Zeitschrift für Personalpsychologie, 4 (4), 159 – 169 © Hogrefe Verlag, Göttingen 2005

Peer-Urteile in einem Assessment Center zur Personalauswahl

Enthalten sie Informationen zum Beurteilten oder zum Beurteiler?

Stefan Höft, Mitra Schümann-Sen und Peter Maschke

Zusammenfassung. Im Rahmen eines Assessment Centers zur Personalauswahl von Nachwuchsflugzeugführern wurde der Informationswert von Peer-Urteilen für die Eignungsdiagnose des Beurteilten und des Beurteilenden an einer Stichprobe von insgesamt 109 Bewerbern untersucht. Hierfür ordneten die Bewerber nach Abschluss der AC-Übungen ihre Mitkandidaten in eine Erfolgsreihenfolge ein (Peer-Ranking) und beurteilten die Leistung von drei Mitkandidaten hinsichtlich der definierten Anforderungsdimensionen (Peer-Rating).
Es ergaben sich generell positive Zusammenhänge zwischen den Peer-Urteilen und den analogen Bewertungen der regulären AC-Kommission mit Werten zwischen r = .16 (Selbstreflexion) und r = .65 (Engagement) bei teilweise sehr geringen Beobachterübereinstimmungen. Die angenommene Beziehung zwischen der Güte der Beurteilung und Personmerkmalen des Beurteilers (allgemeine Intelligenz, eigenes Abschneiden im AC) konnte nicht bestätigt werden. Es wird geschlussfolgert, das Peer-Urteile in dem realisierten Zusammenhang nur wenig differenzierte Informationen zur Person des Beurteilten und keine diagnostisch verwertbaren Informationen zur Person der Beurteilers erfassen.
Schlüsselwörter: Assessment Center, Peer-Beurteilung, Personalauswahl, Verkehrsflugzeugführer

Peer evaluations in an assessment center for personnel selection: Do they contain information indicative for ratees or raters?

Abstract. Within an assessment center for the selection of student pilots at the German Aerospace Center (DLR), a sample of 109 external applicants were subjected to different means of peer evaluation. The applicants each ranked all the participants of the group and rated the performance of three fellow candidates according to defined job requirement dimensions.
Despite low to moderate reliabilities of the peer evaluations, significant relations were found with the evaluation through trained professional observers, ranging from *r* = .16 (Self Awareness) to *r* = .65 (Commitment). The assumed relation between the quality of the evaluation by a candidate and his/her personal features (general intelligence, own assessment center results) could not be confirmed. As a conclusion, peer evaluations in the applied manner seem to deliver only limited information about the rated person and no diagnostically relevant information about the rating person.
Key words: Assessment Center, peer rating, personnel selection, airline pilot

Mit dem Begriff „Peer-Urteil“ wird die Leistungsbeurteilung einer Person durch Gleichgestellte umschrieben (Jeserich, 1995). Im beruflichen Alltag sind dies die Kollegen, im Rahmen einer Personalmaßnahme (z. B. während eines Assessment Centers oder eines Trainings) die übrigen Teilnehmer. Um die unterschiedlichen Anwendungsmöglichkeiten von Peer-Urteilen zu beschreiben, erscheint uns sinnvoll, folgende Randbedingungen für den Einsatz zu berücksichtigen: Urteile von Gleichgestellten können im Rahmen von Personalauswahl- oder von Personalentwicklungsmaßnahmen eingeholt werden. Die Beurteilungen können relativ allgemeine Aussagen zur beurteilten Person treffen (z. B. „Mein Kollege bewahrt einen kühlen Kopf, wenn es im Betrieb mal hoch her geht“) oder sich auf eine genau zu spezifizierende, in der Vergangenheit liegende Situation beziehen („In der Gruppendiskussion hat Teilnehmer X seinen Standpunkt sehr nachdrücklich vertreten“).

Einsatzgebiete von Peer-Beurteilungen

Die Kombination „Generalisierte Peer-Urteile im Rahmen einer Personalentwicklungsmaßnahme“ wird exemplarisch als Bestandteil des so genannten 360-Grad-Feedback-Ansatzes zur Entwicklung von Führungskräften umgesetzt (Neuberger, 2000). Hier dient

DOI: 10.1026/1617-6391.4.4.159

die Beurteilung durch die Kollegen in Kombination mit Urteilen von zugeordneten Mitarbeitern, Vorgesetzten und gegebenenfalls auch Kunden dazu, ein umfassendes individuelles Feedback zu erhalten. Ziel ist es, Abweichungen des Selbstbildes von diesen Fremdbildern aufzudecken und durch eine gezielte Verhaltenssteuerung identifizierte Schwächen zu nivellieren. Die Fremdbeurteilungen erfolgen dabei möglichst verhaltensnah, sind aber zumeist nicht eingeengt auf bestimmte Situationen. Die Urteiler sollen für ihre Bewertung auf ihre individuell gesammelten Erfahrungen mit der Zielperson zurückgreifen.

360-Grad-Feedbacks erfreuen sich in der betrieblichen Praxis zunehmender Beliebtheit (vgl. z. B. Hell, Boramir, Schaar & Schuler, in Druck). Auch die Wissenschaft hat diese Anwendungen zwischenzeitlich als Forschungsgebiet für sich erschlossen (z. B. Cheung, 1999; Mount, Judge, Scullen, Sytsma & Hezlett, 1998; Valle & Bozeman, 2002).

Die Metaanalyse von Conway und Huffcut (1997) zeigt eine im Vergleich zur Vorgesetztenbeurteilung geringe Reliabilität der Peer-Urteile von $r = .37$ (durchschnittliche Beurteilerübereinstimmung bezogen auf einen Beurteiler; ähnliche Werte ermitteln auch Viswesvaran, Ones & Schmidt, 1996). Gleichzeitig ergibt sich ein durchschnittlicher Zusammenhang zwischen Peer-Urteilen und Vorgesetztenbeurteilungen von $r = .34$ (reliabilitätskorrigiert $r_{korr} = .79$).

Anders als beim 360-Grad-Feedback sind Peer-Urteile im Assessment Center-Bereich als spezifische Bewertungen konzipiert, da die Teilnehmer aufgefordert werden, ein Urteil zu Mitkandidaten abzugeben, die sie nur aus dem laufenden Verfahren heraus kennen. Sie können sich bei ihrer Bewertung also nicht auf über mehrere Jahre gesammelte Erfahrungen stützen, sondern müssen die Informationen verwenden, die sie während der Übungen (z. B. während einer Gruppendiskussion) und in den Pausenzeiten zu dem Mitkandidaten gesammelt haben.

