bezeichnet werden. Das Fazit zur Kriteriumsvalidität von Assessment Centern fällt allerdings ungünstiger aus, wenn man die Gültigkeit von Assessment Centern im Vergleich zu alternativen Instrumenten betrachtet.

Vor dem Hintergrund der Leistungsfähigkeit anderer Instrumente der Personalentscheidung fällt die Bilanz zur Kriteriumsvalidität von Assessment Centern eher bescheiden aus und stellt vor allem die Eigenständigkeit der Vorhersagekraft von Assessment Centern in Frage. Der Nachweis der Kriteriumsvalidität allein ist lediglich eine notwendige, nicht aber eine hinreichende Begründung für einen Einsatz von Assessment Centern im Kontext von Personalentscheidungen. Zusätzlich zur Kriteriumsvalidität muss auch die „Nützlichkeit" des Verfahrens geklärt werden. Ein Verfahren ist nach Lienert und Raatz (1994, S. 13) erst dann nützlich, wenn es in seiner Funktion durch kein anderes Verfahren vertreten werden kann, d.h. wenn es gegenüber alternativen Verfahren einen eigenständigen, inkrementellen Beitrag zur Validität leistet. Exakt dies leisten Assessment Center den Ergebnissen von Schmidt und Hunter (1998) zufolge nicht. Setzt man also zusätzlich zu einem Intelligenztest auch noch situative Assessment Center-Übungen ein, so wird nach Schmidt und Hunter (1998) durch die Kombination der Verfahren keine nennenswert bessere Vorhersage beruflicher Leistungen erzielt als durch den Einsatz des Intelligenztests allein. Assessment Center sind nach dem Stand der jüngeren Forschungsergebnisse zur Vorhersage beruflicher Leistungen gegenüber Intelligenztests praktisch unbedeutend und unter diesem Gesichtspunkt gegenüber Intelligenztests nutzlos. Als eine Ursache für die mangelnde inkrementelle Kriteriumsvalidität von Assessment Centern wird deren hohe Korrelation mit Intelligenztestleistungen angesehen. Schmidt und Hunter (1998, S. 11) schätzen die Korrelation zwischen Assessment Centern und Intelligenztests unter Berufung auf Collins auf r=.50. Einer Metaanalyse von Scholz und Schuler (1993) zufolge korreliert die allgemeine Intelligenz zu r=.33 mit dem Abschneiden im Assessment Center. Dies kann u.a. dadurch bedingt sein, dass Intelligenztests häufig ein Bestandteil von Assessment Centern sind.

Eine andere mögliche Erklärung für die im Vergleich zu Intelligenztests defizitäre Kriteriumsvalidität von Assessment Centern rekurriert auf die Kriterien, an denen die Treffsicherheit bestimmt wird. Es ist evident, dass Intelligenz ein exzellenter Prädiktor für alle Arten von Prüfungsleistungen ist. Intelligenztests werden daher eingesetzt, wenn es um die Vorhersage von Ausbildungs- und Studienerfolg geht. Bei der Vorhersage beruflicher Leistungen sollten demgegenüber erwartungsgemäß Verfahren, die ein breiteres Spektrum von Fähigkeiten diagnostizieren und auch die „soft skills" berücksichtigen, überlegen sein. Dies ist aber nach Schmidt und Hunter (1998) dem

Stand der Forschung zufolge nicht der Fall. Intelligenztests sind Assessment Centern auch bei der Vorhersage beruflicher Leistungen überlegen. Dabei können Intelligenztests nicht nur – wie man früher gedacht hat – den kurzfristigen, sondern auch den langfristigen Berufserfolg treffsicher prognostizieren (siehe Hossiep, 1995), die Prädiktionskraft der Tests steigt langfristig sogar tendenziell an.

