



Studienprojekt unter der Leitung von  
Jens Nachtwei und Carsten Schermuly  
WS 2009/2010

# **Qualität in Assessment Centern**

---

Ein Vergleich zwischen Wissenschaft und Praxis

Verfasserinnen:

Valeria Trinko  
Franziska Schölmerich

## **Abstract**

Die Frage nach der Qualität von Assessment Centern (ACn) in der Praxis sorgte in der Vergangenheit für viele Diskussionen (z. B. Kanning et al., 2007; Kersting, 2009; Höft, 2009; Krause, 2009). Ziel dieser Studie ist es daher, a) zu überprüfen, inwieweit wissenschaftlich fundierte Qualitätskriterien von Organisationen umgesetzt werden und b) wie diese Kriterien von Praktikern hinsichtlich ihrer Sinnhaftigkeit und praktischen Umsetzbarkeit bewertet werden. Dabei wurde der Schwerpunkt auf anforderungsdimensionenbezogene Kriterien für AC gelegt, die vorwiegend in der Personalauswahl eingesetzt werden. Von den 55 Organisationen, die an der Befragung teilnahmen, konnten 41 der Gruppe *Unternehmen* und 14 der Gruppe *Beratungen* zugeordnet werden. Die Ergebnisse zeigen nur geringe Unterschiede zwischen den beiden Gruppen. Insgesamt lässt sich ein positiver Trend der Umsetzung wissenschaftlicher AC-Kriterien in der Praxis feststellen, dennoch werden auch einige Defizite deutlich. Darüber hinaus wird die Sinnhaftigkeit der Kriterien von den Praktikern durchgehend höher bewertet als die Umsetzbarkeit. Die Ergebnisse beider Teile werden in Zusammenhang gesetzt und anschließend diskutiert.

**Keywords:** Assessment Center, Anforderungsdimensionen, Personalauswahl, Human Resources, Validität

## **Abstract**

The quality of Assessment Center (AC) has recently been a much discussed topic (e.g. Kanning et al., 2007; Kersting, 2009; Höft, 2009; Krause, 2009). In this study, we aim at investigating the realization of scientific recommendations in the organizational praxis, focusing on recommendations regarding the dimensions in AC that are mainly used for personnel selection. Furthermore, these recommendations will be rated by experts that use AC in their organizations for their reasonableness as well as their realizability. A questionnaire was completed by a sample of 55 organizations that can be divided in two groups: companies (N=41) and consultancies (N=14). Results show only little differences between companies and consultancies. Although a few shortcomings are discovered, an overall positive tendency is observed concerning the realization of scientific recommendations in the organizational praxis. Moreover, recommendations are constantly rated more reasonable than realizable. Results of both parts are subsequently combined and discussed.

**Keywords:** assessment center, assessment dimensions, personnel selection, human resources, validity

## Einleitung

Assessment Center (AC) sind in deutschen Unternehmen eine immer häufiger eingesetzte Methode zur Personalauswahl und –entwicklung. In einer aktuellen Studie stellen Schuler, Hell, Trapmann, Schaar und Boramir (2007) fest, dass in 57,6% der befragten Unternehmen AC eingesetzt werden. In einer weiteren Studie geben sogar 73,4% aller befragten Unternehmen an, ein AC einzusetzen (Obermann, Höft & Janke, 2008). Während die prognostische Validität von ACn unumstritten ist (z.B. Schmidt & Hunter, 1998; Krause, Kersting, Heggstad & Thornton, 2006), gibt es bis heute keine einheitlichen Qualitätsstandards zur Konzeption, Durchführung und Evaluation von ACn. Ein Versuch von Kanning, Pöttker und Gelléri (2007), wissenschaftlich fundierte Qualitätskriterien für das AC in der Praxis zu formulieren, resultierte in einem wissenschaftlichen Diskurs, der in der *Zeitschrift für Arbeits- und Organisationspsychologie* veröffentlicht wurde. Hier bemängeln Kersting (2009) und Höft (2009) unter anderem die Einseitigkeit des Sender-Empfänger-Modells zwischen Wissenschaft und Praxis und betonen die Wichtigkeit des Dialoges und der Diskussion zwischen beiden Seiten anstelle von Belehrungen seitens der Wissenschaft. Krause (2009) kritisiert zudem, dass die ökonomischen Randbedingungen in den Unternehmen bei der Auswahl und Bewertung der Qualitätsstandards nicht beachtet wurden. Diese Kritikpunkte werden in der vorliegenden Studie aufgegriffen.

Die Grundauffassungen von AC-Qualität sind je nach Unternehmen sehr unterschiedlich und dabei auch immer vom Kontext abhängig. Daher ist das Ziel dieser Untersuchung nicht, einheitliche Qualitätsstandards für ein AC festzulegen. Im ersten Teil soll lediglich überprüft werden, inwieweit ausgesuchte wissenschaftlich fundierte Qualitätskriterien in der Praxis umgesetzt werden. Um den Umfang der Untersuchung in einem überschaubaren Maße zu halten, beschränken wir uns dabei größtenteils auf Kriterien im Zusammenhang mit den Anforderungsdimensionen in einem AC. Laut Schuler et al. (2007) wird dem AC im Vergleich zu anderen Auswahlverfahren von Praktikern eine eher schlechte Praktikabilität bescheinigt. Daher werden im zweiten Teil dieser Untersuchung die wissenschaftlichen Kriterien von AC-Experten aus der Praxis nach Sinnhaftigkeit und praktischer Umsetzbarkeit bewertet, um so den Dialog zwischen Wissenschaft und Praxis zu eröffnen und die ökonomischen Randbedingungen in Unternehmen mit einzubeziehen. Auf diese Weise kann analysiert werden, ob die mangelnde Umsetzung einiger Kriterien in der Praxis sich wirklich auf Unkenntnis der Praktiker zurückführen lässt (Kanning et al., 2007) oder ob es dafür andere Gründe gibt. Schließlich wird die Gesamtstichprobe in die Gruppen *Unternehmen* und *Beratungen* unterteilt und die Ergebnisse der Befragung werden verglichen und anschließend diskutiert.

## Theorie

Wie eingangs erwähnt, sind die Grundauffassungen von Qualität je nach Unternehmen sehr unterschiedlich und dabei immer kontextabhängig. Im Bezug auf AC kann man allerdings davon ausgehen, dass alle Organisationen sich ein hoch qualitatives Auswahlverfahren wünschen, damit die Treffsicherheit von Personalentscheidungen und damit die Rationalität dieser Entscheidungen steigt. Die am häufigsten verwendete Kennziffer für die Qualität eines ACs ist die Validität (Kleinmann, 2003). Um den Umfang der Untersuchung überschaubar zu halten, beschränken wir uns größtenteils auf Kriterien, die sowohl mit Anforderungsdimensionen als auch mit der Validität im Zusammenhang stehen. Dabei ist ein Kriterium ein Merkmal, das bei der Auswahl zwischen Personen, Objekten, Eigenschaften, usw. relevant für die Entscheidung ist. Die Formulierung von solchen Kriterien kann dazu beitragen, ein AC zu optimieren und für die Praxis nutzbar zu machen.

*Kriterium 1: Im AC sollten insgesamt zwischen 3 und 6 unterschiedliche Dimensionen erfasst werden.*

Eine Vielzahl von Studien spricht dafür, die Anzahl von Anforderungsdimensionen in einem AC möglichst klein zu halten. Krause und Gebert (2003) sowie Kleinmann (1997) weisen darauf hin, dass mit zunehmender Anzahl der Anforderungsdimensionen die Gefahr wächst, dass diese nicht hinreichend trennscharf voneinander zu unterscheiden sind. Dadurch wird für die Beobachter eine Merkmalsdifferenzierung erschwert, so dass die Tendenz steigt, ein undifferenziertes Globalurteil zu fällen (Krause & Gebert, 2003; Thornton & Rupp, 2006). Laut Gaugler und Thornton (1989) waren die Beobachterurteile dann genauer, wenn 3 bis 6 anstatt 9 Dimensionen in einem AC beobachtet wurden. Woehr und Arthur (2003) haben gezeigt, dass die Konstruktvalidität des ACs steigt, wenn die Anzahl der beobachteten Dimensionen sinkt. Weiterhin sinkt mit zunehmender Anzahl der beobachteten Anforderungsdimensionen die konvergente Validität eines ACs (Gaugler & Thornton, 1989).

*Kriterium 2: Die Dimensionen sollten durch verhaltensorientierte Items erfasst werden.*

Laut Krause und Gebert (2003) ist es unentbehrlich, die Anforderungsdimensionen durch verhaltensorientierte Items zu erfassen. Da ein AC ein verhaltensorientiertes diagnostisches Verfahren ist, sind die protokollierten Verhaltensbeschreibungen das einzige Mittel, zwischen tatsächlichem Teilnehmerverhalten und Interpretation oder Schlussfolgerungen der Beobachter zu unterscheiden (Höft & Funke, 2001). Reilly, Henry und Smither zeigten bereits 1990, dass die Verwendung von verhaltensnahen Beschreibungen (z. B. Verhaltenscheckli-

sten) als Beobachtungssystem Globalurteile verhindern kann und zur Steigerung der Konstruktvalidität sowie konvergenter Validität führt. Wenn aber die Erfassung von Dimensionen durch eigenschaftsorientierte Items erfolgt, so müssen die Beobachter während einer Übung selbst entscheiden, in welche Dimension beobachtetes Verhalten einzuordnen ist, wodurch die Validität eines ACs möglicherweise verringert wird.