In Assessment Centern zur Personalentwicklung werden Peer-Urteile teilweise systematisch als Feedbackmaßnahme eingesetzt. In der beispielsweise von Papon und v. Rüden (2005) beschriebenen Variante übernehmen die vorher hinsichtlich Feedbackmethoden geschulten AC-Teilnehmer selbst die Rolle der Beobachter und geben sich gegenseitig Feedback über ihre Leistung in den abgelaufenen Übungen. Peer-Urteile sind hier also die alleinige Grundlage für das Feedback und werden in sehr vertrauter und geschützter Umgebung gegeben. Die in der wissenschaftlichen Literatur zum Thema „Assessment Center und Peer-Urteile“ zu findenden Studien (vgl. Shore, Shore & Thornton, 1992, p. 43) sind allerdings mehrheitlich dem Potenzialanalyse-Bereich zuzuordnen.

Ein Zwischenfazit zur Nützlichkeit von Peer-Urteilen im AC konnten Thornton, Gaugler, Rosenthal und Bentson bereits 1987 in ihrer Metaanalyse zur kriteriumsbezogenen Validtät von ACs ziehen. Danach wiesen ACs, die zusätzlich Peer-Urteile bei der Urteilsfindung heranzogen, eine bedeutsam höhere Validität auf als ACs ohne diese Beurteilungen. Thornton et al. sprechen sich deshalb nachdrücklich für die Berücksichtigung dieser Zusatzinformationen aus.

In neuerer Zeit (seit 1987) scheint das wissenschaftliche Interesse aber eher nachgelassen zu haben. So fanden wir in einer Literaturrecherche nur zwei Studien, die sich explizit mit dem diagnostischen Wert von Peer-Urteilen im Assessment Center beschäftigten. (Weitere Studien, in denen Peer-Urteile als Randaspekte untersucht wurden, nennen Lievens & Klimoski, 2001.)

Die Studie von Zazanis, Zaccaro und Kilcullen (2001) ist im militärischen Bereich angesiedelt (Auswahl von Spezialkräften mit Hilfe eines dreiwöchigen Outdoor-Assessment Centers) und nur schwer auf den üblichen Assessment Center-Kontext übertragbar. Interessanter ist die Studie von Shore et al. (1992). Hier wurden Peer-Rankings (nach jeder gemeinsam durchgeführten Übung erstellten die Teilnehmer eine Rangreihe zum erfolgreichen Abschneiden der beteiligten Kandidaten) und Peer-Nominationen (am Ende des ACs benannten die AC-Teilnehmer bezogen auf sechs unterschiedliche AC-Dimensionen die ihres Erachtens drei besten Kandidaten) eingesetzt. Sie zeigten deutliche Zusammenhänge mit konzeptionell ähnlich gelagerten kognitiven Fähigkeits- und Persönlichkeitsskalen sowie den analogen Bewertungen der AC-Kommission. Zusätzlich erzielten die Peer-Beurteilungen inkrementelle Validität bei der Vorhersage der fünf bis zehn Jahre später erfassten Karriereentwicklung der Teilnehmer.

Zusammenfassend lässt sich feststellen, dass Peer-Urteile sich als Feedbackinstrument im Rahmen der Personalentwicklung etabliert haben. Sie scheinen bei einer vergleichsweise geringen Reliabilität (erfasst über Beobachterübereinstimmungskoeffizienten) prinzipiell valide Informationen zur beurteilten Person zu enthalten. Anwendungen im Assessment Center-Bereich mit spezifischen Peer-Urteilen zur gezeigten AC-Leistung der Mitkandidaten deuten darauf hin, dass Peer-Urteile inkrementelle Validität gegenüber den Beurteilungen der regulären AC-Kommission aufweisen. Dies wird in der Literatur regelmäßig mit dem Informationsvorsprung erklärt, den die Peers durch informelle Kontakte abseits der regulären AC-Übungen erhalten.

Zielsetzung und Hypothesen der Studie

Zielsetzung

Das am Deutschen Zentrum für Luft- und Raumfahrt e. V. (DLR) durchgeführte Assessment Center stellt im Vergleich zu den berichteten Studien eine Besonderheit dar, da es alleinig für die Personalauswahl externer Bewerber konzipiert ist. Zwar wird in jeder AC-Begrüßung betont, dass keine Quotenauswahl („die vier besten des heutigen ACs werden genommen") betrieben wird. Trotzdem wird die Auswahl von den Bewerbern vielfach als Wettbewerbssituation empfunden. Insoweit stellt sich die Frage, ob die allgemeinen Befunde zu Peer-Urteilen in diesem Zusammenhang repliziert werden können. Möglich wäre beispielsweise, dass die erhobenen Urteile besonders anfällig sind für validitätsmindernde Sympathie-Antipathie-Einflüsse und mögliche Interessenkonflikte.

Konzeptionell knüpft die Studie an eine DLR-Untersuchung von Damitz, Manzey, Kleinmann und Severin (2003) an. Hier wurden die AC-Ergebnisse der erfolgreichen Teilnehmer mit analogen Peer-Urteilen verglichen, die 20 Monate später von Ausbildungskollegen abgegeben wurden. Die höchsten Zusammenhänge ergaben sich für die Einzeldimensionen Engagement ($r_{korr} = .49$) und Konfliktbewältigung ($r_{korr} = .51$). Das AC-Gesamturteil korrelierte zu $r_{korr} = .38$ mit dem aggregierten Peer-Urteil. Für uns stellte sich die Frage, ob die simultan zum AC erhobenen Peer-Urteile möglicherweise ähnliche Zusammenhänge aufweisen wie die Urteile der Ausbildungskollegen.

Schließlich soll noch ein weiterer Aspekt untersucht werden, der unseres Wissens bisher noch nicht im Kontext von AC-Peer-Urteilen analysiert wurde. Möglicherweise lässt die Qualität der Peer-Bewertungen Rückschlüsse auf die Kompetenz des beurteilenden Peers zu. Plausibel erscheint beispielsweise, dass Teilnehmer mit hohen Ausprägungen in den diagnostizierten AC-Dimensionen auch in der sozialen Urteilsbildung besser sind (vgl. hierzu beispielsweise die allgemeineren Studien von Bernardin & Orban, 1990, und Harris, 1994). Angesichts des deutlichen Zusammenhangs von Allgemeiner Intelligenz mit dem Abschneiden im AC (vgl. Höft & Bolz, 2004) sollte auch diese Personvariable nicht unberücksichtigt bleiben.

Hypothesen

Im Mittelpunkt der berichteten Analysen steht die Frage, inwieweit Peer-Urteile, die im Rahmen eines Assessment Centers zur Personalauswahl erhoben wurden, diagnostisch verwertbare Informationen enthalten. Die erste Hypothese konzentriert sich auf die *Informationen bezüglich des Beurteilten*. Ausgehend von der berichteten Befundlage nehmen wir an:

Hypothese 1: Es besteht ein generell positiver Zusammenhang zwischen den Peer-Urteilen und den Bewertungen der regulären AC-Auswahlkommission (1a). Ausgehend von den metaanalytischen Befunden nehmen wir allerdings an, dass die Peer-Urteile weniger reliabel sind als die Kommissionsurteile (1b).