Eine Ursache für das Ausbleiben der erwarteten prädiktiven Überlegenheit von situativen Assessment Center-Übungen könnte der Umstand sein, dass bislang nicht die „richtigen" Kriterien für die Beurteilung des Berufserfolgs genutzt wurden. Möglicherweise sind z.B. 360-Grad-Beurteilungen ein besserer Indikator für den Berufserfolg als die üblicherweise genutzten Vorgesetztenbeurteilungen. Der Großteil der empirischen Untersuchungen stammt zudem aus Nordamerika, die Übertragbarkeit auf europäische Verhältnisse kann nicht ohne weiteres vorausgesetzt werden. In Studien, in denen der Berufserfolg anhand geeigneterer Kriterien bestimmt wird, könnte sich – so die Annahme – die Überlegenheit von Assessment Center-Übungen zeigen. Bislang ist dies aber nur eine Hoffnung, und auch eine begründete Spekulation kann und darf den zu erbringenden Beleg nicht ersetzen.

Der Umstand, dass der Übersicht von Schmidt und Hunter (1998) zufolge (1) Assessment Center gegenüber Intelligenztests keine inkrementelle Kriteriumsvalidität aufweisen und dass (2) mit einem Assessment Center insgesamt im Durchschnitt eine nominell geringere Kriteriumsvalidität erzielt wird als beim isolierten Einsatz einzelner Assessment Center-Bestandteile (z.B. strukturiertes Vorstellungsinterview), indiziert deutlich Forschungs- und Handlungsbedarf. Bei dem jetzigen Erkenntnisstand ist zu fordern, dass Intelligenztests in der Regel zumindest als unverzichtbarer Bestandteil eines auf gültige Aussagen zielenden Assessment Centers vorgesehen werden (siehe z.B. Hossiep, 2001). Durch die Einbeziehung standardisierter Testverfahren in Assessment Center wird nicht nur die Validität erhöht, sondern auch die Leitidee der Assessment Center-Technik, die Methodenvielfalt, verwirklicht. Nach Jeserich (1981, zitiert nach Kleinmann, 1997) ist der Einsatz verschiedenartiger eignungsdiagnostischer Verfahren ein wesentliches Merkmal der Assessment Center-Methode. Uneinigkeit besteht darüber, welche Verfahren im Assessment Center miteinander kombiniert werden können.

Die teilweise vertretene Position, psychometrische Testverfahren aus dem Begriff „Assessment Center" auszuschließen, würde der historischen Begriffsbildung ebenso widersprechen wie einem Teil der aktuellen Assessment Center-Praxis. Der historische Vorläufer des Assessment Centers, die psychologische Prüfung der Offiziersanwärter in der Reichswehr, verwendete sowohl Verhaltensübungen als auch Intelligenztestverfahren. Der Begriff „Assessment Center" wurde in den 30er Jahren von Murray (1938, zitiert nach Stehle, 1982, S. 50) für die Kombination von psychologischen Tests und Übungen eingeführt. In den „klassisch" gewordenen Assessment Centern der britischen Armee und der American Telephone and Telegraph Company (Bray, Campbell & Grant, 1974) wurden situative Übungen stets mit Intelligenztests

kombiniert. Die „Guidelines and Ethical Considerations for Assessment Center Operations" (1989, S. 461) sehen Tests explizit als eine „assessment technique". Auch nach Schuler und Schmitt (1987, S. 264) kann ein Test genauso Bestandteil eines Assessment Centers sein wie z.B. eine Gruppenaufgabe oder eine Postkorb-Übung. Entsprechend umfasst das von Annen im vorliegenden Buch vorgestellte Assessment Center standardisierte Leistungstests. Sarges (2001, S. XVI) fordert, Tests wieder zum selbstverständlichen Bestandteil der multiplen Verfahrenskombination Assessment Center werden zu lassen, und nach Hossiep (2001, S. 59) ist die „Integration solcher Instrumente im AC nicht nur sinnvoll, sondern geradezu zwingend". In Nordamerika sind „skill and ability"-Tests in fast einem Drittel aller Fälle Bestandteil des Assessment Centers (Spychalski, Quinones, Gaugler & Pohley, 1997).

