Wirtschaftspsychologie

Heinz Schuler (Hrsg.)

Assessment Center zur Potenzial-analyse

HOGREFE

Assessment Center zur Potenzialanalyse

Wirtschaftspsychologie

Assessment Center zur Potenzialanalyse

hrsg. von Prof. Dr. Heinz Schuler

Herausgeber der Reihe:

Prof. Dr. Heinz Schuler

Assessment Center zur Potenzial-analyse

herausgegeben von
Heinz Schuler

HOGREFE
GÖTTINGEN · BERN · WIEN · PARIS · OXFORD · PRAG
TORONTO · CAMBRIDGE, MA · AMSTERDAM · KOPENHAGEN

Prof. Dr. Heinz Schuler, geb. 1945 in Wien. Studium der Psychologie und Philosophie in München, Promotion 1973 und Habilitation 1978 in Augsburg. Nach Auslandsaufenthalten 1979 Professor und Institutsvorstand in Erlangen, seit 1982 Inhaber des Lehrstuhls für Psychologie der Universität Hohenheim, daneben Wissenschaftlicher Leiter der S&F Personalpsychologie Managementberatung in Stuttgart. Autor mehrerer eignungsdiagnostischer Verfahren und Standardwerke der Organisations- und Personalpsychologie.

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Göttingen · Bern · Wien · Paris · Oxford · Prag
Toronto · Cambridge, MA · Amsterdam · Kopenhagen
Rohnsweg 25, 37085 Göttingen

http://www.hogrefe.de
Aktuelle Informationen · Weitere Titel zum Thema · Ergänzende Materialien

Gesamtherstellung: AZ Druck und Datentechnik GmbH, Kempten
Printed in Germany
Auf säurefreiem Papier gedruckt

ISBN: 978-3-8017-2035-3

Inhalt

Teil I: Gegenstand und Überblick

Teil II: Anforderungen, Dimensionen, Konstrukte

Teil III: Kriterienbezogene Validität

Teil IV: Reliabilitätssicherung

Teil V: Anwendungsbereiche

Teil I:
Gegenstand und Überblick

1 Assessment Center als multiples Verfahren zur Potenzialanalyse: Einleitung und Überblick

Heinz Schuler

Das Assessment Center nimmt aus unterschiedlichen Gründen eine Sonderstellung unter den berufseignungsdiagnostischen Verfahren ein. Es handelt sich um ein *multiples* Verfahren, es wird häufiger als andere eignungsdiagnostische Methoden auch für *interne* Personalentscheidungen eingesetzt, seine Einsatzhäufigkeit hat in den beiden letzten Jahrzehnten stärker zugenommen als die anderer Verfahren, das Assessment Center wird vornehmlich von eignungsdiagnostischen Laien durchgeführt, und sein Durchführungsaufwand ist erheblich größer als der jedes anderen eignungsdiagnostischen Verfahrens. Deshalb lässt sich seine Popularität weniger durch seine psychometrische Qualität erklären als durch Funktionen, die andere Auswahlverfahren nicht in gleichem Maße zu bieten haben.

Im Folgenden wird zunächst eine kurze Gegenstandsbestimmung vorgenommen, und es werden die Verwendungshäufigkeit, Einsatzzwecke und Verfahrenskomponenten des Assessment Centers aufgezeigt. Im Anschluss daran wird die Herkunft dieses Verfahrenstypus dargelegt; hierbei werden einige Verfahrensdetails aufgezeigt, die für die spätere Diskussion der Frage „Welche Elemente sollen Bestandteil eines Assessment Centers sein?“ von Bedeutung sind. Im Weiteren werden zentrale methodische Probleme angesprochen, die derzeit im Assessment Center-Kontext erforscht und diskutiert werden. Die wichtigsten dieser Probleme werden in den weiteren Kapiteln dieses Bandes behandelt und teilweise neuen Lösungen zugeführt. Deshalb kann sich der letzte Teil dieser Einleitung dem Überblick über die weiteren Beiträge dieses Bandes widmen.

1.1 Gegenstand, Verbreitung und Einsatzzwecke

Assessment Center ist der Name einer multiplen Verfahrenstechnik, zu der mehrere eignungsdiagnostische Instrumente oder leistungsrelevante Aufgaben zusammengestellt werden. Ihr Einsatzbereich ist die Einschätzung aktueller Kompetenzen oder die Prognose künftiger beruflicher Entwicklung und Bewährung. Sie wird deshalb sowohl zur Auswahl künftiger Mitarbeiter wie auch als Beurteilungs- und Förderinstrument eingesetzt. Charakteristisch für Assessment Center ist, dass mehrere Personen (etwa 4 bis 12) gleichzeitig als Beurteilte daran teilnehmen und dass auch die Einschätzungen von mehreren Beurteilern (im Verhältnis etwa 1 Beurteiler: 2 Beurteilten) vorgenommen werden. Die Beurteilergruppe besteht vor allem aus Linienvorgesetzten (typischerweise zwei Hierarchieebenen über der Zielebene der zu Beurteilenden) sowie aus Psychologen und Mitarbeitern des Personalwesens. Die Einschätzungen erfolgen üblicherweise dimensionenbezogen (d. h. Verhaltensbeobachtungen werden Eignungsmerkmalen zugeordnet). Abschließend werden die Beurteilungen durch Diskussion unter den Beurteilern

(„Assessoren“) oder auf statistischem Wege aggregiert und in vielen Fällen als Feedback und evtl. als Entwicklungsempfehlung den Teilnehmern mitgeteilt. Die Durchführungsdauer weist eine große Streubreite von wenigen Stunden bis zu einer ganzen Woche auf. Wie in einigen Beiträgen dieses Bandes berichtet wird (z. B. Kap. 15), sind bei geeigneter Konzeption alle wesentlichen Informationen innerhalb eines Tages erhebbar.

Verschiedentlich wird heute der Begriff Assessment Center auch für Diagnosen verwendet, denen nicht eine Mehrzahl von Personen, sondern allein Individuen unterzogen werden. Dieses Bemühen, von der attraktiven Begriffsassoziation auch für Individualdiagnosen zu profitieren, ist verständlich, aber semantisch unzweckmäßig. Eine gemäßigte terminologische Variante ist das „Einzel-Assessment“, gegen das nichts zu sagen wäre, wenn es nicht nach Unkenntnis dessen klänge, dass psychologische Diagnostik seit Anbeginn ihrer systematischen Existenz, also seit gut einem Jahrhundert, üblicherweise die Form der Individualdiagnostik hat und nicht eine Sonderform der wesentlich jüngeren Gruppenverfahren darstellt.

Gelegentlich werden auch internetgestützte Verfahren als Assessment Center bezeichnet. Gewöhnlich trifft auf sie aber nicht zu, was die Besonderheit des Assessment Centers gegenüber anderen Diagnoseformen ausmacht. Bei technischer Weiterentwicklung – oder auch in konsequenter Nutzung der bereits bestehenden technischen Möglichkeiten – wäre aber eine Durchführung auch interaktiver Verfahren auf diesem Wege – etwa in Form einer Videokonferenz – denkbar (wobei zu berücksichtigen wäre, dass sich hierdurch die Anforderungscharakteristik möglicherweise verändert). Eine systematische Gegenüberstellung des Assessment Centers mit anderen multiplen Diagnoseverfahren nimmt Kleinmann (2003) vor.

Gelegentlich wird das Fehlen einer Eindeutschung des Namens „Assessment Center“ beklagt. „Multiples eignungsdiagnostisches Beurteilungsverfahren“ wäre vielleicht eine Benennung, die das Wesen der Sache trifft; aber weder ist ihr die marktgemäße begriffliche Prägnanz des englischen Wortpaares eigen, noch scheint dieses beim derzeitigen fortgeschrittenen Stand der Verbreitung überhaupt durch eine deutsche Bezeichnung ablösbar zu sein. So wollen wir also, ungeachtet gewisser Ungereimtheiten bei der Deklination im Deutschen, uns auch weiterhin dieses Terminus technicus bedienen und uns mit dem Gedanken trösten, dass auch Beethovens Bemühung erfolglos war, den Namen des zu seiner Zeit noch *Fortepiano* genannten Musikinstruments mit „Starkschwach-Tastenkasten“ einzudeutschen.