In der zweiten Hypothese wird untersucht, inwieweit die Peer-Urteile *Informationen bezüglich des Beurteilenden* enthalten. Die Hypothese lautet folgendermaßen:

Hypothese 2: Die Güte der Peer-Urteile steht im Zusammenhang mit diagnostisch relevanten Personmerkmalen (allgemeine Intelligenz, eigenes Abschneiden im Assessment Center) des jeweils Beurteilenden.

Methode

Stichprobe

Als Analysegrundlage dient eine Bewerberstichprobe, die im September und Oktober 2001 die Hauptuntersuchung im Auswahlverfahren für Nachwuchsflugzeugführer des Deutschen Zentrums für Luft- und Raumfahrt e. V. (DLR), Abteilung Luft- und Raumfahrtpsychologie durchlaufen hat. Auftraggeber waren die Flugbetriebe der Deutschen Lufthansa AG (DLH).

Insgesamt nahmen 109 Bewerber an der Untersuchung teil, davon waren 94 Personen männlich (86.2 %). Das Durchschnittsalter betrug 22.3 Jahre (SD = 3.1 Jahre). Alle Bewerber verfügten über die allgemeine oder fachgebundene Hochschulreife. Die Bewerber verteilten sich auf insgesamt 13 Assessment Center-Gruppen mit Teilnehmerzahlen zwischen fünf und zehn Personen. Bedingt durch teilweise fehlende Werte basieren die berichteten Analysen auf Daten von 105 Personen.

DLR-Auswahlverfahren für Nachwuchsflugzeugführer

Das DLR-Auswahlverfahren ist sequentiell aufgebaut mit mehreren Zwischenauswahlphasen. In der Vorauswahlstufe werden kognitive Verfahren zu Basisfähigkeiten (z. B. Arbeitsgedächtnis, Raumvorstellung, Wahrnehmungsgeschwindigkeit usw.) zusammen mit

anforderungsrelevanten Wissenstests (Mathematik, Physik, Englisch), einem luftfahrtspezifisch konstruierten Persönlichkeitsverfahren und luftfahrtbezogenen Psychomotorik- und Mehrfacharbeitstests computergestützt durchgeführt. Diagnostizierte Schwächen in einem der leistungsbezogenen Merkmalsbereiche führen zum Ausschluss.

Bewerber, die ihre Grundeignung in der ersten Stufe nachgewiesen haben (ca. 30%), werden zur zweitägigen Hauptuntersuchung eingeladen. Am ersten Tag findet neben einem Geräteteamtest („Dyadic Cooperation Test", vgl. Stelling, 2001) ein Assessment Center statt, in das die vorliegende Studie eingebettet war. Nach einer erneuten Zwischenauswahl folgt dann am zweiten Tag eine Einzeltestung an einem flugsimulatorähnlichen Testgerät, an die sich im Positivfall ein abschließendes hypothesengeleitetes Auswahlgespräch anschließt.

Eingesetzte Verfahren

Im Rahmen der berichteten Analysen wird die *Allgemeine Intelligenz* der Bewerber herangezogen. Sie wurde mit Hilfe von sechs DLR-Testverfahren aus der Vorauswahlphase ermittelt, die während des Erhebungszeitraums ohne Unterbrechung eingesetzt wurden und einen guten Querschnitt zu intelligenzbezogenen Inhalten repräsentieren: ein Test zum technischen Verständnis, ein Englischtest, ein akustischer Merkfähigkeitstest, ein optischer Wahrnehmungstest, ein Test zum räumlichen Vorstellungsvermögen sowie ein optischer Aufmerksamkeitstest mit Merkfähigkeitsanteilen. Obwohl die im Rahmen der Studie untersuchten Bewerber bereits die Vorauswahlstufe positiv durchlaufen hatten und somit hinsichtlich Intelligenz vorselegiert waren, ergab eine gemeinsame Hauptkomponentenanalyse der sechs Verfahren einen dominanten ersten Faktor (44,9% Varianzaufklärung). Für jeden Probanden wurde ein Faktorwert ermittelt, der einen guten Indikator für Allgemeine Intelligenz darstellen sollte.

Das *Assessment Center* wird mit maximal zehn Bewerbern gleichzeitig durchgeführt und besteht aus drei interaktiven Verfahren: einem Rollenspiel sowie einer planungsorientierten und einer konfliktorientierten Gruppendiskussion mit jeweils drei bis fünf Diskussionsteilnehmern. Der Beobachterrotationsplan gewährleistet, dass jeder Teilnehmer in jedem Verfahren unabhängig von zwei variierenden Beurteilern (trainierte DLR-Psychologen und DLH-Flugkapitäne) bewertet wird. In jedem Verfahren werden fünf, insgesamt sieben Anforderungsdimensionen erfasst. Jede Dimension wird mindestens zweimal erhoben. Zur Urteilskommunikation werden die Dimensionen in zwei Merkmalsbereiche gruppiert („Soziale Kompetenz" und „Handlungskompetenz"). Eine Kurzbeschreibung zu den Dimensionen ist in Tabelle 1 wiedergegeben. Die frei protokollierten Verhaltensbeobachtungen werden nach Abschluss der jeweiligen Übung von den Beurteilern anhand von Verhaltensankern auf strukturierten Bewertungsbögen zu Merkmalsbeurteilungen auf sechsstufigen Skalen verdichtet. Eine Diskussion der Bewertungen findet erst am Ende des Tages im Rahmen der Beobachterkonferenz statt. Einen Überblick zu allen bis dato zum Assessment Center des DLR vorliegenden Validitätsbefunde geben Höft und Pecena (2004).

Zur *Peer-Beurteilung* wurden zwei Herangehensweisen eingesetzt: Bei dem *Peer-Ranking* erstellten die Teilnehmer eine Erfolgs-Rangliste aller Teilnehmer in absteigender Reihenfolge. Bei dem *Peer-Rating* beurteilten die Peers die Leistung des jeweiligen AC-Teilnehmers separat für jede Anforderung auf einer sechsstufigen Skala. Zur einheitlichen Orientierung wird jedes Merkmal mit Hilfe einer Kurzcharakterisierung genauer umschrieben (z. B. zu Konfliktbewältigung: „Die Bereitschaft und Fähigkeit, Konflikte zu bemerken und sie zu verstehen, Auseinandersetzungen konstruktiv zu gestalten, Lösungen zu finden, die alle Beteiligten mittragen").