Die Integration von psychometrisch hochwertigen standardisierten Testverfahren in das Assessment Center ist unter Validitätsaspekten fachlich geboten. Gerade diesbezüglich erweist sich die Praxis unter dem Gesichtpunkt der Verfahrensgültigkeit als defizitär, wobei die Integration von Tests in Assessment Center einer vom Arbeitskreis Assessment Center initiierten Übersicht zur Assessment Center-Praxis zufolge in den deutschsprachigen Ländern seltener erfolgt als in den Vereinigten Staaten (siehe Krause und Gebert, in Druck). Positiv lässt sich formulieren, dass sich die Vorhersagegüte des Assessment Centers durch eine einfache und kostengünstige Maßnahme (nämlich die Integration von psychometrischen, standardisierten Testverfahren) leicht dramatisch verbessern lässt, bzw. sich die ursprüngliche Verfahrensvielfalt und Validität der traditionellen Assessment Center (z.B. AT&T) wieder erreichen lässt. Notwendig ist natürlich eine tatsächliche Integration; dies erfordert beispielsweise die Ableitung der mit den Tests erfassten Dimensionen aus der Anforderungsanalyse, die Verwirklichung des Redundanz-Prinzips (indem z.B. mehrere Testaufgaben die gleiche Dimension erfassen) und die Integration der Testergebnisse und Assessorenurteile.

Einschränkung der Aussagekraft von Qualitätskontrollen der Personalentscheidung

Hauptanliegen des Artikels ist es zu untersuchen, ob die auf der Basis des Assessment Centers getroffenen Aussagen eine treffsichere Prognose des zukünftigen Ausbildungserfolgs ermöglichen. Vor der Vorstellung dieser Ergebnisse soll kurz auf die Probleme der (1) Streuungseinschränkung und (2) Kriteriumsdefizienz eingegangen werden. Diese beiden Aspekte erschweren Qualitätskontrollen für Personalentscheidungsverfahren.

Unterschätzung der Vorhersageleistung aufgrund von Streuungseinschränkungen

Ein grundsätzliches Problem bei Qualitätskontrollen in der Personalauswahl besteht darin, dass die einstellende Organisation in der Regel den aus dem Verfahren abgeleiteten Empfehlungen weitgehend nachkommt und entsprechend nur positiv eingestufte Bewerber auswählt. Infolgedessen müssen sich die Qualitätskontrollen auf eine Analyse der überwiegend positiv beurteilten Teilnehmer beschränken. In der Begrifflichkeit der Statistik spricht man in diesen Fällen von einer „abgeschnittenen" Verteilung und meint damit, dass ein Teil der Bewerber, nämlich die Gruppe der im Assessment Center relativ schlecht beurteilten Personen, nicht in die Analyse eingeht. Für die nicht zugelassenen Personen lässt sich nicht nachweisen, dass sie zu Recht von der Ausbildung/der Berufstätigkeit ausgeschlossen wurden. Dies ist ein Grund dafür, dass die nachfolgend berichteten Zusammenhänge zwischen den Ergebnissen des Assessment Centers und dem Ausbildungserfolg an der Polizei-Führungsakademie den tatsächlich bestehenden Zusammenhang möglicherweise unterschätzen, weil gerade die im Assessment Center besonders schlechten Kandidaten möglicherweise auch besonders schlechte Ausbildungsergebnisse erzielt hätten.

Unterschätzung der Vorhersageleistung aufgrund der Kriteriumsdefizienz

Die Treffsicherheit der aus dem Assessment Center abgeleiteten Aussagen kann hier lediglich anhand des Ausbildungserfolgs evaluiert werden. Dies ist kein zufriedenstellendes Kriterium, da mit dem Assessment Center nicht nur der Ausbildungserfolg, sondern letztendlich auch der Berufserfolg vorhergesagt werden soll. Einige Testkomponenten, vor allem aber die Übungen des mündlichen Teils des Assessment Centers, wurden im Hinblick auf die Vorhersage des Berufserfolgs maßgeschneidert. Die diesbezüglichen Vorhersagequalitäten des Verfahrens können in der vorliegenden Studie mit dem Kriterium „Ausbildungserfolg" nicht evaluiert werden, der Erfolg an der Polizei-Führungsakademie stellt nicht das „eigentliche" (ultimate) Kriterium dar.

Die Untersuchung

Untersuchungsteilnehmer

Analysiert wurden die Daten von 112 männlichen Polizeibeamten aus vier Bundesländern im Alter zwischen 27 und 43 Jahren (Median 33, Standardabweichung=3.5), die sich in den Jahren 1991 bis 1998 mit Erfolg einem Assessment Center unterzogen haben und zur Ausbildung an der Polizei-Führungsakademie zugelassen wurden.