*Kriterium 3: Eine Dimension sollte durch vier bis acht Items pro Übung erfasst werden.*

Um eine bessere Differenzierungsfähigkeit von Dimensionen für die Beobachter während einer Übung zu ermöglichen, sollen Beurteilungsdimensionen durch vier bis acht Items (Verhaltensanker) pro Übung operationalisiert werden (Eck, Jöri & Vogt, 2007). Weniger als vier Items zu verwenden, kann die Reliabilität und auch die Validität eines ACs gefährden. Laut klassischer Testtheorie wird die Messung reliabler, je länger ein Test ist. Eck et al. (2007) betonen aber, dass eine Dimension durch maximal acht Items pro Übung erfasst werden sollte, damit eine adäquate Bewertung gewährleistet werden kann (wenn bspw. bei fünf Dimensionen je Übung fünf Items zur Verfügung stehen, so müssen Beobachter bereits 25 Aspekte bewerten (Obermann, 2006)). Laut Howard (1997) reichen acht Items pro Dimension aus, um die Struktur und Reliabilität zu gewährleisten und gleichzeitig die Beobachter nicht zu überfordern. Lievens und Conway (2001) weisen auch darauf hin, dass Beobachterfehler teils auf kognitive Überlastungen der Beobachter zurückzuführen sind und somit die Validität eines ACs verringern.

*Kriterium 4: Ein AC sollte zwischen 3 und 6 verschiedene Übungen beinhalten.*

Ein AC ist ein diagnostisches Verfahren, welches sich aus mehreren Einzelverfahren bzw. Übungen zusammensetzen sollte (Kleinmann, 2003). Gaugler, Rosenthal, Thornton und Bentson (1987, 1992) haben gezeigt, dass AC, die eine größere Anzahl unterschiedlicher Übungen aufweisen, prognosestärker und valider sind. Laut Kleinmann (2003) sollte die Anzahl der Übungen zwischen drei und sechs liegen, wenn das Ziel des ACs Potenzialidentifikation oder Selektion ist. Wird ein AC für Trainingszwecke eingesetzt, ist es sinnvoll, die Anzahl der Übungen auf sieben bis zehn zu steigern. Die Standards der AC Technik (Arbeitskreis Assessment Center e.V., 2004) weisen auch darauf hin, dass in einem AC mindestens drei Arbeitssituationen simuliert werden müssen.

*Kriterium 5: Eine Dimension sollte in mindestens zwei unabhängigen Übungen beobachtet werden.*

AC folgen dem Multitrait-Multimethod-Prinzip: Es werden verschiedene Konstrukte (Dimensionen) mit unterschiedlichen Methoden (Übungen) erfasst; zugleich wird die konvergente bzw. diskriminante Validität überprüft. Den „Standards der Assessment Center Technik“ (Arbeitskreis Assessment Center e.V., 2004) zu folge, sollte jede Dimension in mindestens zwei voneinander unabhängigen Übungen untersucht werden (Redundanzprinzip). So wird es möglich, die konfundierenden Umwelteinflüsse, die auf das Verhalten wirken, zumindest ansatzweise von den interessierenden Einflüssen der Person zu trennen (Kanning, 2004). Da laut klassischer Testtheorie eine Messung reliabler wird, je länger ein Test ist, wäre folgendes anzunehmen: die Ausprägung eines Verhaltens lässt sich auf einer bestimmten Dimension genauer erfassen, je häufiger es beobachtet wird, was schließlich zu einer erhöhten Konstruktvalidität führt.

*Kriterium 6: Innerhalb einer Übung sollten nicht mehr als vier Dimensionen gleichzeitig beobachtet werden.*

Während der Durchführung eines ACs werden hohe Anforderungen an die Aufmerksamkeit der Beobachter gestellt, da das Verhalten der Kandidaten innerhalb einer Übung meist hinsichtlich mehrerer Dimensionen beobachtet wird. Lievens und Conway (2001) weisen darauf hin, dass Beobachterfehler teils auf kognitive Überlastungen der Beobachter zurückzuführen sind. Durch die Reduktion der zu beobachtenden Dimensionen kann diese kognitive Überlastung verringert und somit die Qualität der Bewertung erhöht werden (Gaugler & Thornton, 1989). Kleinmann (2003) weist darauf hin, dass mit zunehmender Anzahl der Anforderungsdimensionen diese nicht hinreichend trennscharf voneinander zu unterscheiden sind. Woehr und Arthur (2003) haben gezeigt, dass die Konstruktvalidität des ACs steigt, wenn die Anzahl der beobachteten Dimensionen innerhalb einer Übung sinkt. Eck et al. (2007) betonen zudem, dass innerhalb einer Übung nicht mehr als vier Dimensionen beobachtet und beurteilt werden sollten, damit eine adäquate Bewertung gewährleistet werden kann.

*Kriterium 7: Das AC sollte hinsichtlich der zu erfassenden Dimensionen auf die spezifischen Anforderungen des Unternehmens zugeschnitten sein.*

In der Literatur lassen sich zahlreiche Studien finden, die belegen, dass die Übungen eines ACs hinsichtlich der zu erfassenden Dimensionen auf die spezifischen Anforderungen des Unternehmens zugeschnitten sein sollten. Laut Kersting (1998) weisen AC, die in vollständiger Eigenentwicklung konstruiert wurden, eine erhöhte Inhaltsvalidität auf. Auch Krause, Meyer zu Kniedorf und Gebert (2001) bestätigen, dass die Inhaltsvalidität eines ACs

steigt, je höher der Anteil an spezifisch neu entwickelten Übungen für das Unternehmen ist. Um die wichtigsten Komponenten einer erfolgreichen Arbeitsleistung zu erfassen und daraus die Dimensionen des ACs abzuleiten, sollte eine Anforderungsanalyse durchgeführt werden (International Task Force on Assessment Center Guidelines, 2000).

*Kriterium 8: Das Verhältnis zwischen AC-Teilnehmern und AC-Beobachtern sollte maximal 2:1 sein.*

Mehrere Autoren (z.B. Eck et al., 2007) empfehlen, dass das quantitative Verhältnis zwischen AC-Teilnehmern und AC-Beobachtern in jeder Übung maximal 2:1 sein sollte, um die kognitive Überforderung der Beobachter zu verhindern und so die Qualität der Beobachtung und Beurteilung zu erhöhen. Woehr und Roch (1996) weisen daraufhin, dass die Beobachter Schwierigkeiten haben, Informationen auf der Ebene der Dimensionen für jeden Teilnehmer zu verarbeiten, wenn das Verhältnis zwischen AC-Teilnehmer und AC-Beobachter steigt. Damit kann ein Verhältnis von 2:1 als akzeptabel angesehen werden, um einer Informationsüberlastung der Beobachter vorzubeugen und Beurteilungsverzerrungen weitgehend zu reduzieren (Krause et al., 2001). Eine zunehmende Beobachteranzahl kann mit einer höheren Konstruktvalidität und prädiktiven Validität des ACs korrespondieren (Krause & Gebert, 2003).

*Kriterium 9: Die Beobachter sollten vor dem AC intensiv bezüglich der Dimensionen geschult werden.*

Das Ziel des Beobachtertrainings im Bezug auf die Anforderungsdimensionen ist die Erleichterung des eindeutigen Zuordnens von einzelnen Verhaltensweisen der Teilnehmer zu den entsprechenden Dimensionen (Kleinmann, 2003). Dadurch wird es möglich, die Genauigkeit bei der Beurteilung von Verhaltensaüßerungen sowie die Geschwindigkeit von Zuordnungen der Verhaltensweisen zu Dimensionen zu steigern. Zum Einen führt es zur Reduzierung der Belastung während der Beobachtung (Kleinmann, 2003). Zum Anderen werden laut Woehr und Huffcutt (1994) spezifische Beobachterfehler wie z.B. Halofehler und Mildefehler seltener, was zur Verbesserung der Reliabilität führt. Beobachtertrainings im Bezug auf die Anforderungsdimensionen verbessern die Urteilsakkuratheit (Höft & Melchers, 2008) und können somit zur Steigerung der Validität eines ACs beitragen.

*Kriterium 10: Die zu erfüllenden Ausprägungen auf den Dimensionen sollten durch unterschiedliche Informationsquellen festgelegt werden.*

Thornton und Byham (1982) beschreiben die von ihnen favorisierte Anforderungsermittlung im AC als mehrstufigen Prozess, der sich auf die folgenden Datenquellen bezieht: Dokumentenanalyse, Verhaltensbeobachtungen vor Ort, Interviews mit Stelleninhabern und Vorgesetzten sowie Analyse erfolgskritischer Ereignisse. Auf der Grundlage dieser Informationsquellen werden die ermittelten Verhaltensmerkmale statistisch und/oder intuitiv zu Anforderungsdimensionen zusammengestellt und nach deren Ausprägungsgrad abgestuft. Auch nach Schuler (2001) sowie Krause und Thornton (2009) erweist sich der Einbezug verschiedener Informationsquellen für die Anforderungsanalyse als sinnvoll. Dadurch wird es möglich, die individuellen Ausprägungen auf den einzelnen Dimensionen genauer zu definieren und somit die prognostische Validität des ACs zu erhöhen.