Durchführung

Die Bewerber hatten bereits im Vorfeld der Untersuchung die Möglichkeit, sich mit Hilfe der Internetseite des DLR (http://www.hh.dlr.de) zum Ablauf der Untersuchung und zu den erfassten Anforderungsdimensionen zu informieren. Alle Beurteilungsmerkmale wurden zusätzlich noch einmal bei der Begrüßung erläutert. Die Bewerber wurden darüber informiert, dass sie einige ihrer Mitbewerber beurteilen würden. Es wurde ihnen jedoch nicht gesagt, um welche Personen es sich dabei jeweils handeln würde, um mögliche Bündnisse unter den Bewerbern auszuschließen.

Die Bewerber führten die Peer-Beurteilungen (Peer-Ranking sowie ein Peer-Rating für drei vorher festgelegte Gruppenmitglieder) nach dem Abschluss der Assessment Center-Übungen aus. Durch ein vorgegebenes Schema der Peerzuordnungen wurde gewährleistet, dass nur solche Peers eingeschätzt wurden, mit denen die Beurteiler vorher mindestens eine Übung gemeinsam absolviert hatten. Zum Zeitpunkt der Bearbeitung mussten die Bewerber davon ausgehen, dass ihre Bewertungen mit zur Entscheidungsfindung herangezogen werden können.

Peer-Ranking und Peer-Rating waren in eine größere schriftliche Befragung integriert. Unter anderem

Tabelle 1. Die im AC erfassten Anforderungsdimensionen im Überblick

Titel	Kurzbeschreibung
Merkmalsbereich Soziale Kompetenz (SK)	*Umgang mit anderen sowie die Wahrnehmung der eigenen Person („Ich und Andere")*
Kooperation (KO)	Informationsaustausch und gegenseitige Unterstützung bei der Zusammenarbeit im Team
Konfliktbewältigung (KF)	Meinungsverschiedenheiten konstruktiv gestalten
Empathie (EM)	Interesse und Verständnis für die Sichtweise und die Beweggründe anderer zeigen
Selbstreflexion (SR)	Die eigenen Stärken und Schwächen realistisch einschätzen
Merkmalsbereich Handlungskompetenz (HK)	*Beitrag des Einzelnen zum Problemlöseprozess der Gruppe („Ich und die Aufgabe")*
Engagement (EG)	Intensive Beteiligung am Gruppenprozess durch Fragen, Wortbeiträge oder Ideen
Flexibilität (FL)	Sich rasch auf neue Situationen einstellen und neue Informationen in das Handeln einbeziehen
Belastbarkeit (BE)	Konzentrierte Aufgabenbewältigung trotz ungünstiger Rahmenbedingungen (Stress, Zeitdruck usw.)

mussten einzelne Verhaltensaspekte der Peers und das eigene Abschneiden sowie die einzelnen absolvierten Übungen hinsichtlich unterschiedlicher Kriterien (Unterhaltungswert, Akzeptanz usw.) eingestuft werden. Auf die Ergebnisse zu diesen Zusatzfragen wird in diesem Bericht nicht eingegangen.

Ergebnisse

Die aufgestellten Zusammenhangshypothesen werden mit unterschiedlichen korrelationsstatistischen Analysen geprüft. Soweit es notwendig erscheint, wird im Folgenden ausführlicher auf die Herleitung der eingesetzten Koeffizienten eingegangen.

Peer-Urteile als Informationsquelle zum Beurteilten

Analyse der Peer-Rankings

Um die Peer-Rankings aus den unterschiedlichen Assessment Center-Gruppen zu integrieren, wurde eine Aggregationsstrategie gewählt, die auch bei der metaanalytischen Validitätsgeneralisierung (vgl. z. B. Hunter & Schmidt, 2004) Verwendung findet. Hierfür wurde zunächst für jede beurteilte Person ein mittleres Peer-Urteil aus den vorliegenden Rangplatzierungen ermittelt und anhand dieses mittleren Rangplatzes eine kombinierte Rangreihe der Teilnehmer erstellt. Parallel dazu wurde aus den aggregierten AC-Beurteilungen eine Rangreihe der AC-Teilnehmer gebildet und mit der Peer-Rangabfolge korreliert (Spearmans Rho = r_s). Insgesamt liegen somit aus 13 durchgeführten Assessment Centern 13 Korrelationskoeffizienten zur Ähnlichkeit der Rangreihen vor. Wie bei einer Metaanalyse werden diese Koeffizienten als unabhängige Realisierungen der Hypothesenprüfung interpretiert und stichprobengewichtet aggregiert. Es resultiert eine stichprobengewichtete Effektschätzung von $r_s = .46$ ($p < .01$), basierend auf $k = 13$ Einzelstudien mit insgesamt $n = 105$ Probanden.

Analyse der Peer-Ratings

In Tabelle 2 ist die Korrelationsmatrix zwischen den Peer-Ratings und den dimensionsspezifisch über die Übungen hinweg aggregierten AC-Beurteilungen dargestellt. Ergänzend sind die Zusammenhänge der (durch Aggregation der zugehörigen Dimensionswerte ermittelten) Merkmalsbereiche Soziale Kompetenz und Handlungskompetenz sowie das AC-Gesamtergebnis („overall assessment center rating" = OAR) angegeben. Auf der Hauptdiagonalen sind zudem in Klammern die Beobachterübereinstimmungen in Form von Intraklassenkorrelationskoeffizienten (vgl. Shrout & Fleiss, 1979) eingetragen.

Die Intraklassenkorrelationskoeffizienten (angegeben sind die auf das aggregierte Peer-Urteil bezogenen ICC(1,3)-Koeffizienten) weisen zunächst bei den Peer-Ratings sehr niedrige Übereinstimmungen bei den meisten Dimensionen aus. Nur Engagement und mit deutlichen Abstrichen Konfliktbewältigung, Handlungskompetenz und der AC-Gesamtwert (in