Analysen zur prognostischen Treffsicherheit des Assessment Centers

Bei der Beurteilung im Assessment Center handelt es sich letztendlich um Prognosen. Ausgewählt werden diejenigen Personen, von denen die Assessment Center-Kommission annimmt, dass sie sich mit hoher Wahrscheinlichkeit in der Ausbildung und im Beruf bewähren. Die hier vorgestellte Qualitätskontrolle stellt nun diese „Prognosen" (Ergebnisse des Assessment Centers) dem sich Jahre später tatsächlich einstellenden oder ausbleibenden Erfolg gegenüber. Grundlage der vorliegenden Evaluation ist die Abschlussnote

an der Polizei-Führungsakademie, der Abschluss wurde zwischen 1994 und 2000 erzielt. Die Benotung an der Polizei-Führungsakademie bedient sich der so genannten „15-Punkte-Skala". Abbildung 10 stellt die Umrechnung der Punkte in (Schul-)Notenstufen dar. Drei der 112 analysierten Personen konnten die Ausbildung an der Polizei-Führungsakademie nicht erfolgreich beenden.

von	0,00	bis	1,99 Punkte	=	„ungenügend"
von	2,00	bis	4,99 Punkte	=	„mangelhaft"
von	5,00	bis	7,99 Punkte	=	„ausreichend"
von	8,00	bis	10,99 Punkte	=	„befriedigend"
von	11,00	bis	13,49 Punkte	=	„gut"
von	13,50	bis	15,00 Punkte	=	„sehr gut"

Abb. 10: Umrechnung der Punktwerte in Notenkategorien

Tabelle 6 (s.S. 89) stellt die Verteilung der aus dem Assessment Center abgeleiteten abschließenden Empfehlungsgrade („Gesamtempfehlungsgrad") der Verteilung des Kriteriums (dem Abschluss an der Polizei-Führungsakademie in Notenstufen) in Form einer Kreuztabelle deskriptiv gegenüber. Die Empfehlungsgrade wurden ebenfalls in Kategorien zusammengefasst.

- Für jede „Zelle" der Tabelle sind drei unterschiedliche Angaben verzeichnet, die am Beispiel der äußerst rechten Zelle in der ersten Zeile (Empfehlungsgrad Kategorie „drei" [„3-", „3", „3+"] und dem Ergebnis „nicht bestanden") kurz erläutert werden sollen.
- Die Angabe in der ersten Reihe dieser Zelle bezieht sich auf die absolute Anzahl von zwei Personen. Die erste Zahl ist eine absolute Nennung, die in Klammern aufgeführte Zahl der entsprechende prozentuale Wert. Also: Zwei der insgesamt 112 analysierten Personen (also 1,8% der Untersuchungsgruppe) haben einen Gesamt-Empfehlungsgrad der Kategorie „drei" erzielt und die Prüfung an der Polizei-Führungsakademie „nicht bestanden".
- Die Zahl darunter bezieht sich nur auf die Teilnehmer mit einem Gesamtempfehlungsgrad der Kategorie „drei": 11,1 % der 18 Personen, die im Assessment Center einen Gesamtempfehlungsgrad der Kategorie „drei" erzielten, haben die Prüfung nicht bestanden (Zeilenprozent).
- Die letzte, grau hinterlegte Angabe pro Zelle bezieht sich nur auf die Personen, die in der Prüfung „durchgefallen" sind: 66,7 % derjenigen, die in der Prüfung „durchgefallen" sind, erreichten im Assessment Center einen Gesamtempfehlungsgrad von „drei" (Spaltenprozent).

