

Möglichkeiten der

Spielwiese der Laiendiagnostik“, „Schwätzer bevorzugt“, „Job-anwärter im Affenzirkus“ – so oder ähnlich lauten die Schlagzeilen aktueller Artikel, die sich mit Assessment-Centern (AC) auseinandersetzen und auf die geringe Treffsicherheit dieser Verfahren hinweisen. AC werden seit Jahren wie das viel zitierte Borstenvieh durch das ‚Personaler-Dorf‘ gejagt. Aber welche Kritik ist berechtigt und welche überzogen? Martin Kersting fragt danach, wie die aktuelle AC-Praxis aussieht, welche Trends es gibt und wie es um die durchschnittliche Aussagekraft von AC bestellt ist. Das führt den Autoren zu der entscheidenden Frage, wie man im Einzelfall mit vertretbarem Aufwand ein qualitativ hochwertiges AC realisieren und dessen Qualität überprüfen kann.



Zur Zukunft von Assessment-Center-Verfahren

Qualitätssicherung und -optimierung

Das AC ist tot – es lebe das AC! Entgegen dem anderslautenden Rumor der angeblich nachlassenden Einsatzhäufigkeit erfreuen sich AC einer steigenden Beliebtheit. Dies ergibt eine Studie des Arbeitskreises Assessment Center (AK-AC), die im September 2008 präsentiert wurde (Obermann / Höft 2008). Danach setzen 70 Prozent der DAX-100-Unternehmen AC ein. Sowohl über alle Organisationen hinweg als auch innerhalb einer Organisation nimmt die Einsatzhäufigkeit von AC zu. Die traditionelle Methode wird allerdings häufig hinter modischen Bezeichnungen wie ‚Executive Day‘ oder ‚Talentfinder‘ versteckt, nur noch in der Minderheit der Fälle wird das Kind mit Namen genannt. Dies mag mit der Sorge um die Akzeptanz von AC zu tun haben. Weitere Ergebnisse der Studie, an der 233 Organisationen aus dem deutschsprachigen Raum teilgenommen haben, lassen sich so zusammenfassen:

- Nur 37 Prozent der AC dienen allein der Personalauswahl, die Mehrzahl der AC zielt auf die Potenzialanalyse und / oder soll die Personalentwicklung unterstützen. Angesichts der Schwierigkeiten, neue Mitarbeiter zu rekrutieren, besinnen sich die Organisationen offensichtlich darauf, die eigenen Mitarbeiter zu fördern.
- Beobachtertrainings sind Standard, werden aber (im Vergleich zur Befragung aus dem Jahr 2001) kürzer: In 64 Prozent der Fälle dauert das Training maximal einen halben Tag. Auch das im Anschluss an das AC folgende Feedback an die Teilnehmer – erfreulicherweise ebenfalls für fast alle Organisationen eine Selbstverständlichkeit – wird kürzer und dauert häufig (in 45 % der Fälle) nur maximal 30 Minuten.
- Im Vergleich zur ersten Befragung (2001) hat die Gruppendiskussion ihren bisherigen Spitzenplatz innerhalb der eingesetzten Übungen eingebüßt, demgegenüber werden immer häufiger Tests zur kognitiven Kompetenz (in 31 % aller AC) sowie Persönlichkeitsfragebogen (28 %) in das AC integriert.

- Stark zugenommen hat das sogenannte Einzel-AC, also ein AC mit nur einem Kandidaten. Diese Entwicklung hat möglicherweise ebenfalls mit Akzeptanzerwägungen zu tun, da man den Kandidaten den direkten Vergleich mit internen oder externen Mitbewerbern ersparen will.

Umstrittene Aussagekraft

AC sind also beliebter als je zuvor – aber taugen sie auch etwas, und handelt es sich noch um klassische AC? Von entscheidender Bedeutung ist die Kriteriumsvalidität, also die Frage, ob die Beurteilung im AC mit der später gezeigten Ausbildungs- und / oder Berufsleistung übereinstimmt. Der Grad der Übereinstimmung wird mit dem sogenannten Korrelationskoeffizienten (r) ausgedrückt, der zwischen 0.0 (kein Zusammenhang) und 1.0 (perfekter Zusammenhang) variiert, wobei im Bereich der Personaldiagnostik Korrelationen im Bereich um die $r = .50$ als hervorragend gelten.