In den 70er Jahren waren es in Deutschland erst eine Hand voll Unternehmen, die sich des Assessment Centers zur Auswahl und Entwicklung von Mitarbeitern bedienten. In einer eigenen Umfrage aus dem Jahr 1983 gaben 20 % von 120 befragten großen und mittleren Unternehmen an, diese Methode einzusetzen – vor allem bei Führungskräften und Trainees. Eine erneute Befragung ergab 1990 eine Steigerung auf 39 % (Schuler, Frier & Kauffmann, 1993; N = 105), um sich schließlich in der jüngsten Erhebung (Schuler, Hell, Trapmann, Schaar & Boramir, in Druck) auf 57,6 % zu steigern (Daten von 2003, N = 125). Diese Steigerung ist insofern bemerkenswert, als zu allen drei Befragungszeitpunkten das Assessment Center von den Verwendern als die aufwendigste und damit am wenigsten praktikable Methode von allen verglichenen Verfahrenstypen eingestuft wurde. Durchschnittlich werden in deutschen Großunternehmen Assessment Center 8-mal pro Jahr zur Personalauswahl und 6,5-mal zur Personalentwicklung eingesetzt (Kanning, Pöttker & Gelléri, in Druck).

Während Thornton und Byham (1982) die Auswahl und Entwicklung von Führungskräften als mit 95 % der Anwendungsfälle dominierende Zielsetzung des Assessment Centers angeben, bestand die bevorzugte Zielgruppe in deutschen Unternehmen über die letzten beiden Dekaden hinweg aus Hochschulabsolventen (v. a. Trainees). Die Einsatzzwecke des Assessment Centers sind vielfältig, die wichtigsten dürften sein:
- Auswahl externer Bewerber,
- interne Personalauswahl und -klassifikation,
- Laufbahnplanung,
- Trainingsbedarfsanalyse,
- Beurteilung, insbesondere Potenzialbeurteilung,
- Unternehmerdiagnose und Nachfolgeplanung,
- Ausbildungsberatung,
- Teamentwicklung,
- Berufsberatung,
- berufliche Rehabilitation,
- Arbeitsplatzgestaltung,
- Forschung.

Die Popularität des Assessment Centers würde allerdings nicht ausreichend verständlich, hielte man sich ausschließlich an die genannten „manifesten" Zielsetzungen dieser Diagnosemethode. Denn diese Ziele werden, wie die nachfolgend erläuterten Validitätswerte zeigen, nur in sehr begrenztem Maße erreicht. Vorteile des Verfahrens werden vielmehr auch in „latenten" Funktionen gesehen (vgl. Schuler, 1987), wie im Gewinn eines Überblicks über den Nachwuchs, über Leistungsstand und Defizite im Unternehmen (und zwar nicht nur im Hinblick auf Personen, sondern auch auf Organisationseinheiten, Programme, Führungsstile etc.), in der Gelegenheit zu verhaltensbezogenen Formulierungen von Anforderungen und Leistungsniveaus, in der Betonung der Bedeutung von Personalplanung und Personalentwicklung, der Möglichkeit, Aspekte der Unternehmenskultur zu diskutieren und zu inszenieren, die Teilnehmer mit den Anforderungen – auch sozialpsychologischer Art – einer Führungstätigkeit vertraut zu machen, ihre Selbsteinschätzung zu verbessern und ihnen die Gelegenheit zum sozialen Vergleich zu bieten. Schließlich scheint die Aufgabe des Beobachtens im Assessment Center nicht nur ein gutes Beurteilertraining darzustellen, sondern sogar der Erfüllung weiterer Aufgaben einer Führungskraft dienlich zu sein (Lorenzo, 1984) und überdies deren Selbstverständnis entgegenzukommen. Eine gründliche Erörterung dieser latenten Funktionen des Assessment Centers findet sich bei Neuberger (2002) und – in kritischer Perspektive – bei Kompa (2004).

1.2 Verfahrenselemente

Eine Vielzahl von Aufgabentypen wurde im Assessment Center-Kontext eingesetzt, einige wurden speziell hierfür entwickelt. Ein Großteil dieser Verfahren wird in der Literatur dokumentiert (z. B. Development Dimensions, 1977; Fisseni & Fennekels, 1995; Jeserich, 1981; Kleinmann, 2003; Lattmann, 1989; Obermann, 1992; Sarges, 2001; Sünderhauf, Stumpf & Höft, 2005; Thornton, 1992; Thornton & Byham, 1982). Vor allem

werden in der Literatur simulations- oder arbeitsprobenartige Verfahren beschrieben. Klassischerweise werden im Assessment Center allerdings Aufgaben unterschiedlichster Kategorien eingesetzt, darunter v. a.:

- individuell auszuführende Arbeitsproben und Aufgabensimulationen (v. a. Organisations-, Planungs-, Entscheidungs-, Controlling- und Analyseaufgaben),
- Gruppendiskussionen mit und ohne Rollenvorgabe,
- sonstige Gruppenaufgaben mit Wettbewerbs- und/oder Kooperationscharakteristik,
- Vorträge und Präsentationen,
- Rollenspiele (meist dyadisch, z. B. Verkaufsgespräch, Mitarbeitergespräch),
- Interviews (neuerdings v. a. in strukturierter Form),
- Selbstvorstellung,
- Entscheidungssimulationen (häufig in computergestützter Form),
- Fähigkeits- und Leistungstests,
- Persönlichkeits- und Interessentests,
- biografische Fragebogen.

Die Entscheidung, welche Art von Einzelverfahren im Assessment Center eingesetzt wird, ist nicht nur eine Frage des individuellen Anwendungsfalles, sondern spiegelt auch das Begriffsverständnis wider: Wird „Assessment Center" als multiples Verfahren verstanden, in dessen Rahmen bestmöglicher Gebrauch von der Vielfalt eignungsdiagnostischer Untersuchungsmöglichkeiten gemacht werden soll (z. B. Schuler, 2006), wird man es als unnötige Restriktion ansehen, dieses Instrument nur mit einer bestimmten Klasse von Einzelverfahren zu bestücken und stattdessen das Prinzip der inkrementellen Validität (Welches Verfahren erbringt einen zusätzlichen Validitätsbeitrag?) zum Maßstab der Gestaltung machen. Ordnet man demgegenüber das Assessment Center ganz den simulationsorientierten Verfahren der Personalauswahl zu (z. B. Höft & Funke, 2006), so beschränkt sich die Verfahrenspalette auf arbeitsprobenartige Diagnoseformen wie Diskussionen, Rollenspiele, Präsentationsaufgaben und Entscheidungssimulationen.

Soweit diese Benennungsentscheidung nur als terminologische Frage aufgefasst wird, ist sie relativ unerheblich. (Gegebenenfalls werden die nicht zum engen Assessment Center-Begriff gezählten Verfahrenstypen gesondert durchgeführt wie in dem in Kapitel 16 dieses Bandes vorgestellten Beispiel.) Wird demgegenüber, wie verschiedentlich praktiziert, der normative Anspruch erhoben, die Diagnose der Berufseignung grundsätzlich auf Verfahrenselemente zu beschränken, die der Verhaltensbeobachtung zugänglich und damit auch für eignungsdiagnostische Laien nachvollziehbar sind, so beschränkt man die Qualität berufsbezogener Diagnosen auf unvertretbare Weise.