*Kriterium 11: Die zu erfüllende Ausprägung auf jeder Dimension sollte, je nach dem um welche Dimension es sich handelt, gesondert festgelegt werden.*

Die genauen Ausprägungen der Dimensionen in Anforderungsprofilen sind stark abhängig von der spezifischen Position. So kann es für eine Fachkraft beispielsweise viel wichtiger sein, eine hohe Ausprägung auf der Dimension „Stressresistenz“ zu besitzen als auf der Dimension „Führungsstärke“. Um den spezifischen Anforderungen einer Position gerecht zu werden, sollten die zu erfüllenden Ausprägungen auf jeder Dimension daher gesondert festgelegt werden. Dieses Kriterium wurde logisch hergeleitet, da sich hierzu wenig theoretische und empirische Evidenz finden lässt.

*Kriterium 12: Die zu erfüllende Ausprägung auf jeder Dimension sollte, je nach dem um welche Position und/oder Unternehmenseinheit es sich handelt, gesondert festgelegt werden.*

Die Anforderungsprofile innerhalb desselben Berufszweiges können über verschiedene Unternehmen oder Abteilungen hinweg stark voneinander abweichen, deshalb sollte das Anforderungsprofil und Ausprägungen auf einzelnen Dimensionen spezifisch für die Position erstellt bzw. zumindest auf das Unternehmen und die Abteilung zugeschnitten werden (Arbeitskreis Assessment Center e.V., 2004). Um die Inhaltsvalidität eines ACs zu maximieren, ist eine systematische Ableitung von Anforderungsdimensionen aus der unternehmensspezifischen Arbeits- und Anforderungsanalyse notwendig (Krause et al., 2001). Weiterhin betont Schuler (2001), dass bei einer Anforderungsanalyse neben den tätigkeitsspezifischen Anforderungen auch tätigkeitsübergreifende Anforderungen bestimmt werden müssen.

*Kriterium 13: Die Dimensionen des ACs werden nach der Durchführung statistisch evaluiert.*



Die Durchführung von Evaluationen gehört zu einem der zentralen Qualitätsstandards des Arbeitskreises Assessment Center e.V. Es wird argumentiert, dass insbesondere ein neu entwickeltes AC zunächst nur ein „Hypothesengebäude“ über den Zusammenhang zwischen Anforderungsdimensionen, AC-Komponenten und Bewährungskriterien in der Praxis darstellt. Die statistische Evaluation kann diese Zusammenhänge empirisch absichern und dafür sorgen, dass das Verfahren ständig optimiert wird. Es sollte vor allem darauf geachtet werden, dass einzelne Dimensionen einen ausreichenden Beitrag für das Gesamturteil leisten und gut zwischen den Teilnehmern differenzieren (Arbeitskreis Assessment Center e.V., 2004). In der Literatur lassen sich zahlreiche Studien finden, die belegen, dass die Evaluation eines ACs zur Qualitätssteigerung des Verfahrens beiträgt. Laut Krause und Thornton (2009) ist es sogar strategisch riskant für eine Organisation, ein nicht valides AC einzusetzen.

## **Methode**

### *Stichprobe*

Über das Online-Business-Netzwerk Xing ([www.xing.com](http://www.xing.com)) wurden ca. 1000 Anschreiben mit einem Link zum Fragebogen an Experten zum Thema Personalauswahl, -entwicklung und Human Resource Management versendet. Zudem wurden weitere Anschreiben an Kontakte aus Stellenanzeigen oder an Ansprechpartner bei Firmenkontaktmessen geschickt. Im Anschreiben wurde das Projekt vorgestellt und den Teilnehmern ein Feedback zur Qualität ihres ACs und ein Vergleich ihrer individuellen Ergebnisse mit den anonymisierten Ergebnissen anderer Organisationen in Aussicht gestellt. Durch ein Code-System wurde die Anonymität der Teilnehmer gewährleistet.

Insgesamt nahmen 57 verschiedene Organisationen an der Befragung teil, die jeweils durch einen AC-Verantwortlichen vertreten wurden. Alle Teilnehmer können als Experten aus der Praxis zum Thema AC bezeichnet werden, da sie entweder an der Konzeption, der Durchführung, der Evaluation oder an mehreren dieser Phasen beteiligt waren. Vor der Beantwortung der Fragen wurden sie darauf hingewiesen, dass sie sich bei ihren Antworten auf ein spezifisches AC ihrer Organisation beziehen sollen. Zwei Datensätze wurden von der Auswertung ausgeschlossen, da in einem Fehlbeantwortungen vorlagen und im anderen der Befragte angab, an keiner der Phasen des ACs beteiligt zu sein. Die endgültige Stichprobe bestand somit aus 55 Datensätzen aus verschiedenen Organisationen.

Bei der Analyse des Datenmaterials wurde festgestellt, dass einige Datensätze nicht in ein typisches Unternehmens-Profil passen, obwohl zunächst nur Personalverantwortliche aus

Unternehmen ohne unternehmensberaterischen Bezug adressiert worden sind. Daher wurde beschlossen, die Daten weiter zu differenzieren und zwei Untergruppen durch den Vergleich der Anzahl von AC-Anwendungen pro Jahr und der Anzahl der Mitarbeiter der Organisation zu bilden. Wenn die Anzahl der AC-Anwendungen pro Jahr halb so groß oder größer als die Anzahl der Mitarbeiter war, wurde die Organisation als Beratung klassifiziert (N = 14). Wenn die Anzahl der AC pro Jahr geringer als die Hälfte der Anzahl der Mitarbeiter war, wurde die Organisation als Unternehmen klassifiziert (N = 41). Da es zu den Unterschieden zwischen Unternehmen und Beratungen in der AC-Praxis bisher sehr wenig wissenschaftliche Literatur gibt, werden die Unterschiede in dieser Studie explorativ untersucht. Alle Ergebnisse beziehen sich zunächst auf die gesamte Stichprobe (*Gesamt*) und werden anschließend in die Gruppen *Unternehmen* und *Beratung* unterteilt.

Ausführliche Informationen über die Stichprobe sind in den Tabellen 1.1 und 1.2 dargestellt: Aus Tabelle 1.1 ist ersichtlich, wie viele Mitarbeiter die Organisationen im Mittel haben, seit wie vielen Jahren das AC eingesetzt wird, wie viele Personen im Jahr am AC teilnehmen und mit welcher Schulnote die Befragten das AC beurteilen (*Güte AC*). Tabelle 1.2 zeigt, zu welchem Zweck und für welche Positionen das AC eingesetzt wird sowie die Phase, an der die Befragten beteiligt waren.

*Tabelle 1.1.* Deskriptive Statistik (Mittelwerte)

	Gesamt (N=55)				Unternehmen (N=41)				Beratung (N=14)			
	M	SD	min	max	M	SD	min	max	M	SD	min	max
Anzahl Mitarbeiter	8196,58	23005,13	2	150000	10986,93	26136,47	100	150000	24,86	47,69	2	180
Anzahl Jahre	5,69	5,69	0	24	4,95	5,3	0	24	7,86	4,96	1	19
Anzahl Teilnehmer	105,04	161,04	9	900	103,07	132,67	10	700	110,79	231,34	9	900
Güte AC	2,25	0,78	1	4	2,24	0,77	1	4	2,29	0,83	1	4

*Anmerkungen.* M = Mittelwert, SD = Standardabweichung, min = Minimum, max = Maximum.

Aus Tabelle 1.1 wird deutlich, dass die Größe der befragten Organisationen stark variiert (Anzahl Mitarbeiter: min = 2; max = 150000). Die AC in den Unternehmen (M = 4,95) werden im Mittel kürzer (in Jahren) eingesetzt als in den Beratungen (M = 7,86). Auch die Anzahl der Teilnehmer am AC variiert stark in den befragten Organisationen. Die Güte des ACs wird von Unternehmen wie Beratungen ähnlich positiv beurteilt, der Gesamtdurchschnitt liegt bei 2,25 (Bewertung in Schulnoten).