In der berühmten Studie der American Telephone and Telegraph Company (AT&T) (Bray / Grant 1966) wurde für das AC eine Treffsicherheit in Höhe von $r = .46$ bestimmt. Während sich dieser Wert aus einer Einzelstudie errechnete, ist es seit einigen Jahrzehnten mithilfe der Metaanalyse möglich, Ergebnisse aus verschiedenen Studien quantitativ zusammenzufassen. In der ersten Metaanalyse zur Kriteriumsvalidität von AC, bei der 50 einzelne Bewährungskontrollen berücksichtigt wurden, bestimmten Thornton et al. (1987) einen Koeffizienten von $r = .37$. Demgegenüber konnten in zwei aktuellen Metaanalysen nur



Professor Dr. Martin Kersting ist Diplom-Psychologe und seit 2008 Professor an der Fachhochschule des Bundes in Münster. Kersting ist Mitautor der DIN 33430 und Autor verschiedener Tests zu kognitiven und sozialen Kompetenzen sowie zahlreicher Publikationen.

deutlich geringere Validitäten ermittelt werden.

Dies ist ein Grund dafür, dass AC aktuell eine derart kritische Presse haben. Hermelin et al. (2007) bestimmten anhand von 26 Studien, die in der Zeit von 1985 bis 2005 (also nach der von Thornton et al. erstellten Studie) publiziert wurden, eine AC-Kriteriumsvalidität von $r = .28$. Hardison und Sackett (2007) berücksichtigten in ihrer Metaanalyse ebenfalls aktuellere Studien und kamen nur auf eine (aus 106 Einzelwerten gemittelte) Kriteriumsvalidität von $r = .26$. Der Validitätsverlust (siehe Abb. 1) gegenüber der Studie aus dem Jahr 1966 ist dramatischer, als es auf den ersten Blick aussieht, da die Werte der Metaanalysen durch (inhaltlich begründete) Korrekturrechnungen erhöht sind, während Bray und Grant (1966) keine Optimierungen der Schätzung vorgenommen haben. Gute Nachrichten für AC-Befürworter liefert hingegen eine aktuelle Studie von Holzenkamp, Spinath und Höft (im Druck). Die Autoren fassten mithilfe der Metaanalyse die Ergebnisse aus 19 deutschsprachigen AC-Studien zusammen, die im Zeitraum von 1987 bis 2007 durchgeführt wurden. Dabei ergab sich eine Kriteriumsvalidität von $r = .36$ (unkorrigiert .33).

Selbst bei einer Treffsicherheit von $r = .36$ muss das AC aber unter Aufwand-Nutzen-Gesichtspunkten kritisch betrachtet werden, weniger aufwendige Verfahren wie strukturierte Interviews ($r = .44$; McDaniel et al. 1994) und

vor allem Intelligenztests ($r = .64$; Salgado et al. 2003) erzielen weit aus bessere Werte.

Interessant ist aber nicht nur, was es Neues gibt, sondern auch, ob das Alte schon bekannt ist. Während die enttäuschenden Befunde zur Kriteriumsvalidität die

geben und die Dimensionen innerhalb einer Übung nicht hinreichend differenziert beurteilen.

Die praktische Bedeutung dieses Befundes, der gern als ‚akademisch‘ abgewertet wird, ist sehr weitreichend. Gerade der Einsatz von AC zur Personalentwicklung