Was die Benennung der im Assessment Center eingesetzten Einzelverfahren anbelangt, findet sich in den meisten deutschsprachigen Publikationen die Bezeichnung „Übungen" – ganz analog zum englischen Terminus „exercises". Von Übungen wird insbesondere dann gesprochen, wenn es sich um Aufgaben handelt, die Verhaltensbeobachtungen ermöglichen, also z. B. Rollenspiele; in vielen Fällen wird die Bezeichnung aber auch auf alle in ein Assessment Center einbezogenen Verfahren ausgedehnt, also auch auf Tests, Interviews etc. Demgegenüber werden vom Herausgeber dieses Bandes die Bezeichnungen „Aufgaben" oder „Verfahren" präferiert, was eher dem Prüfungs- oder Diagnosecharakter dieser Einheiten gerecht wird. Nach allgemeinem Sprachgebrauch wird als „Übung" ein Trainingselement, nicht aber eine Prüfungsaufgabe bezeichnet – ungeach-

tet dessen, ob die Konsequenz der Prüfung eine Auswahlentscheidung oder eine Maßnahme der Personalentwicklung ist. Die Vermutung liegt nahe, dass die Fehlbezeichnung in euphemistischer Absicht gewählt wird; sie suggeriert gewissermaßen, es handle sich beim Assessment Center nicht um die diagnostische Grundlage einer Entwicklungsmaßnahme, sondern um die Entwicklungsmaßnahme selbst. Dies trifft gemäß der Datenlage (vgl. Kap. 12) nicht zu. Der Herausgeber und die meisten Autoren dieses Bandes sprechen deshalb im Folgenden von Aufgaben, Verfahren, manchmal auch von Assessment Center-Elementen, Einzelaufgaben etc. Den Autoren wurde aber selbstverständlich die Freiheit eingeräumt, diese Einheiten wahlweise auch als Übungen zu bezeichnen.

1.3 Historische Entwicklung

Vorläufer des heutigen Assessment Centers finden sich erstmals ab 1926/27 in der Weimarer Republik zur Offiziersauswahl der Reichswehr. Das „Rundgespräch", bis heute als – zumeist führerlose – Gruppendiskussion ein Hauptmerkmal des Assessment Centers, wurde 1926 von Rieffert eingeführt, der von 1920 bis 1931 für die „Psychotechnik" im Heer zuständig war. Zu Beginn dieser Entwicklung wurden sowohl „psychotechnische" Methoden, wie sie im militärischen Bereich bereits bekannt und bewährt waren, als auch „charakterologische" Verfahren zur Erfassung der Persönlichkeit angewandt, wobei die charakterologischen Aspekte gegenüber den elementaristischen Teilfunktionsprüfungen später immer wichtiger wurden (Fritscher, 1985). Die Verbreitung psychologischer Methoden zur Auswahl militärischen Führungspersonals war nicht nur auf die Weimarer Republik beschränkt. So berichtet Hofstätter in einem Brief an von Renthe-Fink, dass er im Österreich der frühen 30er Jahre Intelligenztests, Gymnastikaufgaben, Exploration, Fragebogen und Leistungsproben eingesetzt habe (von Renthe-Fink, 1985, S. 98).

Der Ablauf der Prüfung erfolgte gewöhnlich dergestalt, dass gleichzeitig fünf bis acht Offiziersbewerber zwei bis zweieinhalb Tage lang verschiedenen Diagnostikverfahren unterzogen wurden. Hierbei wurden u. a. Intelligenztests, Sprechanalysen, physiognomische Aufnahmen und graphologische Tests sowie, als „situative" Methoden, „Testverfahren", „Befehlsreihe", „Führerprobe", „Exploration" und „Rundgespräch" durchgeführt. Als Datengrundlage für die Urteilsfindung dienten neben biografischen Daten („Lebenslaufanalyse") die „Ausdrucksdiagnostik", Leistungstests („Geistesanalysen") oder Interessenstests (vgl. Fritscher, 1985, S. 434 f.). Eine Exploration hatte zum Ziel, „in einem ganzheitlichen Akt den charakterologischen Gehalt des (Prüflings) zu erfassen" (Kreipe, 1936, S. 105; zitiert nach Fritscher, 1985, S. 436). Neben einer eher neutralen „analytischen" Exploration wurden auch „taraktische" Explorationen vorgenommen, Vorläufer der „Streßinterviews", wie sie dann vor allem vom Office of Strategic Services (OSS) systematischer eingesetzt wurden. Die „Führerprobe" bestand z. B. darin, einer Gruppe im Vortrag ein definiertes Problem zu erklären oder über ein Thema frei zu sprechen. Auch Aufgaben zur Ermittlung körperlicher Belastbarkeit und motorischen Geschicks waren von den Prüflingen zu bewältigen. Das Verfahren endete mit einer Schlussbetrachtung der Prüfergruppe, üblicherweise bestehend aus zwei Offizieren, einem Sanitätsoffizier/Psychiater, zwei Psychologen und einem „Hilfspsychologen".

Über den ersten umfangreichen Einsatz der multiplen Auswahlprozedur in den USA berichtet eine Dokumentation der Beurteilergruppe des Office of Strategic Services, einer

Vorläuferorganisation des CIA. Hauptziel war die Auswahl und das Training von Agenten für Europa und Südostasien (u. a. Ceylon, Indien, China) im militärischen Konflikt mit den Achsenmächten Deutschland und Japan, wobei insbesondere für den Aufbau und die Unterstützung von Widerstandsgruppen Führungsqualitäten unerlässlich schienen. Unter Mitarbeit einer großen Zahl bereits damals oder später namhafter Psychologen – unter ihnen Bronfenbrenner, Fiske, Murray, Newcomb und Tolman – war die Planung und Durchführung des Auswahlprogramms gleichzeitig als Validierungsstudie angelegt. Als Prüfungsaufgaben wurden vornehmlich solche gewählt, die man für Simulationen künftiger Tätigkeiten hielt. Die besondere Schwierigkeit bestand in der Ungewissheit bezüglich dieser Tätigkeiten, zumal vielfältige Einsatzbereiche in unterschiedlichen Kulturkreisen vorgesehen waren. Zusätzlich zu den Simulationsaufgaben wurden eine größere Zahl standardisierter Tests sowie ein biografischer und ein Gesundheitsfragebogen eingesetzt. Die Durchführung der Übungen war für die Kandidaten teilweise dadurch erschwert, dass sie eine falsche Identität aufrecht erhalten mussten. Auch in Belastungssituationen durften sie sich nicht in Widersprüche verwickeln (hier dürfte ebenfalls ein Ursprung des ominösen „Streßinterviews“ liegen). Dieses Datenmaterial wurde allerdings nicht konsequent psychometrisch verwertet, sondern diente vor allem als Grundlage eines Interviews. Als Kriterium der Validierung dienten retrospektiv vorgenommene Leistungsbeurteilungen, deren geringe Reliabilität und Streuung die Autoren des Berichts beklagen (Office of Strategic Services Assessment Staff, 1948). Trotzdem wurden Validitätswerte in jener Höhe erzielt, die sich später als typisch für das Assessment Center herausstellen sollten (Schuler, 1989).

Zur Einsicht, dass multiple Messungen mit verschiedenen Verfahren in verschiedenen Situationen und die Einschätzung durch mehrere unabhängige Beurteiler für die Qualität und Tiefe psychologischer Diagnosen von großem Wert sein können, leistete auch Murrays *Clinical and experimental study of fifty men of college age*, die Basis seiner Persönlichkeitstheorie (Murray, 1938), einen großen Beitrag. Teilweise war diese Arbeit sogar die unmittelbare Grundlage der OSS-Verfahrensweisen, an denen Murray als leitender Psychologe beteiligt war. Für Interessierte stellen sowohl der Bericht des Office of Strategic Services Assessment Staff (1948) als auch Murrays persönlichkeitspsychologisch-diagnostisches Werk (Murray, 1938) eine fruchtbare Lektüre dar. Eine einfacher zugängliche Quelle, in der das OSS-Verfahren sowie weitere militärische Anwendungen relativ ausführlich geschildert werden, ist die Monografie von Thornton und Byham (1982).