Gesamtergebnis des Assessment Centers	***Abschluss der Polizei-Führungsakademie***			
	„gut"	**„befriedigend"**	**„ausreichend"**	**„nicht bestanden"**
den Anforderungen nur teilweise entsprechend („3-", „3" und „3+")		9 (8,0 %)	7 (6,3 %)	2 (1,8 %)
		50,0 %	38,9 %	11,1 %
		12,5 %	31,8 %	66,7%
den Anforderungen weitgehend entsprechend – untere Gruppe („4-" u. „4")	3 (2,7 %)	34 (30,4 %)	10 (8,9 %)	1 (0,9 %)
	6,3 %	70,8 %	20,8 %	2,1 %
	20,0 %	47,2 %	45,5 %	33,3 %
den Anforderungen weitgehend entsprechend – obere Gruppe („4+" u. „5-")	8 (7,1 %)	23 (20,5 %)	5 (4,5 %)	
	22,2%	63,9 %	13,9 %	
	53,3 %	32,0 %	22,7 %	
den Anforderungen entsprechend („5")	4 (3,6%)	6 (5,3 %)		
	40 %	60 %		
	26,7 %	8,3 %		
Gesamt N=112	15	72	22	3
	(13,4 %)	(64,3 %)	(19,6%)	(2,7%)
	(100 %)	(100 %)	(100 %)	(100 %)

Tab. 6: Kreuzklassifikation

Um die Treffsicherheit der im Assessment Center aufgestellten Prognosen zu bestimmen, ist die Zahl in der zweiten Reihe pro Zelle ausschlaggebend. Die Organisation will erfahren, inwieweit sie sich darauf verlassen kann, dass ein Bewerber, der im Assessment Center als „den Anforderungen entsprechend" (= „5") klassifiziert wurde, auch tatsächlich den Anforderungen der Ausbildung

(= „5") klassifiziert wurde, auch tatsächlich den Anforderungen der Ausbildung genügt. Die Kreuztabelle ist diesbezüglich ein Beleg für die Treffsicherheit des Assessment Centers: Von den zehn Personen, die im Auswahlverfahren als „den Anforderungen voll entsprechend" (Gesamtempfehlungsgrad „fünf") beurteilt wurden, gab es keine, die dann den Anforderungen der Polizei-Führungsakademie entgegen der Aussage des Auswahlverfahrens nicht genügte.

Alle diese Personen bestanden die Prüfung entweder mit „gut" oder mit „befriedigend". Selbst ein nur „ausreichendes" Bestehen der Prüfung kam in dieser Gruppe nicht vor. Der Gesamtempfehlungsgrad „voll anforderungsgerecht" impliziert also quasi eine Garantie für ein erfolgreiches Abschneiden bei der Prüfung.

Eine solche „Garantie" gibt es für Personen mit „leichten Einschränkungen" nicht. Die Gesamtempfehlungsgrade „fünf minus" und „vier plus" (bessere Gruppe der Kategorie „den Anforderungen weitgehend entsprechend") signalisieren entsprechend der Skalen-Definition minimale Einschränkungen in Bezug auf den zu erwartenden Erfolg.

In der vorliegenden Untersuchung entsprach auch diese Gruppe noch den Ausbildungsanforderungen, allerdings konnten fünf dieser insgesamt 36 Personen – knapp 14% der so im Assessment Center eingestuften Personen – die Ausbildung nur mit „ausreichend" absolvieren.

Ein höheres Risiko ergibt sich – konform zur diesbezüglichen Skalendefinition – bei den Empfehlungsgraden „vier" und „vier minus" (schlechtere Gruppe der Kategorie „den Anforderungen weitgehend entsprechend"). In dieser Gruppe nimmt die durchschnittliche Prüfungsleistung relativ gesehen ab, sogar eine der drei „durchgefallenen" Personen erzielte im Assessment Center einen Gesamtempfehlungsgrad dieser Kategorie.

Besonders interessant ist die Gruppe der Personen, die das Assessment Center mit der Empfehlungsgrad-Kategorie „drei" („3-", „3" und „3+", „den Anforderungen nur teilweise entsprechend") abgeschlossen hat. Aufgrund der Ergebnisse im Assessment Center bestanden hier seitens der Assessoren deutliche Bedenken gegenüber einer Zulassung. Tatsächlich haben 50% dieser Gruppe die Prüfung entweder nur mit der schlechtesten Note „ausreichend" bestanden oder sind „durchgefallen". Keine einzige Person aus dieser Gruppe konnte die Note „gut" erzielen. Der Einwand, dass ein Großteil (50%) der Kandidaten mit einem Gesamtempfehlungsgrad im Bereich von „drei" die Ausbildung mit der Note „befriedigend" absolviert hat, spricht nicht grundsätzlich gegen die Aussagekraft des Personalauswahlverfahrens.