Praktiker verunsichern, ist ein anderer Befund, der seit Langem bekannt ist, in der Praxis bislang weitgehend unbeachtet geblieben: AC sind nicht konstruktvalide, das heißt, sie messen nicht das, was sie zu messen vorgeben (siehe z. B. Bowler / Woehr 2006). Theoretisch sollen die einzelnen Anforderungsdimensionen (z. B. Kommunikationsfähigkeit und Führungskompetenz) in verschiedenen Übungen, etwa in einer Gruppendiskussion und im Rollenspiel, erfasst werden. Die Analysen zeigen hingegen immer wieder, dass die Assessoren eher globale Urteile über das Verhalten pro Übung ab-

setzt voraus, dass die Diagnose der Stärken und Entwicklungsbereiche ‚stimmt‘, damit die abgeleiteten Fördermaßnahmen greifen. Wenn aber beispielsweise tatsächliche Schwächen im analytischen Denken fehlerhaft als Entwicklungsbedarf in der sozialen Kompetenz diagnostiziert werden, dann ist die Investition in das nachgeschaltete Gruppentraining sozialer Kompetenzen von vornherein vergeblich. Die Problematik der mangelhaften Konstruktvalidität von AC ist nicht nur ein Thema für Bewohner des Elfenbeinturms, sondern sollte vor allem Praktiker konstruktiv beunruhigen.

Andererseits sollten Wissenschaftler es sich nicht zu leicht machen, die Schuldenkonten der AC-Praxis zu saldieren. Während die Wissenschaft bei ihrer Bewertung von AC häufig allein die Validität betrachtet, sind in der Praxis häufig simultan mehrere, teilweise kontradiktorische Kriterien zu berücksichtigen.

Erfolgsfaktoren für das AC

Die Aussagekraft von AC ist im Allgemeinen geringer als die anderer, weniger aufwendiger Verfahren. Wie aber sieht es im Einzelfall aus? In der aktuellen Diskussion wird häufig vergessen, dass es sich bei den meta-analytisch bestimmten Korrelationskoeffizienten um die Mittelwerte aus zahlreichen Einzelwerten handelt,



Die Fotos auf diesen Seiten entstanden im Rahmen eines AC bei der Altana Chemie in Schermbeck.

Dies zeigt sich gerade im AC, welches zusätzlich zur Diagnose einen erheblichen Nebennutzen für die Organisationen erbringt. Mit dem für das AC betriebenen Aufwand sowie der Einbindung der Führungsebene wird zum Beispiel der Stellenwert der Personalarbeit demonstriert; durch das AC werden das Anforderungsprofil und die Führungskultur kommuniziert, außerdem ermöglicht das AC den Teilnehmern eine Selbsteinschätzung im Vergleich zu anderen Teilnehmern. Einer so differenzierten Zielstellung wird man nicht durch die alleinige Bestimmung der Kriteriumsvalidität gerecht, sondern man muss sich die Mühe machen, multiattributive Nutzenmodelle zu erstellen.

die beträchtlich streuen. Insofern gibt es – vereinfacht ausgedrückt – erhebliche Qualitätsunterschiede zwischen den einzelnen AC. Statt nun das Kind mit dem Bade auszukippen, sollte man die Ergebnisse besser wie folgt interpretieren: Bei den meisten AC bleibt Raum für Verbesserungen.

Wie aber sind diese Verbesserungen zu erzielen? AC sind komplexe Verfahren, es ist an dieser Stelle nicht möglich, auf alle Qualitätselemente einzugehen. Orientierung geben unter anderem die Qualitätsstandards 1) des AK-AC (2004), 2) der Task Force on AC (2000) sowie 3) vor allem der DIN 33430 zur Personalbeurteilung, die mittlerweile in Form von praktischen Checklisten aufbereitet ist (Kersting 2008a).¹ Im Folgenden werden ausgewählte, nach der Erfahrung des Autors besonders häufig vernachlässigte Qualitätsaspekte angesprochen, wobei sich die Ausführungen in die Prozessphasen Planung, Durchführung und Evaluation gliedern.

Planung Bezüglich der Planung sollten die Aspekte Übungskonzeption, Beurteilungsdimensionen und Methodenvielfalt betrachtet werden. Die AK-AC-Studie zeigt, dass immer seltener standardisierte AC zum Einsatz kommen, sondern dass Organisationen verstärkt dazu übergehen, AC selbst zu entwickeln. Dies ist allerdings nur dann eine gute Nachricht, falls die Verfahrensentwicklung professionell erfolgt. Wenn aber die Zahl der AC steigt, während gleichzeitig die Zahl der für Konzeption und Durchführung qualifizierten Personaler stagniert, scheint die Sorge von Schuler (2007) berechtigt, der das AC zur Spielwiese der Laiendiagnostik verkommen sieht. Gerade die AC-Entwicklung stellt eine Herausforderung dar, zu deren Bewältigung es umfassender Expertise bedarf. Während es mittlerweile selbstverständlich ist, dass die Assessoren ein Beobachtertraining absolvieren müssen, ist es immer noch unüblich, dass die AC-Entwickler für ihre Aufgabe qualifiziert werden.