*Tabelle 1.2.* Deskriptive Statistik (Häufigkeiten)

		Gesamt (N = 55)		Unternehmen (N = 41)		Beratungen (N = 14)	
		Häufigkeit	%	Häufigkeit	%	Häufigkeit	%
Zweck AC	Personalauswahl (PA)	28	50,9	23	56,1	5	35,7
	Personalentwicklung (PE)	6	10,9	4	9,8	2	14,3
	PA und PE	21	38,2	14	34,1	7	50,0
Position AC	Auszubildende	21	38,2	16	39,0	5	35,7
	Trainees	28	50,9	21	51,2	7	50,0
	Young Professionals	20	36,4	16	39,0	4	28,6
	Fachkräfte ohne Führungsaufgabe	21	38,2	15	36,6	6	42,9
	Mittleres Management	25	45,5	17	41,5	8	57,1
	Höheres Management	11	20,0	7	17,1	4	28,6
Beteiligung AC	Konzeption	45	81,8	35	85,4	10	71,4
	Durchführung	53	96,4	40	97,6	13	92,9
	Evaluation	36	65,5	27	65,9	9	64,3

*Anmerkungen.* Bei den Fragen „Position AC“ und „Beteiligung AC“ waren Mehrfachnennungen möglich.

Tabelle 1.2 zeigt, dass das AC vor allem zur Personalauswahl aber auch oft in Kombination mit der Personalentwicklung eingesetzt wird. Am häufigsten wird das AC sowohl von Unternehmen als auch von Beratungen für die Positionen von Trainees und dem mittleren Management eingesetzt, aber auch die anderen Positionen sind durchgehend vertreten. Die Daten zeigen, dass fast alle Befragten (96,4%) an der Durchführung des ACs und 60% an allen drei Phasen des ACs beteiligt sind.

### *Fragebogen*

Um zu erfassen, welche Qualität die AC in deutschsprachigen Organisationen haben und wie die oben beschriebenen wissenschaftlichen Kriterien von den AC-Experten aus der Praxis eingeschätzt werden, wurde ein zweiteiliger Fragebogen mit insgesamt 35 Fragen in einem mehrstufigen Vorgehen entwickelt. Nach einer ausführlichen Recherche in verschiedenen Literaturdatenbanken (e.g. PsychInfo, Web of Science) wurden die endgültigen Fragen von den Autoren in mehreren aufeinanderfolgenden Diskussionsworkshops ausgewählt. Im ersten Teil wurden neben einigen allgemeinen Angaben zum AC Informationen zur Konzeption, Durchführung und Evaluation des ACs in der Praxis erhoben (22 Fragen). Die Fragen wurden dabei sowohl im Multiple-Choice-Format als auch mit einer offenen Antwortmöglichkeit gestellt. Im zweiten Teil sollten die Praktiker jedes wissenschaftliche Kriterium auf zwei 7-stufigen Skalen mit den Endpunkten „gar nicht sinnvoll (0)“ und „sehr sinnvoll (6)“ sowie „sehr schwer umsetzbar (0)“ und „sehr leicht umsetzbar (6)“ bewerten (13 Fragen). Die

Bearbeitung des Fragebogens dauerte insgesamt etwa 30 Minuten. Einen Überblick über die wissenschaftlichen Kriterien und die dazugehörigen Fragen liefert Tabelle A (Anhang).

Einige Teilnehmer gaben in der Befragung keine konkreten Zahlenwerte, sondern eine Zahlenspanne an. Bei der Auswertung wurde aus diesen Zahlenspannen das arithmetische Mittel gebildet. Die Mittelwertsunterschiede zwischen den Beurteilungen der Skalen „sinnvoll“ und „umsetzbar“ wurden in der Gesamtgruppe aufgrund der verletzten Normalverteilungsangabe mithilfe des Wilcoxon-Tests für zwei abhängige Gruppen berechnet.

### *Statistisches Vorgehen*

Alle Daten wurden auf Normalverteilung überprüft. Daraus ergab sich, dass die Ergebnisse der beiden Stichproben nicht normal verteilt sind. Die Unterschiede zwischen Unternehmen und Beratungen hinsichtlich der Mittelwerte wurden mit dem Mann-Whitney-U-Test berechnet, da es sich hier um zwei unabhängige, nicht parametrische Datensätze handelt. Die Unterschiede zwischen Unternehmen und Beratungen hinsichtlich der Häufigkeitsverteilungen wurden mit dem Pearson-Chi-Quadrat-Test berechnet. Alle statistischen Berechnungen wurden mit SPSS 16.0 durchgeführt.

### **Ergebnisse**

Im ersten Teil der Befragung wurde untersucht, inwieweit die wissenschaftlichen AC-Kriterien in der Praxis umgesetzt werden. Die Ergebnisse werden in den Tabellen B.1 und B.2 im Anhang dargestellt.

Die befragten Organisationen erfassen im Mittel 8,76 Dimensionen ( $SD=5,56$ ) in ihrem AC, dabei erfüllen 36,59% der Unternehmen und 50% der Beratungen die wissenschaftliche Vorgabe von 3-6 Dimensionen (*Kriterium 1*). 96,3% der befragten Organisationen verwenden in ihrem AC nur verhaltensorientierte oder verhaltens- und eigenschaftsorientierte Items und erfüllen somit das *Kriterium 2* vollständig oder teilweise. Davon verwenden 29,3% der Unternehmen und 50% der Beratungen ausschließlich verhaltensorientierte Items. Die Anzahl der Items pro Dimension in einer Übung liegt sowohl bei Unternehmen ( $M=7,16$ ;  $SD=10,85$ ) als auch bei Beratungen ( $M=7,82$ ;  $SD=3,48$ ) im Mittel innerhalb der Spanne von 4-8 Items (*Kriterium 3*). 60% der Unternehmen und 71,4% der Beratungen erfüllen das Kriterium. Dabei ist zu beachten, dass es bei den Unternehmen einen statistischen Ausreißer (70) gab, der in die Berechnung integriert wurde. Die Anzahl der verschiedenen Übungen im AC liegt in der Gesamtgruppe bei  $M=5,64$  ( $SD=2,69$ ), dabei erfüllen 82,93% der Unternehmen

und 64,29% der Beratungen die wissenschaftliche Vorgabe von 3-6 Übungen (*Kriterium 4*). Nähere Informationen über die Art der Übungen können Abbildung A und Abbildung B im Anhang entnommen werden. *Kriterium 5* wird von fast allen befragten Organisationen erfüllt, lediglich in einem Unternehmen wird eine Dimension nicht in mindestens zwei unabhängigen Übungen beobachtet. Die Mittelwerte liegen hier deutlich über dem vorgegebenen Wert 2 (Unternehmen:  $M=2,59$ ;  $SD=0,87$ ; Beratungen:  $M=3,11$ ;  $SD=1,36$ ). *Kriterium 6* wird von 62,5% der Unternehmen und 71,43% der Beratungen erfüllt, in der Gesamtgruppe liegt der Mittelwert ( $M=4,24$ ;  $SD=1,35$ ) leicht über dem vorgegebenen Wert von maximal 4 Dimensionen pro Übung.

Die Ergebnisse zu *Kriterium 7* zeigen, dass Unternehmen signifikant häufiger (Unternehmen: 90,2%; Beratungen: 64,3%) ihr AC neu konzipieren und so auf ihre spezifischen Anforderungen zuschneiden. Beim Verhältnis von Teilnehmern zu Beobachtern (*Kriterium 8*) zeigt sich, dass in der Gesamtgruppe die Vorgabe von maximal 2:1 im Mittel erfüllt wird, dabei halten 74,36% der Unternehmen und 76,92% der Beratungen dieses Kriterium ein. Informationen zur Anzahl der Beobachter in Einzelübungen können der Tabelle C im Anhang entnommen werden. Die Ergebnisse zu *Kriterium 9* zeigen, dass die Beobachter vor dem AC im Mittel 2,07 Stunden ( $SD=2,59$ ) bezüglich der Dimensionen geschult werden. Dabei werden die Beobachter von Beratungen ( $M=3,71$ ;  $SD=3,37$ ) signifikant länger geschult als von Unternehmen ( $M=1,49$ ,  $SD=2$ ).

Ein Großteil der Unternehmen und Beratungen beruft sich bei der Festlegung der Ausprägungen auf den verschiedenen Dimensionen auf 2-3 Informationsquellen (56,1% bei den Unternehmen und 64,3% bei den Beratungen). 18,2% der Unternehmen (0% der Beratungen) verwenden keine oder nur eine Informationsquelle und erfüllen somit nicht das *Kriterium 10*. *Kriterium 11* wird von mehr als der Hälfte der befragten Organisationen erfüllt, 58,2% der Gesamtgruppe geben hier an, die Ausprägungen auf den Dimensionen je nach Dimension variabel festzulegen. *Kriterium 12* wird von 29,1% der Gesamtgruppe nicht erfüllt, da diese Befragten die Ausprägungen für die verschiedenen Dimensionen für das gesamte Unternehmen festlegen. Fast die Hälfte aller Befragten (45,5% der Gesamtgruppe) führt keine statistische Evaluation ihres ACs durch und verstößt somit gegen *Kriterium 13*. Dabei wird in Unternehmen (48,8%) häufiger auf eine statistische Evaluation verzichtet als in Beratungen (35,7%).

Tabelle 2 gibt an, wie viele wissenschaftliche Kriterien von der Gesamtgruppe sowie von den Unternehmen und Beratungen im Mittel jeweils erfüllt worden sind.