Die Einstufung dieser Personen als „Risikokandidaten" bestätigt sich vielmehr während der Ausbildung an der Polizei-Führungsakademie. Das Risiko, die Ausbildung nicht zu bestehen, ist bei diesen Personen größer als bei den im Assessment Center positiver beurteilten Personen. In der Regel ist man bei wichtigen und kostspieligen Personalentscheidungen nicht risikofreudig. Unsichere Kandidaten werden zugunsten aussichtsreicherer Bewerber zurückgewiesen, auch wenn man ihnen damit eine potenziell vorhandene Bewährungschance verwehrt. Die im Assessment Center deutlich schlechter (mit „2" und „1") beurteilten Personen wurden erst gar nicht zur Ausbildung zugelassen. Setzt man gedanklich den klaren Zusammenhangstrend fort, so ist zu erwarten, dass die zahlreichen Bewerber, die aufgrund ihrer schlechten Leistung im Assessment Center mit „2" oder „1" beurteilt wurden, auch in der Prüfung versagt hätten.

In einem weiteren Analyseschritt wurde der statistische Zusammenhang zwischen den Ergebnissen des Assessment Centers und dem Punktwert der Abschlussprüfung als Kriterium in korrelativer Form bestimmt. Um mögliche Missverständnisse bei der Interpretation zu vermeiden, werden die logisch positiven Korrelationen berichtet, obwohl rein nominell aufgrund der unterschiedlichen Polung der beiden Skalen der Zusammenhang zur Punktwertskala positiv und der Zusammenhang zur Notenskala negativ ausfällt.

Der in Tabelle 7 dargestellte Zusammenhang entspricht einer Rangkorrelation von r=.46. In der Tabelle 7 sind sowohl die Ergebnisse des Assessment Centers als auch die Ergebnisse der Prüfung zu Kategorien zusammengefasst worden. Dem Vorteil der Übersichtlichkeit steht der Nachteil gegenüber, dass möglicherweise bestehende oder ausbleibende Zusammenhänge auf einer feiner auflösenden Betrachtungsebene nicht entdeckt werden. Außerdem sind die Werte auf dieser groben Auflösungsstufe nicht normal verteilt. Schließlich bietet die Kategorienbildung Raum zur Manipulation, indem jeweils die „passenden" Werte zu Kategorien gebündelt werden. Die nachstehenden Analysen beschränken sich aus diesen Gründen auf die exakten Zusammenhänge zwischen den Ergebnissen im Assessment Center auf der Ebene aller möglichen Empfehlungsgradstufen einerseits und den exakten Punktwerten (mit zwei Nachkommastellen) auf der Seite der Kriterien andererseits.

	Empfehlungsgrad-Skala			**z-Skala**
Ausbildungserfolg (exakte Punktwerte)	**Test**	**Verhalten**	**Gesamt**	**Test**
Gesamtgruppe (N=112)	.41**	.28*	.47**	--
Teilgruppe (N=91)	43**	.30**	.48**	.54**

** p<.01; * p<.05; einseitig;
Tab. 7: Korrelativer Zusammenhang

Die korrelativen Zusammenhänge sind in der Tabelle 3 wiedergegeben. Die Korrelation zwischen der Assessment Center-Gesamtbewertung in den Empfehlungsgradstufen und dem Erfolg an der Polizei-Führungsakademie betrug r=.47 für die Vorhersage der exakten Punktwerte. Die Korrelation zwischen der Testleistung in den Empfehlungsgradstufen und dem Erfolg an der Polizeiführungsakademie (in Punktwerten) betrug r=.41. Demgegenüber war der

„Empfehlungsgrad Verhalten" nur zu r=.28 mit dem in der Prüfung erzielten Gesamtpunktwert assoziiert.