Ein Qualitätsmerkmal von AC ist laut der Analyse von Thornton et al. (1987) der Einsatz einer großen Anzahl verschiedener Übungen. Diese Verfahrensvielfalt lässt sich aber nicht mit dem Wunsch nach immer kürzeren Verfahren vereinbaren.

Wichtig ist darüber hinaus, dass nur wenige, trennscharfe, aus der Anforderungsanalyse abgeleitete und gut operationalisierte Anforderungsdimensionen bewertet werden. Um Urteilsfehler durch Überlastung und Komplexität zu vermeiden, empfiehlt der AK-AC in seinen Standards (2004), in jeder Übung maximal fünf (besser aber nur maximal drei) Anforderungsdimensionen bewerten zu lassen. Insgesamt ergeben sich für die Übungskonstruktion folgende Erfolgsfaktoren:

- Die Übungskonstruktion sollte anforderungsorientiert sein, erfolgskritisch und realitätsorientiert (Simulationsprinzip) erfolgen.
- Die Übungskonstruktion sollte dimensionsorientiert erfolgen (ist die Übung geeignet, das gewünschte Verhalten, zum Beispiel „Durchsetzungsstärke“, zu provozieren?);
- maximal drei bis fünf zu bewertende Dimensionen pro Übung,
- Vermeidung der Bevorzugung einzelner Teilnehmer (z. B. durch zu fachspezifische Gestaltung),
- sorgfältige Gestaltung der Materialien und Instruktionen,
- bei Rollenspielen Einsatz von ‚Drehbüchern‘ und Trainings für die Rollenspieler,
- Wahrung einer angemessenen Schwierigkeit (weder zu leicht noch zu schwer).

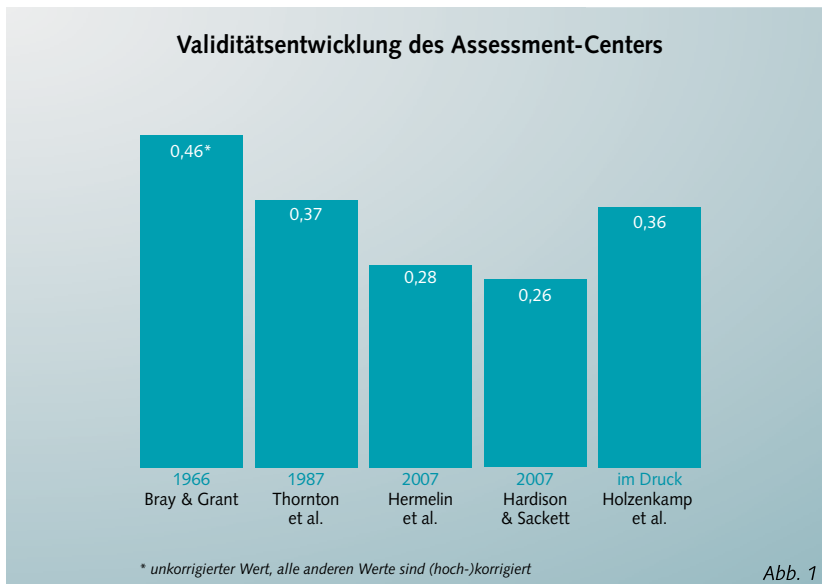
Ob die Übungskonstruktion gelungen ist, ist durch Probedurchgänge zu prüfen, wobei solche Personen in die Rolle der Teilnehmer schlüpfen, die den Zielpersonen des AC möglichst ähnlich sind. Die Auswertung der Probedurchgänge liefert erste Hinweise auf die Funktionstüchtigkeit, Differenzierungsfähigkeit und Schwierigkeit der Übungen.