Tabelle 2. Interkorrelation der aggregierten Peer- und Kommissions-Ratings

		Peer-Ratings										Kommissions-Ratings									
		KO	**KF**	**EM**	**SR**	**EG**	**FL**	**BE**	**SK**	**HK**	**OAR**	**KO**	**KF**	**EM**	**SR**	**EG**	**FL**	**BE**	**SK**	**HK**	**OAR**
Peer-Ratings	**KO**	(.16)																			
	KF	**.69**	(.47)																		
	EM	**.61**	**.59**	(.12)																	
	SR	**.35**	**.31**	**.42**	(.20)																
	EG	**.34**	**.55**	.16	.04	(.62)															
	FL	**.58**	**.62**	**.46**	*.20*	**.54**	(.23)														
	BE	**.31**	**.54**	**.26**	*.24*	**.54**	**.57**	(.19)													
	SK	**.85**	**.84**	**.83**	**.63**	**.36**	**.60**	**.44**	(.32)												
	HK	**.48**	**.68**	**.33**	.17	**.88**	**.81**	**.81**	**.54**	(.49)											
	OAR	**.77**	**.86**	**.66**	**.46**	**.69**	**.79**	**.69**	**.89**	**.86**	(.43)										
Kommissions-Ratings	**KO**	**.26**	**.33**	**.27**	.04	.16	*.20*	.03	**.29**	.16	**.26**	(.61)									
	KF	**.28**	**.39**	**.31**	.08	**.39**	**.33**	*.23*	**.34**	**.39**	**.42**	**.71**	(.72)								
	EM	**.26**	**.28**	**.28**	**.22**	.13	.17	.10	**.33**	.16	**.29**	**.65**	**.69**	(.79)							
	SR	*.20*	**.31**	**.31**	.16	**.32**	**.30**	.17	**.31**	**.32**	**.37**	**.48**	**.68**	**.56**	(.67)						
	EG	**.30**	**.55**	*.21*	**.06**	**.64**	**.44**	**.35**	**.37**	**.60**	**.52**	**.42**	**.68**	**.36**	**.50**	(.87)					
	FL	*.16*	**.26**	*.24*	.01	**.33**	**.34**	*.22*	*.22*	**.36**	**.33**	**.44**	**.73**	**.57**	**.57**	**.52**	(.74)				
	BE	**.25**	**.30**	**.28**	**.25**	*.22*	*.20*	.19	**.34**	*.25*	**.34**	**.39**	**.52**	**.37**	**.50**	**.37**	**.40**	(.76)			
	SK	**.30**	**.38**	**.34**	.15	**.29**	**.29**	.15	**.38**	**.29**	**.39**	**.84**	**.90**	**.87**	**.79**	**.57**	**.68**	**.52**	(.80)		
	HK	**.30**	**.48**	**.31**	.13	**.52**	**.42**	**.33**	**.40**	**.52**	**.52**	**.53**	**.82**	**.55**	**.66**	**.82**	**.81**	**.73**	**.75**	(.86)	
	OAR	**.33**	**.46**	**.34**	.14	**.44**	**.38**	**.23**	**.41**	**.43**	**.48**	**.75**	**.92**	**.76**	**.79**	**.73**	**.78**	**.65**	**.95**	**.92**	(.85)

Anmerkungen: vgl. Tabelle 1 für eine Erläuterung der Dimensionsabkürzungen. Soziale Kompetenz (SK) ist ein Aggregat aus: Kooperation (KO), Konfliktbewältigung (KF), Empathie (EM), Selbstreflexion (SR). Handlungskompetenz (HK) ist ein Aggregat aus: Engagement (EG), Flexibilität (FL), Belastbarkeit (BE). Der AC-Gesamtwert (OAR) ist ein Aggregat aus SK und HK. N = 105; alle Korrelationen $> .19$ sind signifikant mit $p < .05$ und *kursiv* markiert, alle Korrelationen $> .25$ sind signifikant mit $p < .01$ und **fett** markiert. In der Hauptdiagonalen sind die Beobachterübereinstimmungen in Form von Intraklassenkorrelationskoeffizienten angegeben (Peer: ICC (1,3); Kommission: ICC (1,2)).

den letzteren beiden Aggregatwerten ist Engagement als Summand enthalten) zeigen Werte jenseits .40. Unter anderem bedingt durch die übungsübergreifende Aggregation fallen die Übereinstimmungskoeffizienten bei den Kommissions-Ratings deutlich besser aus. Auch hier nimmt Engagement mit ICC(1,2) = .87 einen Spitzenplatz ein.

Angesichts der schlechten Beobachterübereinstimmungen sind die Konvergenzen zwischen den Peer- und Kommissions-Ratings fast erstaunlich. Die jeweils zusammengehörigen Dimensionsbewertungen (im Sinne der Nomenklatur von Campbell & Fiske, 1959, die Monotrait-Heteromethod-Koeffizienten) sind grau unterlegt. Alle Zusammenhänge sind positiv und bis auf die Dimensionen Selbstreflexion und Belastbarkeit auch statistisch bedeutsam ($p < .01$). Bei allen Dimensionen außer Engagement gibt es allerdings immer Heterotrait-Heteromethod-Koeffizienten, die größer ausfallen. Bevorzugt sind dies die Zusammenhänge mit Engagement (bei vier der sechs anderen Dimensionen), gefolgt von Konfliktbewältigung (zwei Dimensionen). Bei den Heterotrait-Monomethod-Koeffizienten ergeben sich bei den Peer-Ratings mehrheitlich Koeffizienten zwischen .3 und .6, auffällig ist hier der konsistent geringere Zusammenhang des Selbstreflexionsratings mit den übrigen Beurteilungen. Die Heterotrait-Monomethod-Koeffizienten bei den Kommissionsratings fallen homogen positiv mit Werten zwischen .36 und .71 aus.

Peer-Urteile als Informationsquelle zum Beurteiler

Die Analysen im Rahmen dieses Untersuchungsabschnitts werden in drei Unterabschnitten dargestellt. Zunächst wird überprüft, ob der Grad der Bekanntheit zwischen Beurteiler und Beurteiltem einen Einfluss auf die Konvergenz mit dem Kommissions-Rating hat. Danach werden die Abweichungen zwischen den Peer-Urteilen auf beurteilerspezifische Einflüsse untersucht. Im letzten Unterabschnitt wird dann die Akkuratheit der individuellen Urteile analysiert, indem ein direkter Bezug zum Referenzurteil der Kommission hergestellt wird.

Analyse der unterschiedlichen Bekanntheit zwischen Beurteiler und Beurteiltem

Wie bereits erwähnt, wurde durch einen genauen Zuordnungsplan der Peer-Beurteilungen gewährleistet, dass alle Beurteiler den Beurteilten in mindestens einer Übung erlebt haben. Gleichzeitig ist damit auch geregelt, dass sich die Wartezeiten der jeweiligen Bewerber deutlich überschneiden und somit der informelle Kontakt in den Pausen erleichtert wird. Innerhalb der Peer-Zuordnungen streut die Anzahl der gemeinsam erlebten Übungen aber deutlich zwischen einer gemeinsamen Übung (bei 32.7% der Peer-Urteile) bis hin zu drei gemeinsamen Übungen (12.8% der Peer-Urteile). Sollte Vertrautheit relevant sein, so sollte innerhalb der Dreiergruppe der beurteilenden Peers die Konvergenz mit dem Kommissionsurteil für die Peers mit dem höchsten Bekanntheitsgrad höher ausfallen als für die Peers mit dem geringsten Bekanntheitsgrad. Diese Annahme musste allerdings verworfen werden, da die Analysen einen anderen Trend zeigten: Bei den untersuchten zehn Konvergenzen erwies sich nur in einem Fall (zur Dimension Kooperation) die Konvergenz bei den bekannteren Peers höher, während in allen anderen Fällen die weniger vertrauten Peers geringfügig höhere Konvergenzen erzielten.