Nicht die erste, aber fraglos die für die spätere Verbreitung dieses Instruments bedeutendste frühe Assessment Center-Studie im zivilen Bereich war die ab 1956 in der American Telephone and Telegraph Company (AT & T) durchgeführte *Management Progress Study* (Bray, Campbell & Grant, 1974). Hierbei wurden 422 bereits beim Unternehmen beschäftigte Nachwuchs-Führungskräfte mit einer großen Anzahl psychologischer Tests untersucht und mit Simulationsaufgaben wie Postkorbaufgabe, Wirtschaftsspiel und führerloser Gruppendiskussion konfrontiert, daneben kamen Interviews und biografische Fragebogen zur Anwendung.

Ausgangspunkt der Aufgabensammlung in der AT & T-Studie war eine Liste von 25 mutmaßlich wichtigen Eigenschaften, Fähigkeiten und Werthaltungen erfolgreicher Manager – allerdings nicht auf der Basis von Anforderungsanalysen erstellt. Die Beurteiler stuften die Kandidaten nach gründlicher Diskussion bezüglich jedes dieser 25 Merkmale ein und gaben Einschätzungen ab, ob die Beurteilten im Laufe der darauffolgenden

10 Jahre in das mittlere Management aufrücken würden und sollten. Beurteilungen und Karriereerwartungen wurden später mit dem tatsächlichen Karriereerfolg verglichen. Die Ergebnisse zeigten hohe Vorhersageleistungen des Gesamtverfahrens, wobei die prognostische Validität vor allem auf die Arbeitsproben und die kognitiven Leistungstests zurückgingen, während Persönlichkeitstests und Interviews nur einen geringen Beitrag leisteten (vgl. Tab. 1 aus Schuler, 1989, S. 119).

Tabelle 1: Trefferquoten und Validitätskoeffizienten für die Prognosen der Management Progress Study bei AT & T (Daten kompiliert aus Howard, 1981 [nach Thornton & Byham, 1982] sowie aus Bray & Grant, 1966)

Prädiktor: Assessment Center		**Kriterium: Führungsposition**			
Einschätzung der Beurteiler: Erreicht der Kandidat das Mittlere Management innerhalb von 10 Jahren?	N	nach 8 Jahren		nach 16 Jahren	
		mit College	ohne College	mit College	ohne College
Ja	103	64 %	40 %	89 %	63 %
Nein oder fraglich	166	32 %	9 %	66 %	18 %
Validitätskoeffizient		.46	.46	.33	.40

Im Anschluss an die Verfahrens- und Validitätsdemonstration bei AT & T nahm die Nutzung des Assessment Centers zur Auswahl von Führungskräften, später auch für andere Positionen, speziell in nordamerikanischen Unternehmen, stetig zu. Auch in Großbritannien fand das Assessment Center in den 70er und 80er Jahren bereits rege Verbreitung, während es in anderen europäischen Ländern zu dieser Zeit noch weitgehend unbekannt war.Über die Verwendungshäufigkeit in deutschen Unternehmen wurde in Abschnitt 1.1 bereits berichtet; der Vergleich einiger europäischer Länder zum Zeitpunkt 1990 ist Tabelle 2 zu entnehmen (berichtet wird hier über die Auswahl externer Bewerber; bei internen Personalentscheidungen bietet sich ein ähnliches Bild).

Die Literatur zum Assessment Center hat in den vergangenen Jahrzehnten rasch an Umfang gewonnen. Hervorzuheben ist insbesondere das einflussreiche Werk von Thornton und Byham (1982), das weltweit Anregung für den Einsatz dieses Verfahrens sowie Grundlage anderer Veröffentlichungen war. Die ersten deutschsprachigen Publikationen zum Thema erfolgten etwa ab 1980, wobei das Buch von Jeserich (1981) viele potenzielle Anwender auf den Verfahrenstypus Assessment Center aufmerksam machte. Schuler und Stehle (1983) wiesen unter dem Stichwort „soziale Validität“ auf eine Qualität eignungsdiagnostischer Verfahren hin, die bis dato wenig Aufmerksamkeit gefunden hatte und die ihnen beim Assessment Center günstiger ausgeprägt schien als bei den meisten anderen Auswahlverfahren: die Information der Kandidaten über Tätigkeitsanforderungen, die Transparenz von Diagnose und Entscheidung, Partizipation und Möglichkeiten der Verhaltenskontrolle sowie offene und faire Urteilskommunikation (als Möglichkeit zum sozialen Vergleich sowie als direktes, konstruktives Feedback). In dieser Hinsicht wurde das Assessment Center wegweisend auch für andere eignungsdiagnostische Verfahren und für die Gestaltung von Personalauswahlprozessen.

Tabelle 2: Einsatzhäufigkeiten des Assessment Centers im Jahr 1990 (Daten nach Schuler, Frier & Kauffmann, 1993)

	N	Ungelernte Arbeiter	Auszubildende		Facharbeiter	Angestellte ohne Führungsaufgaben	Trainees	Führungskräfte		
			technisch	kaufmännisch				untere	mittlere	obere
Deutschland	105	0	0	11	3	9	40	14	15	12
Großbritannien	19	0	14	15	0	0	56	41	35	24
Benelux	21	0	11	9	0	0	0	13	12	20
Frankreich	21	0	10	0	0	0	33	16	5	0
Spanien	25	0	0	0	0	0	0	0	10	20

Anmerkungen: Befragt wurden große und mittlere Unternehmen. Angaben in Prozent der Unternehmen, die die jeweilige Berufsgruppe beschäftigen.

1.4 Die methodische Wende

Wissenschaftlich am Assessment Center Interessierte wurden 1982 durch eine Mitteilung von Sackett und Dreher aufgeschreckt. Kerngehalt ihrer Entdeckung war, dass es den Beurteilern nicht gelingt, innerhalb einer Aufgabe zwischen den verschiedenen Anforderungsdimensionen zu unterscheiden, dass sie also ein Gesamturteil pro Aufgabe abgeben anstelle der erwarteten differenzierten Einschätzung. Gleichzeitig zeigte sich geringe Übereinstimmung der Einschätzungen des jeweils gleichen Merkmals über verschiedene Aufgaben hinweg: Beispielsweise ergab sich eine Nullkorrelation zwischen den Beurteilungen des „Geschicks im Umgang mit Mitarbeitern", das in den verschiedenen Aufgaben gezeigt wurde. Die Originaldaten von Sackett und Dreher (1982) sind in Tabelle 3 wiedergegeben. Zu erkennen ist die minimale durchschnittliche Korrelation („Grand M") für die Beurteilungsdimensionen ($r = .074$) über die Aufgaben hinweg, die eigentlich hoch ausfallen sollte, sowie die theoriewidrig hohe durchschnittliche Korrelation der dimensionsbezogenen Beurteilungen innerhalb jeder der Einzelaufgaben ($r = .638$).

Campbell und Fiske hatten bereits 1959 ein Verfahren vorgeschlagen, die Konstruktvalidität diagnostischer Messungen zu bestimmen, die sog. Multitrait-Multimethod-Matrix, abgekürzt MTMM (Campbell & Fiske, 1959). Im Sinne dieser MTMM ist bei dem von Sackett und Dreher analysierten Datensatz weder das Erfordernis der *konvergenten* noch das der *diskriminanten Validität* erfüllt, die *Konstruktvalidität* ist also gering.