Die unterschiedliche Vorhersagekraft der Leistungstests einerseits und der situativen Übungen andererseits zeigt sich insbesondere, wenn man für den Testwert nicht die Empfehlungsgradskala, sondern die differenziertere z-Wert-Skala zugrunde legt. Wie oben ausgeführt, wurde zu diesem Zweck der Mittelwert aus den Leistungen in den vier Testdimensionen gebildet. Diese Ergebnisse lagen nur für eine Teilgruppe im Umfang von 91 Personen vor. Um diese Ergebnisse mit den Ergebnissen der Empfehlungsgradskala vergleichen zu können, werden zusätzlich auch die bereits berechneten Korrelationen für die Gruppengröße von N=91 berichtet. Die Vorhersagekraft des Testmittelwerts fiel in Bezug auf die Vorhersage der in der Ausbildung erzielten Punktwerte mit r=.54 nominell deutlich höher aus als die des Testempfehlungsgrades (r=.43). Dies bedeutet, dass bei der Transformation der Testwerte in den Empfehlungsgrad ein Informationsverlust stattfindet. Als Ursache für diesen Informationsverlust kommen entweder die geringere Differenzierungsfähigkeit der Testempfehlungsgradskala und/oder aber defizitäre Beurteilungsprozesse in Frage. Der Testempfehlungsgrad wird nicht mathematisch bestimmt, sondern von den Psychologen intuitiv erfahrungsgeleitet festgelegt. Diese so genannte „klinische" Vorgehensweise der Beurteilung ist einer statistischen Verrechnung qualitativ unterlegen (siehe z.B. Wiggins, 1973).

Das aufgrund der Daten im standardisierten psychometrischen Test erzielte Assessment Center-Ergebnis ist mit r=.54 deutlich aussagekräftiger als das Assessment Center - „Verhaltensurteil", welches nur mit r=.30 mit dem Ausbildungserfolg assoziiert ist. Die Differenz zwischen den beiden Korrelationen ist statistisch bedeutsam (t=2.11, p<.05; Berechnung nach Cohen & Cohen, 1975, S. 53 f.). Somit bestätigt auch diese Studie, dass standardisierte psychometrische Testverfahren unter Vorhersagegesichtspunkten ein zumindest unverzichtbarer Bestandteil von Assessment Centern sind.

Fazit

Die aus dem Assessment Center-Verfahren abgeleiteten Prognosen/Personalentscheidungen haben sich in der Praxis – gemessen am Kriterium Ausbildungserfolg – insgesamt bewährt. Die vorliegende Bewährungskontrolle unterstreicht einmal mehr die Bedeutung standardisierter psychometrischer Testverfahren. Durch die Berücksichtigung von Testverfahren kann die Aussagekraft von Assessment Centern gesteigert werden.

Kritisch ist zu bewerten, dass der einfache Mittelwert der vier kognitiven Testdimensionen eine nominell (noch) bessere Vorhersage ermöglicht als die von den Psychologen vorgenommene Einstufung der Testleistung auf der zur Verfügung stehenden Empfehlungsgradskala. Diesbezüglich sollte nach Möglichkeiten gesucht werden, den durch die erfahrungsgeleitete Methode der Testbewertung und/oder den durch die Reduktion der Skalenbreite bedingten Informationsverlust zu reduzieren.

Unter betriebswirtschaftlichen Kostenrechnungen steht der Nutzen einer qualitativ hochwertigen Personalentscheidung außer Frage (siehe z.B. Höft, 2001). Zu beachten sind außerdem zusätzliche Nutzenaspekte, etwa die Gelegenheit zur Selbstreflexion und Begünstigung der Selbstselektion für die Bewerber, die Erkenntnisse für die Assessoren (Erkenntnisse über die Qualität und den Trainingsbedarf des internen Personalmarktes, Erkenntnisse und Erfahrungen zum Themenkomplex „Beobachtung und Beurteilung" usw.) sowie die insgesamt positive, Akzeptanz erzeugende Wirkung, die von einer professionellen Gestaltung interner Personalentscheidungen auf die Organisation ausgeht. Dieser „Zusatznutzen" basiert wesentlich auf den situativen Verfahrenskomponenten des Assessment Centers, so dass diese Elemente trotz ihrer im Vergleich zu psychometrischen Testverfahren geringen Aussagekraft ihre Berechtigung haben. Der entscheidende Grund für oder gegen ein Personalentscheidungsverfahren ist aber dessen nachweisliche Aussagekraft.