Analyse der Abweichungen zwischen den beurteilerspezifischen Peer-Urteilen

Die im Folgenden berichteten Analysen zu beurteilerspezifischen Einflüssen konzentrieren sich auf die Abweichungen zwischen den individuellen Peer-Urteilen. Die ersten vier Koeffizienten leiten sich aus einem generalisierbarkeitstheoretischen Analyseansatz ab (vgl. Shavelson & Webb, 1991). Für jeden beurteilenden Peer wird die Gesamtvarianz der von ihm abgegebenen Dimensionsurteile bestimmt (σ^2(total)). Diese Gesamtvarianz wird in drei Komponenten aufgeschlüsselt: σ^2(d) gibt den Anteil der von den Beurteilten unabhängigen dimensionsspezifischen Unterschiede in den Urteilen des Peers an. σ^2(r) beschreibt den Varianzanteil, der auf generelle Bewertungsunterschiede zwischen den Beurteilten zurückgeht. σ^2(dr,e) bezeichnet die Restvarianz, die in den spezifischen Dimensionsurteilen zu den Beurteilten verbleibt und nicht durch dimensions- und beurteiltenspezifische Effekte erklärt werden kann. Bewertungen eines guten Peer-Urteilers sollten generell differenziert sein (hohes σ^2(total)), wenig beurteiltenunabhängige Dimensionsunterschiede aufweisen (geringes σ^2(d)), gut zwischen den Beurteilten differenzieren (hohes σ^2(r)) und gleichzeitig dimensionsspezifische Besonderheiten des Beurteilten erfassen (hohes σ^2(dr,e)).

Im linken Teil der Tabelle 3 werden diese Koeffizienten in Verbindung gesetzt zu dem persönlichen Abschneiden des beurteilenden Peers im AC und seiner erfassten Allgemeinen Intelligenz. Ersichtlich ist, dass Peers, die ein gutes AC-Ergebnis erzielen (OAR), eine größere Gesamtvarianz in ihren Peer-Urteilen aufweisen. Sie ist auf einen erhöhten Anteil der „erwünschten“ Varianzkomponenten σ^2(r) und

Tabelle 3. Zusammenhänge zwischen Personvariablen des Peer-Raters und unterschiedlichen Gütekoeffizienten zu seinen Beurteilungen

		Abweichungsanalyse Varianzkomponentenschätzungen					*Akkuratheitsanalyse* Cronbach(1955)-Koeffizienten			
		σ^2(total)	σ^2(d)	σ^2(r)	σ^2(dr,e)	Rater-differenz	D^2	E^2	DE^2	SA^2	DA^2
AC-Ergebnis	**KO**	.11	−.06	.16	.00	.01	.02	.01	.01	−.04	−.02
des Beurteilenden	**KF**	.23*	−.22*	.18	.22*	−.01	.04	−.07	.12	−.06	.15
	EM	.14	−.09	.15	.07	−.04	.13	.11	.05	−.06	.11
	SR	.10	−.16	.09	.13	−.06	.10	.07	.07	−.06	.11
	EG	.35**	−.10	.27**	.23*	.00	.07	−.13	.22*	.04	.19*
	FL	.19*	−.03	.19*	.06	.02	.10	.06	.13	−.01	.00
	BE	.28**	−.10	.20*	.22*	.11	.06	−.03	.10	.08	.17
	SK	.17	−.16	.17	.12	−.03	.08	.04	.07	−.07	.10
	HK	.36**	−.10	.28**	.22*	.05	.10	−.05	.19*	.05	.16
	OAR	.27**	−.14	.24*	.18	.01	.10	.00	.13	−.02	.13
Allgemeine Intelligenz		.02	−.19	.17	−.07	.08	.14	.13	.15	−.16	−.06

Anmerkungen: * Die Korrelation ist auf dem Niveau von .05 (2-seitig) signifikant. ** Die Korrelation ist auf dem Niveau von .01 (2-seitig) signifikant. Vgl. Tabelle 1 für eine Erläuterung der Dimensionsabkürzungen; vgl. Text für eine detaillierte Beschreibung der eingesetzten Gütekoeffizienten.

σ^2(dr,e) sowie auf einen tendenziell erniedrigten σ^2(d)-Anteil zurückzuführen. Eine Detailanalyse zu den in Tabelle 2 nach Einzeldimensionen aufgeschlüsselten Zusammenhängen zeigt, dass die generellen Effekte auch bei den Beurteilern auf Konfliktbewältigung und Engagement zurückzuführen sind. Zusätzlich erweist sich Belastbarkeit als bedeutsam. Zur Allgemeinen Intelligenz ergibt sich keine substanzielle Beziehung.

Der fünfte Koeffizient (vereinfacht als „Raterabweichung" tituliert) erfasst als Variante der Beurteilerübereinstimmungserhebung den quadrierten Abstand des jeweiligen Peer-Urteils vom Mittelwert der beiden konkurrierenden Peer-Urteile. Hohe Werte in diesem Koeffizienten weisen auf Beurteiler hin, die deutlich anders geartete Bewertungen vornehmen als ihre Peers. Der über alle Dimensionen und Beurteilten hinweg gebildete Koeffizient weist keinen Zusammenhang mit Personmerkmalen der Beurteilenden auf. Auch eine (hier nicht explizit aufgeführte) nach Dimensionen und Beurteilten aufgeschlüsselte Analyse zeigt keine relevanten Beziehungen.

Analyse der Akkuratheit der Peer-Urteile

Während sich die Analysen des letzten Abschnitts auf die alleinigen Peer-Urteile stützten, sollen im Folgenden die Beurteilungen der regulären AC-Kommissionen herangezogen und der Abstand der Peer-Urteile zu diesen „wahren" Werten untersucht werden. Bei der Bestimmung dieser Akkuratheit wird auf die Systematik von Cronbach (1955) zurückgegriffen, der insgesamt fünf Koeffizienten bildet. Für eine genauere statistische Herleitung verweisen wir beispielsweise auf Sulsky und Balzer (1988). Die Koeffizienten haben folgende inhaltliche Bedeutung:

- Die *Gesamtdifferenz D^2* beschreibt die durchschnittliche quadrierte Abweichung zwischen jedem einzelnen Peer-Urteil und dem entsprechenden „wahren" Wert der Kommission.
- *Elevation E^2* bildet die quadrierten Abweichungen des mittleren Peer-Urteils zum mittleren Kommissionsurteil ab. Elevation ist damit ein Maß der individuellen Verwendung der Bewertungsskala, da sich darin lediglich ausdrückt, ob ein Beobachter generell zu streng oder zu milde urteilt.
- *Differential Elevation DE^2* berechnet varianzanalytisch gesprochen den Haupteffekt „Beurteilte Person". DE^2 ist folglich ein Maß für die Genauigkeit der Einschätzung der Leistung eines Beurteilten in Relation zu den anderen bewerteten Personen.