**Tabelle 2.** Erfüllte Kriterien (Mittelwerte)

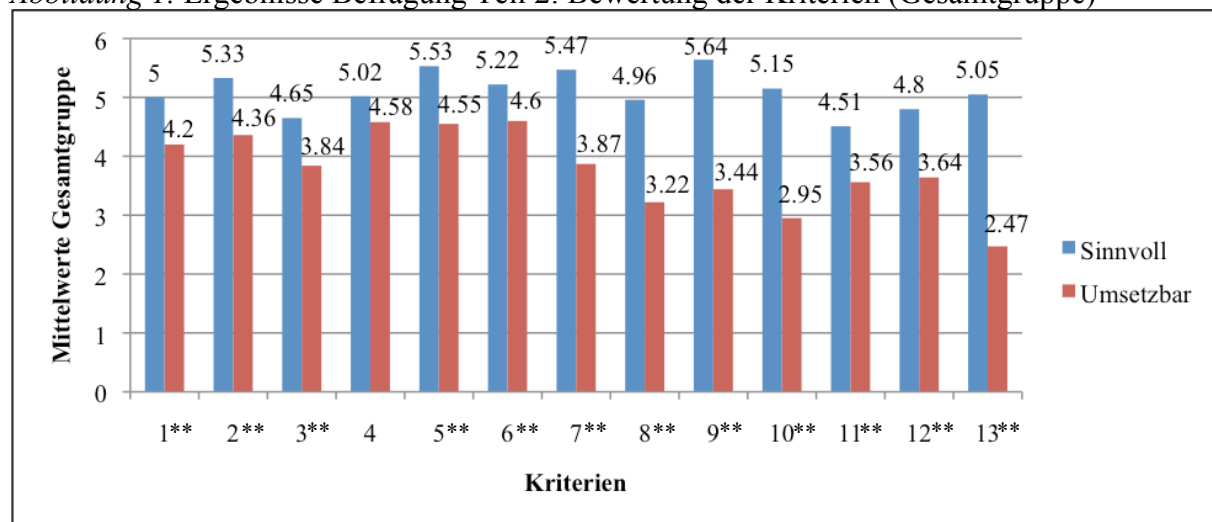
Gesamt					Unternehmen					Beratung				
N	M	SD	min	max	N	M	SD	min	max	N	M	SD	min	max
55	7,93	1,91	2	11	41	7,78	2,03	2	11	14	8,36	1,5	5	10

Anmerkungen. M = Mittelwert, SD = Standardabweichung, min = Minimum, max = Maximum.

Es zeigt sich, dass von der Gesamtgruppe im Mittel knapp 8 Kriterien erfüllt werden. Die Ergebnisse der Untergruppen sind vergleichbar (Unternehmen: M=7,78, Beratungen: M=8,36) und unterscheiden sich nicht signifikant voneinander. Dabei muss beachtet werden, dass Kriterium 9 von dieser Analyse ausgeschlossen wurde, da nicht eindeutig festgelegt werden konnte, was unter einer „intensiven Schulung der Beobachter bezüglich der Dimensionen“ zu verstehen ist. Das Maximum an erfüllbaren Kriterien beträgt hier also 12.

Im zweiten Teil der Befragung sollten die AC-Experten aus der Praxis die Sinnhaftigkeit sowie die praktische Umsetzbarkeit der wissenschaftlichen Kriterien bewerten. Die Ergebnisse werden in Abbildung 1 (Gesamtgruppe) und Tabelle 3 (Untergruppen) dargestellt.

**Abbildung 1.** Ergebnisse Befragung Teil 2: Bewertung der Kriterien (Gesamtgruppe)



Anmerkungen. \*\* =  $p < .01$  (zweiseitig); Unterschiede berechnet mit dem Wilcoxon-Test.

Abbildung 1 zeigt, dass alle Kriterien in der Gesamtgruppe auf der Skala „Sinnvoll“ höher bewertet wurden als auf der Skala „Umsetzbar“. Dieses Ergebnis lässt sich auch bei den einzelnen Bewertungen der Unternehmen und Beratungen wiederfinden (Tabelle 3). Ausgenommen Kriterium 4 sind alle Unterschiede zwischen den Skalen in der Gesamtgruppe signifikant. Als besonders sinnvoll wurden Kriterium 9 (M = 5,64) und Kriterium 5 (M = 5,53) bewertet. Für weniger sinnvoll hielten die Befragten Kriterium 11 (M = 4,51) und Kriterium 3 (M = 4,65). Als eher leicht praktisch umsetzbar bewerteten die Befragten das Kriterium 6 (M = 4,6) und das Kriterium 4 (M = 4,58). Als weniger leicht umsetzbar wurden das Kriterium 13 (M = 2,47) und das Kriterium 10 (M = 2,95) bewertet.

**Tabelle 3. Ergebnisse Befragung Teil 2: Bewertung der Kriterien (Untergruppen)**

Kriterium	Unternehmen (N = 41)				Beratung (N = 14)			
	sinnvoll		umsetzbar		sinnvoll		umsetzbar	
	M	SD	M	SD	M	SD	M	SD
1	5,15	1,11	4,2	1,52	4,57	1,22	4,21	1,76
2	5,27	1,1	4,24	1,36	5,5	0,76	4,71	1,68
3	4,49	1,47	3,59	1,77	5,14	1,1	4,57	1,22
4	5,27	1,1	4,59	1,4	4,29	1,86	4,57	1,22
5	5,44	1,14	4,39	1,58	5,79	0,58	5	1,47
6	5,15	1,26	4,49	1,50	5,43	0,76	4,93	1,0
7	5,61	0,83	3,88	1,47	5,07	1,54	3,86	1,70
8	5,15	1,3	3,12	1,6	4,43	1,95	3,5	1,79
9	5,66	0,62	3,27	1,76	5,57	0,85	3,93	1,27
10	5,15	0,96	2,98	1,42	5,14	1,51	2,86	1,70
11	4,56	1,83	3,68	1,85	4,36	1,34	3,21	1,58
12	4,85	1,53	3,59	1,83	4,64	1,15	3,79	1,48
13	4,98	1,21	2,39	1,6	5,29	0,99	2,71	1,98

Anmerkungen. M = Mittelwert, SD = Standardabweichung.

Aus Tabelle 3 wird deutlich, dass sich bei der differenzierten Betrachtung der Daten in den Gruppen *Unternehmen* und *Beratungen* ähnliche Ergebnisse wie in der Gesamtgruppe zeigen. Auch hier sind die Beurteilungen auf der Skala „Sinnvoll“ durchgehend höher als die der Skala „Umsetzbar“. Die Unterschiede zwischen den beiden Gruppen sind dabei von geringem Ausmaß.

## Diskussion

Ziel dieser Untersuchung war es, einige dimensionsbezogene Aspekte von ACn in der Praxis aufzugreifen und anschließend ihre Relevanz und praktische Umsetzbarkeit von AC-Experten aus Unternehmen und Beratungen beurteilen zu lassen. Insgesamt zeigt sich ein positiver Trend, da die befragten Organisationen einen Großteil der wissenschaftlichen Kriterien vollständig einhalten oder nur leicht von den empfohlenen Richtwerten abweichen. Damit sind diese Ergebnisse vergleichbar mit denen von Kanning et al. (2007). Passend zur Kritik von Kersting (2009), Höft (2009) und Krause (2009) wird zudem deutlich, dass die Sinnhaftigkeit der Kriterien von den Praktikern durchgehend höher als die Umsetzbarkeit bewertet wird. Auch nach Bereinigung der Alpha-Inflation durch Mehrfachvergleiche waren fast alle Unterschiede zwischen den Bewertungen der Sinnhaftigkeit und der Umsetzbarkeit der Kriterien statistisch signifikant (ausgenommen Kriterium 4). Dies zeigt, dass die wissenschaftlichen Kriterien in der Praxis durchaus bekannt sind und als sinnvoll eingeschätzt werden, die praktische Umsetzung jedoch teilweise schwierig ist. Die Ergebnisse zu den einzelnen Kriterien werden im Folgenden diskutiert.