Nun ist der Begriff der Konstruktvalidität – der Frage danach, was mit einem diagnostischen Verfahren eigentlich gemessen wird – nicht mit der Konvergenz und Differenzierung von Korrelationsdaten im Sinne der MTMM erschöpft (der Schluss also vielleicht etwas zu weitgehend), aber konvergente und diskriminante Validität stellen zweifellos essenzielle Erfordernisse dar, und die Entdeckung stellt geradezu die rationale Rekonstruierbarkeit dessen in Frage, was im Assessment Center getan wird. Sollte die Seman-

Tabelle 3: Konvergente und diskriminante Validität im Datensatz A von Sackett und Dreher (1982, S. 404)

Dimension or exercise	Mean *r*
Dimension	
Oral communication	.21
Written communication	–.09
Interpersonal skills with subordinates	.00
Giving work assignments	.00
Analytical skills	.08
Organizing skills	.13
Organizational acumen	.19
Grand M	.074
Exercise	
In-basket	.56
Group	.74
Oral communication	.71
Written communication	.61
Role play (1)	.64
Role play (2)	.57
Grand M	.638

tik der Merkmalsbezeichnungen eine Differenzierung suggerieren, die nicht der Realität der Urteilsbildung entspricht?

Bemerkenswerterweise wurde die Praxis der Assessment Center-Anwendung dadurch nicht besonders beeinflusst. Heute, nach 25 Jahren weiterer Forschung, muss man sagen, die Praxis hatte nicht Unrecht, sich nicht allzu sehr beunruhigen zu lassen, denn einige nachfolgende Arbeiten – z. B. Guldin und Schuler (1997) sowie Kleinmann (1997) – zeigten, dass ein Teil der verwirrenden Ergebnisse darauf zurückgeht, dass die Anforderungsdimensionen den verschiedenen Aufgaben nicht gleichermaßen angemessen sind, sich also auch nicht gleich gut beobachten lassen (siehe hierzu auch Lance, in diesem Band). Auch ist nicht auszuschließen, dass die verschiedenen Dimensionen etwas Unterschiedliches bedeuten, wenn sie mit verschiedenen Verfahren erfasst werden, dass es also eine Verflechtung von Merkmal und Methode gibt.

Eine weitere mögliche Schlussfolgerung klingt möglicherweise noch etwas radikaler: Könnte es nicht sein, dass in den Assessment Center-Matrizen und erst recht auf den Beurteilungsblättern der Assessoren gewissermaßen nur „Oberflächenmerkmale“ verzeich-

net sind, deren Benennungen relativ beliebige Vereinbarungen darstellen? Was stattdessen wirklich beurteilt wird, sind vielleicht „Grundmerkmale“, die gewissermaßen „hinter dem beobachteten Verhalten stehen“. Dies würde jedenfalls neueren Theorien der sozialen Urteilsbildung entsprechen, die annehmen, der Eindruck von einem Menschen komme nicht in Form einer Synthese vieler Einzelbeobachtungen – Verhaltensbeobachtungen – zu Stande, sondern als Globalurteil in Form weniger genereller Kategorien. Die Beurteilung der Verhaltensmerkmale könnte dann bereitwillig den terminologischen Vorgaben folgen, die sich einer oberflächlichen Phänomenologie leicht fügen, ohne deshalb den Fokus der Eindrucksbildung wesentlich anpassen zu müssen.

Eine naive landläufige Meinung scheint darin zu bestehen, die Assessoren im Assessment Center beobachteten und beurteilten gerade das, was ihnen auf den Beurteilungsblättern vorgegeben wird. Steht da „Kooperationsfähigkeit“, beurteilen sie Kooperationsfähigkeit, lautet die Anweisung, beschränken Sie sich auf singuläre beobachtbare Verhaltensweisen, trennen Sie Beobachtung und Bewertung und achten Sie nur auf den Blickkontakt, so fügen sich die Assessoren auch dieser Anweisung und tun, wie ihnen geheißen. Könnte es nicht sein, dass sie (1) nicht willens oder nicht in der Lage sind, Beobachtung und Beurteilung zu trennen (eine Erkenntnis, die schon auf die Gestaltpsychologie zu Anfang des zwanzigsten Jahrhunderts zurückgeht), und/oder dass sie sich (2) vielleicht auch an den genereller formulierten Merkmalen („Kooperationsfähigkeit“) weniger orientieren als an ihren eigenen Annahmen darüber, welche Eigenschaften erfolgsrelevant sind – zumal wenn sie als Führungskräfte des betreffenden Unternehmens mit den Anforderungen und Erfolgsvoraussetzungen auf ihre Weise vertraut sind. Diese Annahmen wiederum könnten Unternehmensspezifisches enthalten, dürften zu einem guten Teil aber an einigen allgemeinen, zuvor als Grundmerkmale apostrophierten Eigenschaften orientiert sein.

Scholz und Schuler (1993) führten eine Metaanalyse durch – also eine statistische Reanalyse einer größeren Zahl von Einzelstudien –, um diese Frage zu klären.

In die Analysen gingen 51 Studien mit 66 Datensätzen und insgesamt 22.106 Teilnehmern ein. Den ersten Teil des Ergebnisses zeigt Tabelle 4, nämlich das Ausmaß der Übereinstimmung der Assessment Center-Beurteilungen mit dem Ergebnis von Intelligenztests.

Der Wert $\rho = .43$ bedeutet: Das Merkmal, das am stärksten in das Assessment Center-Gesamtergebnis eingeht, ist Intelligenz. Nun gehen in das Assessment Center-Gesamtergebnis auch Ergebnisse von Planspielen, Postkörben u. Ä. ein, teilweise sogar die Ergebnisse der Intelligenztests selbst. Dadurch wäre es nicht sehr verwunderlich, wenn im Gesamtergebnis auch die Intelligenz der Teilnehmer zum Ausdruck kommt. Also gilt die zweite Prüfung einer Aufgabe, die in aller Regel nicht zur Messung der Intelligenz angelegt ist, nämlich der Gruppendiskussion.

Der zweite Teil der Tabelle zeigt, dass das bemerkenswerte Ergebnis offenbar auch für das Resultat der Gruppendiskussion gilt. Das heißt, die Assessoren haben die Kandidaten nach ihren geistigen Fähigkeiten beurteilt, obwohl sie hierzu in aller Regel nicht aufgefordert werden. (In wenigen Fällen wird „Ausdrucksfähigkeit“ eingeschätzt. Dass zur Gruppendiskussion nicht nur verbale, sondern auch numerische Intelligenz besonders viel beitragen soll, mag zunächst erstaunen. Es klärt sich aber, wenn man daran denkt, dass innerhalb des g-Faktors der Intelligenz diese beiden Hauptfaktoren am meisten bei-

Tabelle 4: Ergebnisse der Metaanalyse für Intelligenztests (Scholz & Schuler, 1993, S. 77)

	N	Zahl unabhängiger Stichproben	*r*	ρ	SD	90 % Konfidenzintervall (artefaktkorrigiert)	Varianzaufklärung durch Artefakte in %
Assessment Center (Overall Assessment Rating)							
Allgemeine Intelligenz	17.373	28	.33	.43	.12	.24–.62	16.73
Numerische Intelligenz	11.525	17	.28	.36	.08	.24–.48	30.78
Verbale Intelligenz	12.957	22	.30	.40	.08	.26–.53	30.00
Gruppendiskussion							
Allgemeine Intelligenz	1.591	12	.32	.46	.11	.28–.64	62.70
Numerische Intelligenz	758	8	.21	.30	.17	.03–.57	46.26
Verbale Intelligenz	1.323	12	.28	.41	.08	.27–.55	75.53

tragen und mittelhoch korreliert sind. Assessment Center und Gruppendiskussion erfassen offenbar diesen g-Faktor, also allgemeine Intelligenz.)

Welche nicht kognitiven Persönlichkeitsmerkmale sind ausschlaggebend für das Abschneiden im Assessment Center? Das Ausmaß der Übereinstimmung der Beurteilungen mit dem Ergebnis von Persönlichkeitstests gibt Tabelle 5 wieder.