- *Stereotype Accuracy SA²* betrachtet den Haupteffekt „Dimension" und bildet die Genauigkeit der Einordnung einer Dimension auf einem Bewertungskontinuum in Relation zu anderen bewerteten Dimensionen ab. Dieser Koeffizient fokussiert damit die Vertrautheit der Beobachter mit einer implizit geteilten dimensionsspezifischen Bewertungsnorm.
- Die Interaktion der beiden Faktoren „Dimension" und „Beurteilte Person" wird durch *Differential Accuracy DA²* abgebildet. Der Koeffizient bildet damit die Fähigkeit des Beurteilers ab, Unterschiede in der Leistung der beurteilten Personen hinsichtlich verschiedener Dimensionen zu erfassen.

Für alle Koeffizienten gilt: Je kleiner der Koeffizient, desto akkurater das Urteil.

Die im rechten Teil der Tabelle 3 dargestellten Zusammenhänge der Cronbach-Koeffizienten mit den Personmerkmalen der Beurteiler zeigen, dass es kaum substanzielle Beziehungen gibt. Einzig bei den besonders engagierten Bewerbern ergibt sich ein signifikanter, vergleichsweise geringer Zusammenhang: Ihr personbezogenes Urteil (DE^2) und ihr personbezogen-dimensionsspezifisches Peer-Urteil (DA^2) weichen eher vom analogen Urteil der AC-Kommission ab. Auch hier zeigt sich kein Zusammenhang mit Allgemeiner Intelligenz.

Diskussion

In der berichteten Studie wurde die Aussagekraft von Peer-Urteilen untersucht, die in einem Assessment Center zur Personalauswahl erhoben wurden. Die Urteile wurden aus zwei unterschiedlichen Perspektiven betrachtet: Im ersten Untersuchungsabschnitt standen die diagnostischen Informationen zur beurteilten Person im Mittelpunkt. Im zweiten Abschnitt wurde der umgekehrte Weg gegangen und analysiert, inwieweit die abgegebenen Urteile diagnostisch verwertbare Informationen zur beurteilenden Person enthalten.

Die erste Hypothese konnte bestätigt werden. Es zeigt sich sowohl beim Peer-Ranking als auch beim Peer-Rating ein generell positiver Zusammenhang mit den Bewertungen der AC-Kommission (Hypothese 1a). Die Peer-Rating-Zusammenhänge fallen allerdings für die beurteilten Dimensionen deutlich unterschiedlich aus. Eine mögliche Ursache hierfür liegt in der heterogenen Reliabilität der dimensionsspezifischen Peer-Urteile (Hypothese 1b). Die Beurteilerübereinstimmungskoeffizienten fallen nur für Engagement und Konfliktbewältigung akzeptabel aus. Dies sind aber genau die Dimensionen, für die sich die höchsten Zusammenhänge mit den analogen Kommissionsbewertungen ergeben (konvergente Validität) und gleichzeitig weniger hohe Zusammenhänge mit Beurteilungen zu anderen Dimensionen auftreten (diskriminanten Validität).

Die plausibelste Erklärung dürfte sein, dass diese Dimensionen auch für untrainierte Beurteiler leicht zu beobachten und zu bewerten sind. Die hohen Zusammenhänge zwischen den beiden Dimensionsbewertungen deuten darauf hin, dass hier im Wesentlichen die durch den AC-Teilnehmer gezeigte Aktivität als Bewertungsanker dient. Sie wird in Engagement direkt abgebildet und ist für eine konstruktive Lösung von Problemen und Konflikten zwingend notwendig. Dimensionen wie Empathie und Selbstreflexion sind deutlich abstrakter und komplexer gestaltet. Ihre Verhaltensindikatoren sind weniger eindeutig identifizierbar und die Urteile hierzu fallen entsprechend unsicher aus.

Interessanterweise decken sich die Konvergenzbefunde mit den Ergebnissen aus der Peer-Studie von Damitz et al. (2003). Auch hier wiesen die Peer-Urteile zu Engagement und Konfliktbewältigung die höchsten Zusammenhänge mit dem 20 Monate vorher erhobenen AC-Urteil auf. Zwar ergaben sich hier höhere Beurteilerübereinstimmungen, die Beurteiler hatten bis auf einen kurzen Hinweis zu möglichen Beobachterfehlern aber ebenfalls kein reguläres Training erhalten. Das dort verwendete Rating war praktisch identisch mit dem in der vorliegenden Studie eingesetzten Verfahren. Somit können die in der damaligen Studie gefundenen differenziellen Befunde möglicherweise ebenfalls teilweise darauf zurückgeführt werden, dass Engagement und Konfliktbewältigung einfacher zu beobachten und zu bewerten waren als beispielsweise die übrigen Dimensionen aus dem Merkmalsbereich Soziale Kompetenz.

Die in Hypothese 2 angenommene Beziehung zwischen der Güte der Beurteilung und Personmerkmalen des Beurteilers konnte nicht bestätigt werden. Zunächst zeigte sich keine Beziehung zwischen dem Bekanntschaftsgrad von Peer und Beurteiltem (operationalisiert über die Anzahl miteinander absolvierter Übungen) und der Konvergenz der Peer-Urteile mit dem Kommissionsurteil. Zwar wies die Abweichungsanalyse darauf hin, dass engagierte, flexible und belastbare AC-Teilnehmer eine breitere Urteilsstreuung aufweisen. Die Akkuratheitsanalyse deutete aber an, dass diese Personen dabei ungenauer beurteilen. Die Allgemeine Intelligenz wies keine substanziellen Beziehungen auf. Auf eine Korrektur der Varianzeinschränkungen wurde deshalb verzichtet.

Einschränkungen im Geltungsanspruch der Studie ergeben sich durch die unterschiedliche Herleitung der Peer-Urteile und der Kommissionsbewertungen. Die Peers bewerten auf der Basis eines Gesamtein-

drucks, während das dimensionale Kommissionsurteil auf aggregierten Beurteilungen beruht, die von verschiedenen Beobachtern in unterschiedlichen Übungen vorgenommen wurden. Der in der Akkuratheitsanalyse der Beurteiler als Bezugspunkt dienende „wahre Wert" weist also durch die Aggregation tendenziell eine geringere Varianz auf als die dazu in Beziehung gesetzten einzelnen Peer-Urteile. Um hier eine bessere Angleichung zu erreichen, hätte beispielsweise das Kommissionsurteil erst am Ende des AC-Tages nach Abschluss aller Verfahren erfolgen können (die so genannte „within dimensions"-Auswertungsvariante; vgl. Silverman, Dalessio, Woods & Johnson, 1986).

Möglicherweise hätten auch höhere Beurteilerübereinstimmungen bei den Peers erzielt werden können, wenn die Dimensionen noch genauer beschrieben und der Bewertungsprozess noch ausführlicher erläutert worden wäre.