Bei der Mehrzahl der Befragten werden im AC mehr Dimensionen als vorgeschlagen erfasst, obwohl *Kriterium 1* als sinnvoll und durchaus umsetzbar bewertet wird. Vermutlich

gehen die Organisationen oft nach dem Prinzip „Je mehr desto besser“ vor und berücksichtigen dabei nicht, dass es mit steigender Anzahl immer schwieriger wird, zwischen den einzelnen Dimensionen zu differenzieren. *Kriterium 2* wird von Beratungen (50%) häufiger erfüllt als von Unternehmen (29,3%). Beide Gruppen bezeichnen das Kriterium als sehr sinnvoll, jedoch bewerten die Beratungen es als etwas leichter umsetzbar. Der Großteil der Befragten verwendet sowohl verhaltens- als auch eigenschaftsorientierte Items obwohl es sehr wichtig ist, zwischen diesen Arten von Items zu differenzieren. Mehr als die Hälfte der befragten Organisationen erfüllen *Kriterium 3*, wobei die Werte innerhalb der Stichprobe stark streuen. Im Vergleich zu den anderen Kriterien wird es als eher weniger sinnvoll und als weniger leicht umzusetzen bewertet. *Kriterium 4* wird von Unternehmen (82,93%) häufiger erfüllt als von Beratungen (64,29%), da einige Beratungen mehr als 6 Übungen im AC einsetzen. Dieses Kriterium wird von Unternehmen auch für sinnvoller gehalten. Beide Gruppen bewerten es gleichermaßen als durchaus umsetzbar. In der Gesamtgruppe zeigt sich allein bei diesem Kriterium kein signifikanter Unterschied zwischen den Skalen „Sinnvoll“ und „Umsetzbar“. Fast alle befragten Organisationen erfüllen *Kriterium 5*, dementsprechend wird es auch als besonders sinnvoll und eher leicht umzusetzen bewertet. Da nur ein Unternehmen dieses Kriterium nicht erfüllt, ist es als wenig trennscharf zu betrachten. Die Mehrzahl der befragten Organisationen erfüllt *Kriterium 6*, tendenziell werden jedoch zu viele Dimensionen pro Übung beobachtet. Grund dafür könnte sein, dass im AC generell zu viele Dimensionen erfasst werden (siehe Ergebnisse Kriterium 1). Beide Gruppen bewerten dieses Kriterium sowohl als sinnvoll als auch als leicht umsetzbar. Unternehmen konzipieren ihr AC deutlich häufiger neu als Beratungen (*Kriterium 7*). Dementsprechend wird dieses Kriterium auch von Unternehmen als etwas sinnvoller beurteilt. Beide Gruppen beurteilen es gleichermaßen als eher schwierig umsetzbar. Vermutlich sind in Beratungen häufiger schon Standardlösungen vorhanden, die dann an die jeweilige Organisation angepasst werden. Rund drei Viertel der Befragten erfüllen *Kriterium 8*, dabei beurteilen Unternehmen es als sinnvoller aber gleichzeitig als schwierig umzusetzen. Dies zeigt, dass die Befragten zwar die Sinnhaftigkeit des Kriteriums anerkennen, der Umsetzungsaufwand vor allem in Unternehmen jedoch sehr hoch ist. Die Beratungen schulen ihre Beobachter deutlich länger hinsichtlich der Dimensionen (*Kriterium 9*). Das Kriterium wird von beiden Gruppen als sehr sinnvoll aber relativ schwer umsetzbar bewertet, wobei die Unternehmen es, entsprechend ihrer kürzeren Schulungszeit, als noch schwieriger umsetzbar einschätzen. *Kriterium 10* wird von allen Beratungen erfüllt, einige Unternehmen verwenden zu wenige oder gar keine Informationsquellen. Trotzdem wird es von beiden Gruppen als sehr sinnvoll und eher schwierig umsetzbar bewertet. Insgesamt werden die Kon-



sultationen von Vorgesetzten und HR-Verantwortlichen am häufigsten als Informationsquellen genutzt. Es wäre wünschenswert, dass weitere Informationsquellen (wie z. B. Konsultationen von Stelleninhabern sowie standardisierte Methoden der Arbeitsplatzanalyse) öfter herangezogen würden, um die Anforderungsdimensionen und deren Ausprägungen genauer definieren zu können. Mehr als die Hälfte der befragten Organisationen erfüllen *Kriterium 11*. Im Vergleich zu den anderen Kriterien wird es jedoch von beiden Gruppen als wenig sinnvoll und als relativ schwer umsetzbar bewertet. *Kriterium 12* wird von rund 71% der Gesamtgruppe erfüllt, obwohl es im Vergleich zu den anderen Kriterien als eher weniger sinnvoll und als relativ schwer umzusetzen bewertet wird. Knapp die Hälfte der Unternehmen führt keine statistische Evaluation durch (*Kriterium 13*), im Gegensatz zu 35,7% der Beratungen. Dementsprechend beurteilen die Beratungen dieses Kriterium als etwas sinnvoller und leichter umzusetzen, dennoch wird es von beiden Gruppen als das am schwierigsten umsetzbare Kriterium bewertet. Hier zeigt sich deutlich ein Defizit in der AC-Praxis, da eine statistische Evaluation die Grundvoraussetzung für die kontinuierliche Qualitätskontrolle und –verbesserung eines ACs ist.

Insgesamt zeigt sich, dass die befragten Organisationen im Mittel 7,93 Kriterien erfüllen. Tendenziell schneiden die Beratungen hier besser ab als die Unternehmen, dieser Unterschied ist jedoch statistisch nicht signifikant. Dieses Ergebnis zeigt einen positiven Trend der Qualitätssteigerung in der AC-Praxis und repliziert überwiegend die Befunde von Kanning et al. (2007). Dennoch lassen sich auch in einigen Bereichen Defizite feststellen, wie zum Beispiel bei der Nutzung unterschiedlicher Informationsquellen zur Festlegung eines Anforderungsprofils oder bei der statistischen Evaluation des ACs. Hier gibt es vor allem bei Unternehmen Verbesserungsbedarf.

Bei der Interpretation der Ergebnisse müssen einige Einschränkungen der Studie beachtet werden. Zum einen wurden die befragten Organisationen nicht randomisiert ausgewählt und die Rücklaufquote des Fragebogens war sehr gering, sodass die Ergebnisse möglicherweise ein verzerrtes Bild der Realität wiedergeben. Organisationen, in denen die Qualität des ACs ohnehin eine große Rolle spielt, nehmen mit höherer Wahrscheinlichkeit an einer Befragung zur Qualität ihres ACs teil. Diese Selbstselektion der teilnehmenden Organisationen kann zur Verzerrung der Ergebnisse in positiver Richtung beitragen. Es handelt sich hier also um eine anfallende Stichprobe, die nicht als repräsentativ für alle Organisationen im deutschsprachigen Raum betrachtet werden kann. Zum anderen bezieht sich die Beurteilung eines Kriteriums meist nur auf ein einzelnes Item aus dem Fragebogen und aus jeder Organisation wurde jeweils nur eine Person befragt, daher konnte keine Beurteilungsübereinstim-

mung berechnet werden und es kann nicht ausgeschlossen werden, dass die Auskunftspersonen unter Umständen die Realität der AC-Praxis in ihrer Organisation verzerrt wiedergegeben haben. Zudem ist die Auswahl der hier untersuchten Kriterien mitnichten vollständig. Um den Umfang der Befragung nicht zu sprengen, wurde der Schwerpunkt auf Kriterien gesetzt, die im Bezug zu den Anforderungsdimensionen im AC stehen. Schließlich ergaben sich durch Probleme bei der Fragebogenkonstruktion Inkompatibilitäten bei Kriterium 3 und Item 17 sowie bei Kriterium 7 und Item 28 (siehe Tabelle A). Die Formulierung der Items stimmt hier nicht vollständig mit der der Kriterien überein, so dass diese Ergebnisse mit Vorsicht interpretiert werden müssen.

Zukünftige Studien sollten im Optimalfall mindestens zwei AC-Experten aus jeder Organisation befragen, um die Beurteilungsgenauigkeit zu erhöhen. Außerdem wäre es interessant, die Gründe für eine eventuell mangelnde Umsetzbarkeit in der Praxis genauer zu ergründen. Schließlich sollte die Befragung von AC-Experten aus der Praxis in Zukunft auch mit dem Schwerpunkt auf Kriterien aus anderen Bereichen ausgeweitet werden.

## Literatur

- Arbeitskreis Assessment Center e.V. (2004). *Standards der Assessment-Center-Technik*. Verfügbar unter: <http://www.arbeitskreis-ac.de/index.php/projekte/standards.html> [12.01.2010].
- Eck, C.D., Jöri, H. & Vogt, M. (2007). *Assessment-Center*. Heidelberg: Springer.
- Gaugler, B. B., Rosenthal, D. B., Thornton, G. C., III, & Bentson, C. (1987). Meta-analysis of assessment center validity. *Journal of Applied Psychology*, 72, 493–511.
- Gaugler, B.B. & Thornton, G. C., III. (1989). Number of Assessment Center Dimensions as Determinant of Assessor Accuracy. *Journal of Applied Psychology*, 74(4), 611-618.
- Hennessy, J., Mabey, B., & Warr, P. (1998). Assessment centre observation procedures: An experimental comparison of traditional, checklist and coding method. *International Journal of Selection and Assessment*, 6, 222–231.
- Howard, A. (1997). A Reassessment of Assessment Centers: Challenges for the 21th Century. *Journal of Social Behavior and Personality*, 12, 13-52.
- Höft, S. (2009). Gutes AC, schlechtes AC? Ein kritischer Kommentar zu den wissenschaftlichen Qualitätskriterien für Assessment Center von Kanning et al. (2007). *Zeitschrift für Arbeits- und Organisationspsychologie*, 53(2), 70-84.
- Höft, S. & Funke, U. (2001). Simulationsorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (S. 135-173). Göttingen: Hogrefe.
- Höft, S. & Melchers, K. (im Druck). Training von AC-Beobachtern: Worauf kommt es an? In Arbeitskreis Assessment Center e.V. (Hrsg.), *Diagnostische Kompetenz: Entwickeln und Anwenden. Dokumentation zum 7. Deutschen Assessment-Center-Kongress 2008* (S. 80-96). Lengerich: Pabst Science Publishers.
- International Task Force on Assessment Center Guidelines (2000). Guidelines and ethical considerations for assessment center operations. *Public Personnel Management*, 29, 315–331.
- Kanning, U. P. (2004). *Standards der Personaldiagnostik. Ein Lehrbuch für die Praxis*. Göttingen: Hogrefe.
- Kanning, U.P., Pöttker, J. & Gelléri, P. (2007). Assessment Center-Praxis in deutschen Großunternehmen. Ein Vergleich zwischen wissenschaftlichem Anspruch und Realität. *Zeitschrift für Arbeits- und Organisationspsychologie* 51(4), 155-167.
- Kersting, M. (1998). Differentielle Aspekte der sozialen Akzeptanz von Intelligenztests und Problemlöseszenarien. *Zeitschrift für Arbeits- und Organisationspsychologie*, 42, 61-75.
- Kersting, M. (2009). Qualität im Assessment Center. *Zeitschrift für Arbeits- und Organisationspsychologie*, 53(2), 70-84.
- Kleinmann, M. (1997). *Assessment-Center. Stand der Forschung - Konsequenzen für die Praxis*. Göttingen: Hogrefe.
- Kleinmann, M. (2003). *Assessment-Center. Reihe Praxis der Personalpsychologie*. Göttingen: Hogrefe.
- Krause, D. E., Meyer zu Kniedorf, C. & Gebert, D. (2001). Aktuelle Trends in der Assessment Center Anwendung. *Wirtschaftspsychologie*, 3, 47– 55.