Die große Bedeutung kognitiver Fähigkeiten für den Eindruck der Beurteiler im Assessment Center kommt auch dann zum Ausdruck, wenn man nicht, wie bei Scholz und Schuler (1993), die Beziehung zwischen Testergebnissen und OAR ermittelt, sondern die prognostische Validität von Beurteilungsdimensionen prüft: In einer metaanalytischen Auswertung dieser Fragestellung kommen Arthur, Day, McNelly und Edens (2003) auf einen korrigierten Validitätskoeffizienten von $\rho = .39$ für die Dimension „Problem solving“ (was einer Einschätzung der allgemeinen Intelligenz durch die Beurteiler nahekommt). Die weiteren Urteilsdimensionen konnten diesem Wert nur wenig inkrementelle Validität hinzufügen (den einzigen nennenswerten Beitrag leistete die Dimension „Influencing others“ mit 3 % zusätzlich aufgeklärter Kriterienvarianz).

Nicht die breiten Persönlichkeitsmerkmale spiegeln sich im Gesamtergebnis wider, die faktorenanalytisch bestimmten Globalmerkmale Extraversion, Neurotizismus etc., sondern v. a. einige enger definierte Eigenschaften, nämlich Dominanz, Leistungsmotivation, soziale Kompetenz und Selbstvertrauen. Diese Aspekte scheinen den Assessoren aufzufallen, und sie sind es, von denen das Abschneiden im Assessment Center abhängt.

Interessanterweise gehören alle diese Merkmale zu jenen, die sich als allgemein berufserfolgsrelevante Eigenschaften herausgestellt haben, die also relativ unabhängig von den spezifischen Anforderungen von Nutzen sind, sowohl Intelligenz als auch die

Tabelle 5: Ergebnisse der Metaanalyse für Persönlichkeitstests (Scholz & Schuler, 1993, S. 79)

	N	Zahl unabhängiger Stichproben	*r*	ρ	SD	90 % Konfidenzintervall (artefaktkorrigiert)	Varianzaufklärung durch Artefakte in %
Assessment Center (Overall Assessment Rating)							
Neurotizismus	909	8	–.12	–.15	.00	–.15 – –.15	100,00
Extraversion	1.328	10	.10	.14	.13	–.08 – .36	43,40
Offenheit	631	5	.07	.09	.08	–.04 – .22	69,62
Verträglichkeit	871	7	–.05	–.07	.00	–.07 – –.07	100,00
Gewissenhaftigkeit	494	4	–.05	–.06	.00	–.06 – –.06	100,00
Maskulinität	335	2	.09	.12	.00	.12 – .12	100,00
Dominanz	909	8	.23	.30	.06	.21 – .40	82,39
Locus of Control	176	2	.12	.16	.00	.16 – .16	100,00
Leistungsmotivation	613	5	.30	.40	.14	.18 – .63	41,91
Soziale Kompetenz	572	7	.31	.41	.17	.13 – .70	39,02
Selbstvertrauen	601	6	.24	.32	.08	.18 – .45	71,67

genannten nicht kognitiven Persönlichkeitsmerkmale (Schuler, 1996). Wenn sich also die Assessoren nur sehr eingeschränkt an die Beurteilungsvorgaben halten, so könnte es sein, dass sie dabei die prognostische Validität auf ihrer Seite haben.

Die Entdeckung von Sackett und Dreher (1982) war also insofern bemerkenswert, als sie aufzeigte, dass Urteilsprozesse im Assessment Center nicht gemäß den Vorgaben der Verfahrenssteuerer erfolgen. Dieses Ergebnis wurde nachfolgend viele Male bestätigt (vgl. Höft & Bolz, 2004). Die Behauptung allerdings, hiermit einen Mangel an Konstruktvalidität nachzuweisen, ist nur in dem eingeschränkten Sinne richtig, als sich dieser Konstruktbegriff auf die terminologischen Vorgaben der Beobachtungs- oder Beurteilungsformulare bezieht. Deren Angemessenheit steht aber durchaus in Frage: Guldin und Schuler (1997) konnten zeigen, dass konvergente wie diskriminante Validität bei aufgabenrelevanten Merkmalen höher ausfallen als bei weniger relevanten. Bezeichnet man als Konstrukte diejenigen Merkmale (Eigenschaften und Verhaltensweisen), von denen der Berufserfolg – und bei validen Aufgaben also auch das Abschneiden in diesen – abhängt, so besteht keine Veranlassung, mangelnde Konstruktvalidität zu beklagen. Die typischen Assessment Center-Beurteilungsdimensionen sollten jedenfalls nicht mit Eigenschaften gleichgesetzt werden, wie sie in der Persönlichkeitstheorie gemessen werden (Höft & Schuler, 2001).

Die vorherrschende Interpretation unter Assessment Center-Forschern ist allerdings bis heute die der geringen Konstruktvalidität. Dementsprechend gilt ihr methodisches Bemühen Versuchen, durch schärfere Begriffsfassung, engere Aufmerksamkeitsfokussierung der Beobachter (Beschränkung auf wenige Merkmale und/oder Teilnehmer), gründlicheres Training und andere Maßnahmen die Konvergenz und Diskriminanz der Urteile zu verbessern. Die Metaanalyse von Woehr, Arthur und Meriac (in diesem Band) zeigt, dass diesen Versuchen nur sehr eingeschränkt Erfolg beschieden ist. Lance (in diesem Band) plädiert deshalb dafür, den mehrfach – auch schon von Sackett und Dreher (1982) – vorgebrachten Vorschlag ernst zu nehmen und nicht die üblichen Dimensionen, sondern *Aufgaben* oder *Rollen* als Einheiten der Teilnehmerbeurteilung im Assessment Center zu verwenden.

1.5 Die Bedeutung der Multimodalität

Für verschiedene eignungsdiagnostische Verfahrenstypen hat die Forschung der letzten Jahrzehnte sowie die Nutzung der Anwendungserfahrung zu methodischen Verbesserungen geführt. Vor allem das Einstellungsinterview ist dadurch zu einem Auswahlverfahren geworden, das hinsichtlich seiner Validität in Konkurrenz zu den besten anderen Diagnosemethoden treten kann (zusammenfassend Schuler, 2002). Für das Assessment Center gilt dies leider nicht: Die oben beschriebene klassische AT & T-Studie hat mit einem (unkorrigierten!) Validitätswert von $r = .46$ Maßstäbe gesetzt, an denen man sich gern orientiert hat. Demgegenüber zeigte schon die Metaanalyse von Thornton et al. (1987 und in diesem Band) mit dem messfehlerkorrigierten Wert $\rho = .37$ auf, dass die Erwartungen etwas enger gesteckt werden müssen. Hardison und Sackett (in diesem Band) kommen schließlich bei der Analyse neuerer Studien nur noch zu einem Durchschnittswert von $\rho = .26$ (die besser vergleichbaren unkorrigierten Werte lauten .29 und .22). Aktuelle Reliabilitätsschätzungen (z. B. Kleinmann, 1997; Kelbetz & Schuler, 2002) liegen mit $r =$ um .40 erheblich unter den in der Literatur der 70er Jahre genannten von $r =$ um .70. Überdies wird dem Assessment Center-Gesamtwert nur eine minimale inkrementelle Validität ($r = .02$) über Intelligenztests hinaus attestiert (Schmidt & Hunter, 1998); ähnlich gering fiel die inkrementelle Validität simulationsorientierter Einzelaufgaben untereinander in einer Studie von Goldstein, Yusko, Braverman, Brent Smith und Chung (1998) aus. Die verbreitete Behauptung, das Assessment Center weise zwar gute prognostische, aber unzulängliche konstruktbezogene Validität auf, ist also nicht nur im zweiten Aussageteil fragwürdig, sondern, entscheidender noch, auch im ersten.

Wie ist erklärlich, dass ein eignungsdiagnostisches Verfahren nach fünfzig Jahren Forschung und weitester Anwendungserfahrung heute schlechter dasteht als vor einem halben Jahrhundert? Dass die Literatur – insbesondere die für die Praxis verfasste – überreich ist an Innovationsangeboten, die empirischen Daten hingegen immer unzulänglicher ausfallen?