Schließlich konnte mangels Außenkriterien keine Aussage zur inkrementellen Validität von Peer-Urteilen getroffen werden. Die vorliegenden Daten lassen unseres Erachtens aber einen solchen Informationszugewinn nicht erwarten. Ein wichtiger Unterschied im Vergleich zur Studie von Shore et al. (1992), in der inkrementelle Validität der Peer-Urteile für die Karriereentwicklung nachgewiesen werden konnte, besteht im Übrigen in der Stichprobencharakteristik: Bei Shore et al. wurden Bewerber aus einem unternehmensinternen Assessment Center befragt. Durch die Größe des Unternehmens (ca. 50000 Mitarbeiter) kannten sie sich vor dem AC zwar nicht persönlich, konnten bei ihren Erfolgsprognosen aber auf ihr implizites Wissen zu im Sinne der Unternehmenskultur erfolgreichem Verhalten zurückgreifen (vgl. Klimoski & Strickland, 1977). Dieses Insiderwissen steht den Pilotenbewerbern nicht zur Verfügung.

Ebenfalls nicht eingegangen wurde auf ethisch-rechtliche Probleme beim Einsatz von Peer-Urteilen im Umfeld von Personalauswahlsituationen. Sie dürften beträchtlich sein.

Unser abschließendes Fazit auf der Grundlage unserer Studienergebnisse lautet: Im Personalentwicklungsbereich erfüllen Peer-Urteile sicherlich eine wichtige Feedbackfunktion. Im Kontext des untersuchten Personalauswahl-ACs lieferten sie allerdings nur wenig differenzierte Informationen zur Person des Beurteilten. Sie beinhalteten anscheinend keine diagnostisch relevanten Informationen zur Person des Beurteilenden. Das DLR wird deshalb zunächst auf eine Integration von Peer-Urteilen in den Auswahlprozess verzichten, bis weiterführende Studien positivere Befunde liefern.

Literatur

Bernardin, H. & Orban, J. A. (1990). Leniency effect as a function of rating format, purpose for appraisal, and rater individual differences. *Journal of Business and Psychology, 5,* 197–211.

Campbell, D. & Fiske, D. W. (1959). Convergent and discriminant validitation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cheung, G. W. (1999). Multifaceted conceptions of self-other ratings agreement. *Personnel Psychology, 52,* 1–36.

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

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

Damitz, M., Manzey, D., Kleinmann, M. & Severin, K. (2003). Assessment center for pilot selection: Construct and criterion validity and the impact of assessor type. *Applied Psychology: An International Review, 52,* 193–212.

Harris, M. M. (1994). Rater motivation in the performance appraisal context: A theoretical framework. *Journal of Management, 20,* 737–756.

Hell, B., Boramir, I., Schaar, H. & Schuler, H. (in Druck). Interne Personalauswahl und Personalentwicklung in deutschen Unternehmen. *Wirtschaftspsychologie*.

Höft, S. & Bolz, C. (2004). Zwei Seiten derselben Medaille? Empirische Überlappungen zwischen Persönlichkeitseigenschaften und Assessment Center-Anforderungsdimensionen. *Zeitschrift für Personalpsychologie, 3,* 6–23.

Höft, S. & Pecena, Y. (2004). Behaviour-oriented assessment for the selection of aviation personnel. In K.-M. Goeters (Ed.), *Aviation psychology: Practice and research* (pp. 153–170). Aldershot: Ashgate.

Hunter, J. & Schmidt, F. L. (2004). *Methods of meta-analysis. Correcting error and bias in research findings*. Thousand Oaks, CA: Sage.

Jeserich, W. (1995). Kollegenurteile. In W. Sarges (Hrsg.), *Management-Diagnostik* (S. 671–676). Göttingen: Hogrefe.

Klimoski, R. & Strickland, W. J. (1977). Assessment centers: Valid or merely prescient. *Personnel Psychology, 30,* 353–363.

Lievens, F. & Klimoski, R. (2001). Understanding the assessment center process: Where are we now? In C. L. Cooper & I. T. Robertson (Eds.), *International review of industrial and organizational psychology (Vol. 16,* pp. 245–286). Chicester: John Wiley.

Mount, M. K., Judge, T. A., Scullen, S. E., Sytsma, M. R. & Hezlett, S. A. (1998). Trait, rater and level effects in 360-degree performance ratings. *Personnel Psychology, 51,* 557–576.

Neuberger, O. (2000). *Das 360°-Feedback: Alle fragen? Alles sehen? Alles sagen?* München: Hampp.

Papon, A. & von Rüden, R. (2005). Assessment Center als Chance, nicht als Schicksal – Ein Beispiel eines Orientation Centers (OC) bei der Lilly Pharma Holding GmbH. In K. Sünderhauf, S. Stumpf, & S. Höft (Hrsg.), *Assessment Center: Von der Auftragsklärung bis zur*

Qualitätssicherung (S. 324–333). Lengerich: Pabst Science Publishers.

Shavelson, R. & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage.

Shore, T. H., Shore, L. & Thornton, G. C. (1992). Construct validity of self- and peer evaluations of performance dimensions in an assessment center. *Journal of Applied Psychology, 77,* 42–54.

Shrout, P. & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86,* 420–428.

Silverman, W. H., Dalessio, A., Woods, S. B. & Johnson, R. L. (1986). Influence of assessment center methods on assessor's ratings. *Personnel Psychology, 39,* 565–578.

Stelling, D. (2001). DCT – Dyadic Cooperation Test. In W.Sarges & H. Wottawa (Hrsg.), *Handbuch wirtschaftspsychologischer Testverfahren* (S. 197–199). Lengerich: Pabst Science Publishers.

Sulsky, L. & Balzer, W. K. (1988). Meaning and measurement of performance rating accuracy: Some methodological and theoretical concerns. *Journal of Applied Psychology, 73,* 497–506.

Thornton, G. C., Gaugler, B. B., Rosenthal, D. B. & Bentson, C. (1987). Die prädiktive Validität des Assessment Centers – eine Metaanalyse. In H. Schuler & W. Stehle (Hrsg.), *Assessment Center als Methode der Personalentwicklung* (S. 36–77). Göttingen: Hogrefe.

Valle, M. & Bozeman, D. (2002). Inter-rater agreement on employee job performance: Review and directions. *Psychological Reports, 90,* 975–985.

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

Zazanis, M. M., Zaccaro, S. J. & Kilcullen, R. N. (2001). Identifying motivation and interpersonal performance using peer evaluations. *Military Psychology, 13,* 73–88.

Eingegangen: 01. 12. 2004
Revision eingegangen: 01. 07. 2005

Dr. Stefan Höft

Deutsches Zentrum für Luft- und Raumfahrt e. V. (DLR)
Abteilung Luft- und Raumfahrtpsychologie
Sportallee 54a
22335 Hamburg
E-Mail: stefan.hoeft@dlr.de