- Krause, D. E. & Gebert, D. (2003). A comparison of Assessment Center practice in organizations in german-speaking regions and the United States. *International Journal of Selection and Assessment*, 11, 297– 312.
- Krause, D. E., Kersting, M., Heggstad, E.D. & Thornton III, G. C. (2006). Incremental Validity of Assessment Center Ratings Over Cognitive Ability Tests: A Study at the Executive Management Level. *International Journal of Selection and Assessment*, 14(4), 360-371.
- Krause, D.E. (2009). “Ob Kritik konstruktiv wird, entscheidet der Kritisierte”. Kritische Reflexion der AC-Standards. *Zeitschrift für Arbeits-und Organisationspsychologie*, 53(2), 70-84.
- Krause, D. E. & Thornton III, G. C. (2009). A Cross-Cultural Look at Assessment Center Practices: Survey Results from Western Europe and North America. *Applied Psychologie: An International Review*, 58(4), 557-585.
- Lievens, F. & Conway, J. M. (2001). Dimension and exercise variance in assessment center scores: A large evaluation of multitrait-multimethod studies. *Journal of Applied Psychology*, 86, 2 – 12.
- Obermann, C. (2006). *Assessment Center. Entwicklung, Durchführung, Trends*. Wiesbaden: Gabler.
- Obermann, C., Höft, S. & Janke, O. (in Druck). Was machen die DAX 100-Unternehmen? Die große AC-Studie 2008. In Arbeitskreis Assessment Center e.V. (Hrsg.), *Personal-diagnostik: Entwickeln und Anwenden. Dokumentation zum 7. Deutschen Assessment-Center-Kongress 2008*. Lengerich: Pabst Science Publishers.
- Reilly, R. R., Henry, S. & Smither, J.W. (1990). An examination of the effects of using behavior checklists on the construct validity of assessment center dimensions. *Personnel Psychology*, 43, 71– 84.
- Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practice and theoretical implications of 85 years of research findings. *Psychological Bulletin*, 124, 262 – 274.
- Schuler, H. (2001). *Lehrbuch der Personalpsychologie*. Göttingen: Hogrefe.
- Schuler, H., Hell, B., Trapmann, S., Schaar, H. & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie*, 6(2), 60-70.
- Thornton, G. C., III, Byham, W.C. (1982). *Assessment centers and managerial performance*. New York: Academic Press.
- Thornton, G. C., III, Gaugler, B. B., Rosenthal, D. B. & Bentson, C. (1992). *Die prädiktive Validität des Assessment Centers - eine Metaanalyse*. In H. Schuler & W. Stehle (Hrsg.), *Assessment Center als Methode der Personalentwicklung* (2. Aufl., S. 36-77). Göttingen: Hogrefe.
- Thornton, G.C. III, & Rupp, D.R. (2006). *Assessment centers in human resource management: Strategies for prediction, diagnosis, and development*. Mahway, NJ: Erlbaum.
- Woehr, D. J. & Huffcutt, A. I. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67(3), 189-205.
- Woehr, D. J. & Roch, S. G. (1996). Context effects in performance evaluation: The impact of ratee sex and performance level on performance ratings and behavioral recall. *Organizational Behavior and Human Decision Processes*, 66(1), 31-41.

Woehr, D.J., & Arthur, W. (2003). The construct-related validity of assessment center ratings: A review and meta-analysis of the role of methodological factors. *Journal of Management*, 29, 231–258.

## Anhang

*Tabelle A. Kriterien und Fragen*

Kriterium	Frage
1 Im AC sollten insgesamt zwischen 3 und 6 unterschiedliche Dimensionen erfasst werden. (Gaugler & Thornton, 1989)	7 Wie viele Dimensionen werden in Ihrem AC erfasst?
2 Die Dimensionen sollten durch verhaltensorientierte Items erfasst werden. (Krause & Gebert, 2003)	16 Welche Art von Items wird in Ihrem AC verwendet? a) verhaltensorientierte Items b) eigenschaftsorientierte Items c) sowohl verhaltens- als auch eigenschaftsorientierte Items
3 Eine Dimension sollte durch vier bis acht Items pro Übung erfasst werden. (Eck et al., 2007)	17 Mit wie vielen Items werden die einzelnen Dimensionen in Ihrem AC im Durchschnitt gemessen?
4 Ein AC sollte zwischen drei und sechs verschiedene Übungen beinhalten. (Thornton et al. 1992)	18 Wie viele Übungen sind in Ihrem AC enthalten?
5 Eine Dimension sollte in mindestens zwei unabhängigen Übungen beobachtet werden. (Kanning, 2004)	21 In wie vielen Übungen wird eine Dimension in Ihrem AC durchschnittlich beobachtet?
6 Innerhalb einer Übung sollten nicht mehr als vier Dimensionen gleichzeitig beobachtet werden. (Eck et al., 2007)	20 Wie viele Dimensionen werden in Ihrem AC durchschnittlich pro Übung erfasst?
7 Das AC sollte hinsichtlich der zu erfassenden Dimensionen auf die spezifischen Anforderungen des Unternehmens zugeschnitten sein. (Krause et al., 2001)	28 Wie ist das AC auf die spezifischen Anforderungen des Unternehmens zugeschnitten worden? a) das AC wurde für das Unternehmen neu konzipiert b) eine Standardlösung wurde an das Unternehmen angepasst c) eine Standardlösung wurde übernommen
8 Das Verhältnis zwischen AC-Teilnehmern und AC-Beobachtern sollte maximal 2:1 sein. (Kanning et al., 2007)	23 Wie viele Beobachter gibt es durchschnittlich bei einer typischen Gruppenübung? 24 Wie viele Teilnehmer gibt es durchschnittlich bei einer typischen Gruppenübung?
9 Die Beobachter sollten vor dem AC intensiv bezüglich der Dimensionen geschult werden. (Lievens & Conway, 2001)	24 Z Zusatzitem: Verhältnis Teilnehmer : Beobachter 25 Führen Sie ein Beobachtertraining durch? 26 Wie viele Stunden dauert das Beobachtertraining insgesamt? 27 Wie viele Stunden entfallen dabei auf die Beschäftigung mit den Dimensionen?
10 Die zu erfüllenden Ausprägungen auf den Dimensionen sollten durch unterschiedliche Informationsquellen festgelegt werden. (Schuler, 2001)	29 Wie wird festgelegt, welche Ausprägung erfolgreiche Teilnehmer hinsichtlich der Dimensionen im AC haben sollen? a) Konsultation von Stelleninhabern b) Konsultation von Vorgesetzten c) Konsultation von HR-Verantwortlichen im Unternehmen d) Konsultation von externen HR-Beratern e) standardisierte Methode der Arbeitsplatzanalyse f) empirische Analyse g) Nutzung von Forschungsbefunden h) andere Vorgehensweise i) die Ausprägung wird nicht festgelegt 29 Z Zusatzitem: Summe der Quellen
11 Die zu erfüllende Ausprägung auf jeder Dimension sollte, je nach dem um welche Dimension es sich handelt, gesondert festgelegt werden.	30 Die Ausprägungen für die verschiedenen Dimensionen sind a) für alle Dimensionen identisch oder b) je nach Dimension variabel.
12 Die zu erfüllende Ausprägung auf jeder Dimension sollte, je nach dem um welche Position und/oder Unternehmenseinheit es sich handelt, gesondert festgelegt werden. (Krause et al., 2001)	31 Die Ausprägungen für die verschiedenen Dimensionen werden für a) jede Position oder Funktion b) jede Unternehmenseinheit c) das gesamte Unternehmen definiert.