Die Antwort, die hier angeboten werden kann, lautet: Das Assessment Center bleibt aufgrund der methodischen Unzulänglichkeiten, durch die seine Konzeption und Anwendung vielfach gekennzeichnet ist, weit hinter den Möglichkeiten zurück, die dieses Diagnoseverfahren grundsätzlich bietet. Zu den methodischen Unzulänglichkeiten gehört Verschiedenes, was in den nachfolgenden Beiträgen angesprochen wird – Verzicht

auf eine fundierte Anforderungsanalyse, geringe Reliabilität der Einzelverfahren, Missverhältnis der Anzahl von Aufgaben und Beurteilungsdimensionen, unzureichende diagnostische Qualifikation der Beurteiler, Verzicht auf Verfahrensevaluation. (Die Orientierung an den Qualitätsstandards der DIN 33430 ist für Assessment Center ebenso relevant wie für andere eignungsdiagnostische Verfahren. Neuerdings wird von Kersting, 2006, hierfür ein „DIN SCREEN" angeboten.) Verbesserungen dieser Faktoren lassen bereits erheblichen Nutzenszuwachs erwarten.

Das wichtigste Defizit üblicher Assessment Center aber ist ein darüber hinausgehendes, grundsätzliches: die methodische Einseitigkeit der eingesetzten Aufgaben. Präsentationsaufgaben, Gruppendiskussionen und ähnliche Verfahren bieten die Gelegenheit zur Verhaltensbeobachtung, was insbesondere von den als Beobachtern beteiligten Führungskräften und Personalleuten geschätzt wird. Erlaubt diese Beobachtungsgelegenheit ausreichend reliable (d. h. von Fehlereinflüssen nicht stark verzerrte) Einschätzungen und ist sie inhaltsvalide (d. h. entspricht sie den tatsächlichen späteren Tätigkeitsanforderungen), so kann sie *eine* wichtige Quelle des diagnostischen Urteils beisteuern.

Komplexe Phänomene sind aber niemals durch *eine* Informationsquelle allein zu erfassen. Untersucht man beispielsweise in der Astrophysik einen galaktischen Nebel, so wird man durch ein leistungsfähiges Teleskop Aufschlüsse darüber bekommen, dass er aus einer Vielzahl einzelner Sterne, vielleicht auch Doppel- und Dreifach-Sternsystemen besteht. Die Erkenntnismöglichkeit ist allerdings auf den Bereich des sichtbaren Lichts beschränkt. Andere Frequenzen des elektromagnetischen Spektrums – etwa infrarotes Licht, Röntgen- oder Gammastrahlung – liefern ganz andere Information, die in Ergänzung zu dem, was das Fernrohr sichtbar macht, unsere Erkenntnis über das Himmelsobjekt erweitert und vertieft. Die ergänzende Information ist allerdings nur mithilfe zusätzlicher Instrumente zu gewinnen und nicht allein durch angestrengteres Gucken oder durch die Verbesserung optischer Teleskope.

Dieses Prinzip gilt für alle Erkenntnisbereiche. Sucht man hierfür eine philosophisch-methodologische Basis, so findet man sie im erkenntnistheoretischen Multiplizismus (vgl. Schulze & Holling, 2004). Diese Erkenntnistheorie zeigt nicht nur auf, dass wir durch unterschiedliche methodische Zugänge zu einander ergänzenden Einsichten kommen, sondern auch, dass wir Hypothesen so „operational" zu formulieren haben, dass das Messverfahren in ihre Formulierung eingeht. Jedes Messverfahren ist nämlich nicht nur auf einen bestimmten Erkenntnisausschnitt beschränkt, sondern ist auch durch eine besondere Art von Messfehlern charakterisiert.

Ermittelt man beispielsweise die Überzeugungskraft einer Person durch einen Fragebogen zur Selbsteinschätzung, so gehen eine Vielzahl einschlägiger Erlebnisse in den Testwert ein, die sich in ihrer Gesamtheit zu einem Selbstbild geformt haben. Allerdings kann das Bemühen um positive Selbstdarstellung das Ergebnis beeinflussen. Besteht die Messgröße dagegen in der Anzahl der einer Führungskraft zugeordneten Mitarbeiter, so liegt zwar ein von der Selbsteinschätzung relativ unabhängiges Maß vor, dafür ist aber in Rechnung zu stellen, dass dieser Wert zwar durch die Überzeugungskraft einer Person mitbedingt ist, aber zusätzlich andere, von ihr nicht zu verantwortende Einflussgrößen widerspiegelt. Eine dritte Datenquelle könnte ein Verhandlungsrollenspiel sein, bei dem die erreichten Zugeständnisse das Überzeugungsmaß darstellen. Messfehler können in diesem Fall durch das unstandardisierte Verhalten des Rollenspielpartners oder durch die Wahl des Verhandlungsgegenstands in das Ergebnis einfließen.

Die grundsätzliche Bedeutung multimodalen oder multimethodalen Vorgehens in der Personalpsychologie wurde von Schuler und Schmitt (1987) dargelegt, wobei aufgezeigt wurde, dass das Prinzip der Multimodalität nicht nur in der Berufseignungsdiagnostik, sondern auch bei der Leistungsbeurteilung und Personalentwicklung sowie bei der allen diesen Bereichen vorgelagerten Analyse der Tätigkeitsanforderungen zu berücksichtigen ist. In der weiteren Entwicklung wurden drei Modalitäten als maßgebliche Ansätze für die Berufseignungsdiagnostik herausgearbeitet, die durch drei Verfahrenstypen charakterisiert sind:

– eigenschafts- oder konstruktorientierte,
– simulationsorientierte,
– biografieorientierte Verfahren.

Gemeinsam bilden diese Verfahrenszugänge den *trimodalen Ansatz der Berufseignungsdiagnostik* (Schuler, 2000).

Sie unterscheiden sich nicht nur hinsichtlich der Diagnosemethoden, die prototypisch diese Ansätze repräsentieren – Tests, Arbeitsproben und biografische Indikatoren –, sondern auch bezüglich der Validierungslogik, die ihnen innewohnt (siehe Abb. 1): Das korrespondierende Validierungsprinzip von Testverfahren, die primär auf die Messung homogener Konzepte ausgerichtet sind, ist das der *Konstruktvalidierung*. Die primäre Fragestellung lautet also, welche psychologische Bedeutung die Messung bzw. das gemessene Merkmal besitzt. Den Simulationsverfahren entspricht die Inhaltsvalidierung, die zu ermitteln hat, inwieweit Elemente der Berufstätigkeit durch die Aufgabe repräsentiert werden. Das den biografieorientierten Verfahren entsprechende Prüfungsprinzip schließlich ist das der kriterienbezogenen Validierung, im Falle der Prognose operationalisiert als Genauigkeit der Vorhersage eines Kriteriums (Verhalten, Leistung, Zufriedenheit u. Ä.) aufgrund eines Prädiktors (v. a. vergangenes Verhalten und Verhaltensergebnisse wie Zensuren).