Kriterium	Frage
13 Die Dimensionen des AC werden nach der Durchführung statistisch evaluiert (Arbeitskreis Assessment Center e.V., 2004)	32 Wie werden die AC Dimensionen und deren Items nach der Durchführung statistisch evaluiert? a) Faktorenanalyse b) Trennschärfe c) Beobachterübereinstimmung d) Schwierigkeit e) Zusammenhang mit Berufserfolg f) Zusammenhang mit Arbeitszufriedenheit g) andere h) keine statistische Evaluation

Tabelle B.1. Ergebnisse Befragung Teil 1 (Mittelwerte und Mittelwertsunterschiede)

Kriterium	Item	Gesamt					Unternehmen					Beratung					Teststatistik	
		N	M	SD	min	max	N	M	SD	min	max	N	M	SD	min	max	U	p
1	7	55	8,76	5,56	3	37	41	9,04	6,04	3	37	14	7,93	3,89	3	16	266,5	,691
3	17	54	7,33	9,47	0	70	40	7,16	10,85	0	70	14	7,82	3,48	4	16	160,5	,017*
4	18	55	5,64	2,69	1	16	41	5,33	2,41	1	16	14	6,53	3,3	3	12	238,5	,341
5	21	53	2,73	1,04	1.5	6	39	2,59	0,87	1.5	5	14	3,11	1,36	2	6	216,5	,213
6	20	54	4,24	1,35	2	8	40	4,34	1,42	2	8	14	3,96	1,12	2	6	257,5	,647
8	23	52	3,36	1,53	2	8	39	3,49	1,64	2	8	13	2,96	1,13	2	6	213	,374
	24	52	5,15	1,92	1	10	39	5,08	1,86	1	10	13	5,39	2,14	3	10	246,5	,879
9	26	55	4,07	4,01	0	16	41	3,23	3,72	0	16	14	6,54	3,93	2	16	118	,001**
	27	54	2,07	2,59	0	13	40	1,49	2,0	0	10	14	3,71	3,37	0,5	13	116	,001**

Anmerkungen. \* =  $p < ,05$ ; \*\* =  $p < ,01$  (zweiseitig). M = Mittelwert, SD = Standardabweichung, min = Minimum, max = Maximum, U = Prüfgröße des Mann-Whitney-Tests (Vergleich Unternehmen vs. Beratungen),  $p$  = Signifikanzniveau (zweiseitig).

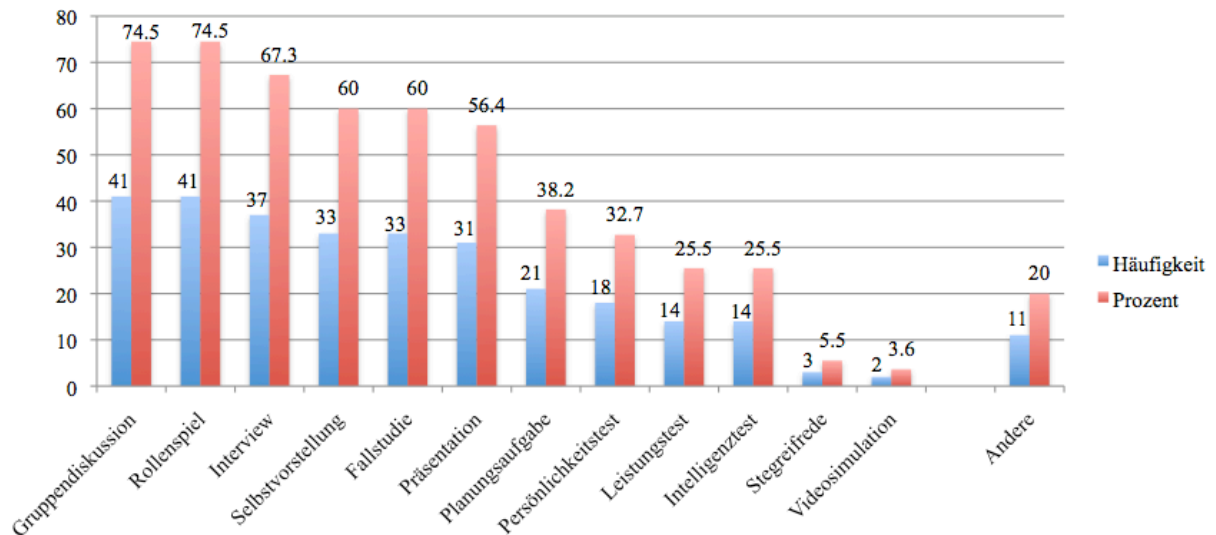
Tabelle B.2. Ergebnisse Befragung Teil 1 (Häufigkeiten)

Kriterium	Item	Gesamt		Unternehmen		Beratung		
		H	%	H	%	H	%	
2	16	a	19	34,5	12	29,3	7	50
		b	2	3,6	2	4,9	0	0
		c	34	61,8	27	65,9	7	50
7	28	a	46	83,6	37	90,2	9	64,3
		b	8	14,5	3	7,3	5	35,7
		c	1	1,8	1	2,4	0	0
10	29	a	17	30,9	10	24,4	7	50
		b	41	74,5	30	73,2	11	78,6
		c	42	76,4	30	73,2	12	85,7
		d	15	27,3	10	24,4	5	35,7
		e	11	20	8	19,5	3	21,4
		f	6	10,9	5	12,2	1	7,1
		g	6	10,9	5	12,2	1	7,1
		h	5	9,1	5	12,2	0	0
		i	5	9,1	5	12,2	0	0

11	30	a	23	41,8	17	41,4	6	42,9
		b	32	58,2	24	58,5	8	57,1
12	31	a	38	69,1	29	70,7	9	64,3
		b	7	12,7	5	12,2	2	14,3
		c	16	29,1	12	29,3	4	28,6
13	32	a	10	18,2	6	14,6	4	28,6
		b	10	18,2	8	19,5	2	14,3
		c	30	54,5	21	51,2	9	64,3
		d	6	10,9	5	12,2	1	7,1
		e	15	27,3	10	24,4	5	35,7
		f	7	12,7	5	12,2	2	14,3
		g	0	0	0	0	0	0
		h	25	45,5	20	48,8	5	35,7

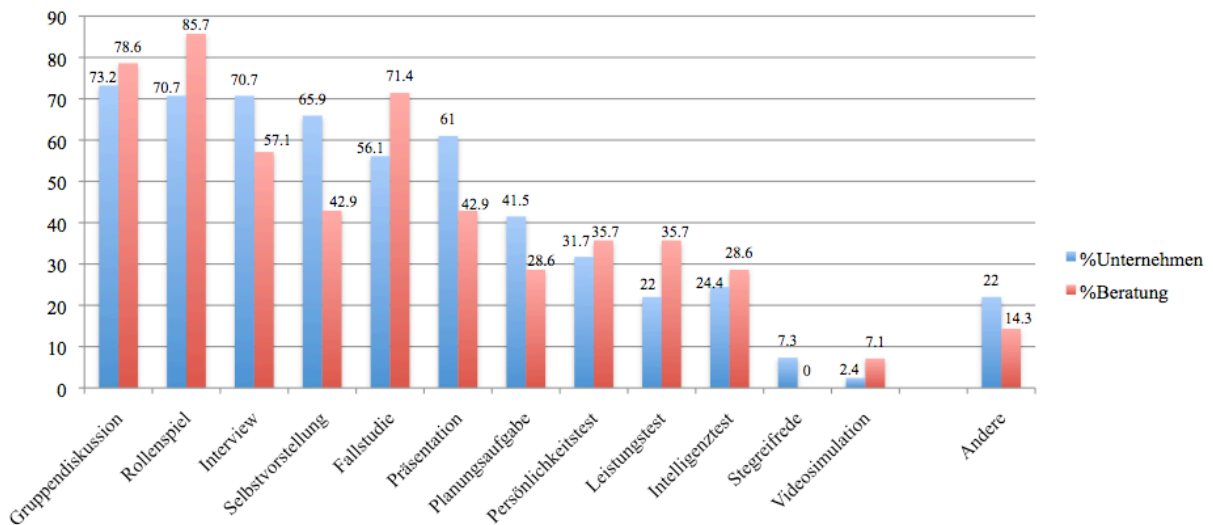
Anmerkungen. H = Häufigkeiten, bei den Items 29, 31 und 32 waren Mehrfachnennungen möglich.

Abbildung A. Übungen im AC gesamt



Anmerkung. Mehrfachnennungen möglich

Abbildung B. Übungen AC Vergleich Unternehmen und Beratungen



Anmerkung. Mehrfachnennungen möglich



*Tabelle C. Anzahl Beobachter in Einzelübung*

	Gesamt					Unternehmen					Beratungen				
	N	M	SD	min	max	N	M	SD	min	max	N	M	SD	min	max
Beobachter in Einzelübung	55	2,66	1,31	1	8	41	2,72	1,40	1	8	14	2,46	1,01	1	4

*Anmerkungen.* M = Mittelwert, SD = Standardabweichung, min = Minimum, max = Maximum.