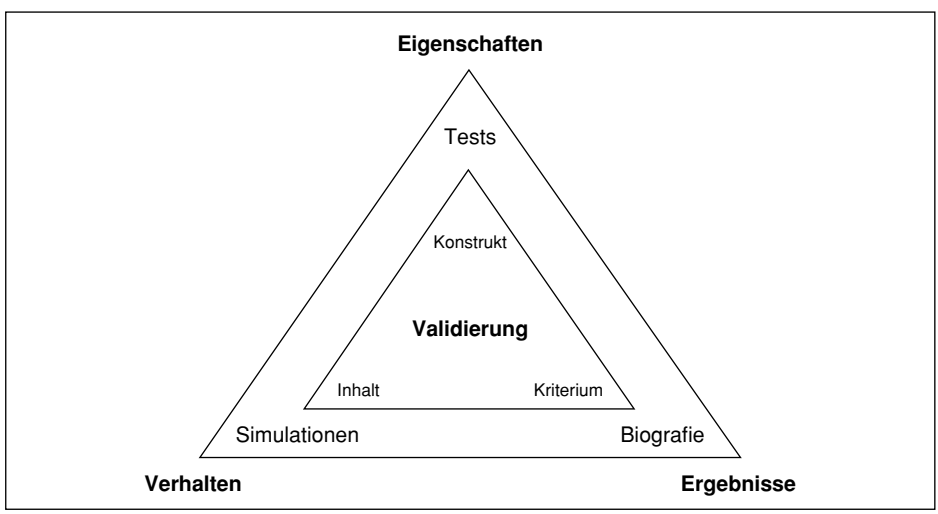

Abbildung 1: Der trimodale Ansatz der Berufseignungsdiagnostik (Abb. aus Schuler & Höft, 2006, S. 103)

Für die praktische Berufseignungsdiagnostik bedeutet dies, dass für komplexere Anforderungskonstellationen zumeist ein multiples Verfahren angemessen ist, das verschiedene Erfassungsmethoden kombiniert.

Das Assessment Center bietet nun die besten Möglichkeiten, die drei diagnostischen Verfahrenstypen zu kombinieren und damit den Erkenntnisbereich zu erweitern, Eindrücke aus der einen Informationsquelle durch andere zu ergänzen sowie einseitige Fehler zu korrigieren.

Was demgegenüber in der gängigen Praxis betrieben wird, ist der gleichermaßen hartnäckige wie aussichtslose Versuch, durch die Vermehrung gleichartiger – ausschließlich simulationsorientierter – Verfahren die Validität zu verbessern. Wie die Arbeit von Goldstein et al. (1998) zeigte, ist die inkrementelle Validität situativer Einzelverfahren untereinander sehr gering, d. h. man stößt rasch an eine Grenze des Informationsgewinns. Der Gesamtwert, das sogenannte OAR (Overall Assessment Rating), ist deshalb nicht valider als eine einzelne gute Arbeitsprobe. Aber selbst hierüber fehlt qualifizierte Forschung, die zeigen würde, an welchen Stellen arbeitsprobenartige Verfahren untereinander ergänzenden Informationsgewinn bieten würden.

Beispiel:

Anfang der 80er Jahre hatte der Verfasser Gelegenheit, im deutschen Tochterunternehmen eines internationalen Konzerns eine Evaluation des dort praktizierten Assessment Centers durchzuführen. Eine der Fragestellungen lautete, welchen Beitrag jede der eingesetzten Aufgaben zum OAR leistet. An jedem der vier Durchführungstage dieses reichhaltigen (also in heutiger Terminologie „multimodalen“) Verfahrens wurde eine Gruppendiskussion durchgeführt. Obwohl die Diskussionsrunden mit wechselnden Beteiligten zusammengesetzt waren, nahm ihr Informationswert rasch ab: Die erste Gruppendiskussion klärte noch 25 % der Ergebnisvarianz im OAR auf, die zweite nur noch 4 %, und schon die dritte Aufgabe dieser Art konnte nur noch einen Beitrag zum Gesamtergebnis leisten, der unter 1 % Varianzaufklärung lag, war also praktisch wertlos. Im Anschluss an die Evaluation konnte das Assessment Center ohne Informationsverlust von vier Tagen Durchführungsdauer auf zwei Tage reduziert werden.

Die Forschung zur inkrementellen Validität eignungsdiagnostischer Verfahren steht erst am Anfang. Ihre Ergebnisse bestätigen aber schon, was theoretisch zu erwarten war: Zusätzlicher Informationsgewinn ist nur zu erzielen, wenn unterschiedliche Verfahrenskomponenten eingesetzt werden. Hierbei stößt die Phänomenologie allerdings an enge Grenzen, denn nach unserem intuitiven Eindruck mag man ja ein dyadisches Rollenspiel und eine Gruppendiskussion für zwei recht unterschiedliche Verfahren halten. Von ihrer Anforderungsstruktur und von der Art unserer sozialen Urteilsbildung her sind sie es allerdings nicht oder nur zu einem geringen Anteil. Deshalb findet man gewöhnlich bei Faktorenanalysen die interaktiven Assessment Center-Aufgaben untereinander annähernd in Höhe ihrer Reliabilität korreliert und auf einem einzigen Faktor vereint.

Wir können also den Schluss ziehen, dass ein Assessment Center, das den investierten Aufwand lohnen soll, Einzelverfahren enthalten sollte, die sich durch Verschiedenartigkeit auszeichnen. Hierbei ist auch daran zu denken, dass die Verschiedenartigkeit

sowohl die Stimuluskomponente (also die Situation, die direkte Anweisung etc.) als auch die Reaktionskomponente (etwa die Aufgabenlösung, die Handlungsweise etc.) berücksichtigen sollte (Funke & Schuler, 1998). Eine Zusammenstellung nachweislich oder mutmaßlich validitätsrelevanter Merkmale eignungsdiagnostischer Verfahren bietet der folgende Kasten (modifiziert aus Schuler, 2000, S. 67).

Unterscheidungsmerkmale eignungsdiagnostischer Verfahren:

- interaktive vs. nicht interaktive Aufgaben
- offenes Verhalten vs. Verhaltensbeschreibungen
- konkretes Verhalten vs. Verhaltenspräferenz
- spezifisch vs. generalisierend
- berufsbezogen vs. berufsfern
- Fremdbeurteilung vs. Selbstbeurteilung
- maximales Verhalten vs. typisches Verhalten
- schriftliche Ausdrucksform vs. mündliche Ausdrucksform
- bildliche Aufgabenvorgabe vs. sprachliche Aufgabenvorgabe
- schriftliche Aufgabenvorgabe vs. mündliche Aufgabenvorgabe
- offene Reaktionsform vs. geschlossene Reaktionsform
- Eigenschaftsansatz – Simulationsansatz – biografischer Ansatz

Noch ist kaum bekannt, welche Merkmalsvariation bei welchen Verfahrenstypen am aussichtsreichen ist und durch welche Kombination die höchste Validität des gesamten Assessment Centers zu erzielen ist. Selbstverständlich können im Einzelfall nicht alle denkbaren Variationen erprobt und eingesetzt werden. Es ist zu vermuten, dass mit dem trimodalen Ansatz „Eigenschaften – Verhalten – Ergebnisse" ein erheblicher Teil der Eignungskomponenten erfasst werden kann. Wenn zusätzlich darauf geachtet wird, innerhalb jeder dieser drei Modalitäten auf Unterschiedlichkeit der eingesetzten Einzelverfahren hinsichtlich der genannten Merkmale (vgl. obigen Kasten) zu achten, kann man zuversichtlich sein, ein gutes, d. h. valides Assessment Center zusammenstellen. Wo konsequent nach diesem Prinzip verfahren wurde, bestätigt jedenfalls das Ergebnis den Ansatz (z. B. Schuler, Funke, Moser & Donat, 1995, oder die Verfahren, deren Gestaltung und Evaluation in den Kapiteln 11 und 15 dieses Bandes beschrieben werden).

Die drei Modalitäten Eigenschaften, Verhalten und Ergebnis sind selbstverständlich nicht unabhängig voneinander, sondern stehen in einer Bedingungsrelation, deren einfachste Vorstellung so aussieht:

Eigenschaften → Verhalten → Ergebnis

„Hinter" einem beobachteten Verhalten (z. B. Verhandeln) steht eine Eigenschaft (z. B. Kontaktfähigkeit). Das Verhandlungsverhalten führt zu einem Ergebnis (z. B. einem Vertrag). Wenn es sich bei dieser Sequenzreaktion um monokausale und deterministische Beziehungen handeln würde, wäre es ausreichend, die Ergebnisse festzustellen, da in ihnen alle Information enthalten wäre. Aus Zeugnissen, Verkaufsergebnissen und hergestellten Produkten ließe sich alles erschließen, was zur Prognose künftigen Verhaltens