Darüber hinaus sollte die Grundidee des AC, nämlich die Methodenvielfalt, umgesetzt werden, indem die Übungen (Rollenspiel, Gruppendiskussion usw.) durch psychometrische Tests ergänzt werden. Zu nennen sind hier in erster Linie Intelligenztests, die nachweislich besonders aussagekräftig sind. Sofern Intelligenztests, wie es der aktuelle Trend zeigt, in AC integriert werden, gelingt es vielleicht wieder, zu den Qualitä-

ten der historischen AC aufzuschließen. Früher war es üblich, im AC situative Übungen mit Intelligenztests zu kombinieren. Tatsächlich beruht die oben zitierte, mittlerweile legendär gewordene Treffsicherheit des klassischen AT&T-AC, wie Hossiep (2001) zeigt, wesentlich auf den Fähigkeitstests. Das AC wäre – anders ausgedrückt – wesentlich weniger nützlich gewesen, hätte AT&T auf Tests verzichtet. Dies zeigte sich auch in der aktuellen Studie von Holzenkamp et al. (im Druck). AC mit Intelligenztests erzielen eine Treffsicherheit von $r = .52$, während die AC, die auf Intelligenztests verzichten, auf einen Wert von $r = .23$ zurückfallen.

Ein Grund für die in Deutschland (und fast nur in Deutschland) vorherrschende Test-Phobie ist die Sorge um die mangelhafte Akzeptanz. Dabei schätzen Personen, die schon einmal getestet wurden, die hohe Messqualität von Tests (siehe Kersting 2008b) und begrüßen es, dass eine Organisation sich um ein objektives Bild der Stärken und des Entwicklungsbedarfs einer Person bemüht. Testergebnisse sind, anders als das Interview- und AC-Ergebnis, wirklich unabhängig von Sympathie und Antipathie, von Aussehen und Geschlecht. Aus Akzeptanzgründen sollte man allerdings möglichst auf Tests mit einem semantischen Bezug auf den beruflichen Kontext zurückgreifen (anstelle von Tests mit abstraktem Aufgabenmaterial) und nicht von Intelligenztests, sondern zum Beispiel von Tests zur kognitiven Kompetenz sprechen.

Vielversprechend ist darüber hinaus die Integration von Persönlichkeitsfragebogen in das AC, zum Beispiel als Indikator für das Selbstbild und / oder als Mittel zur Hypothesengenerierung für sich an-



schließende teilstrukturierte Interviews. (Interviews sollten in den meisten Fällen ebenso ein AC-Bestandteil sein.) Obgleich die Validitätswerte für Persönlichkeitstests absolut betrachtet gering ausfallen und noch hinter den Werten für das AC zurückbleiben, können sie in Kombination mit den AC-Ergebnissen sowie mit den Intelligenztestergebnissen zu einer verbesserten Treffsicherheit beitragen. Während AC-Übungen und Intelligenztests eher das maximale Verhalten (Können) fokussieren, erfassen Interviews und Persönlichkeitsfragebogen eher die Motive (Wollen). Definiert man – wie üblich – Leistung als Produkt von Können und Wollen, darf keines der beiden Konstrukte vernachlässigt werden. Als Nachteil von Persönlichkeitsfragebogen galt lange Zeit die Verfälschbarkeit. Die Forschung hat aber gezeigt, dass die früher so sehr gefürchtete Verfälschung der Antworten die Aussagekraft von Persönlichkeitsfragebogen im Durchschnitt nicht senkt (Ones et al. 1996).

Durchführung Zu diesem Themenbereich werden im Folgenden die Aspekte Beurteilertraining, Hilfsmittel zur Unterstützung der Bewertung sowie Da-

tenintegration näher beleuchtet. Die Durchführung des AC ist so zu gestalten, dass die Assessoren ihren schwierigen Aufgaben gerecht werden können. Dies setzt zunächst ein Training voraus, wobei sich die Kombination eines Beurteilerfehlertrainings mit einem Bezugsrahmentraining oder einem Verhaltensbeobachtungstraining bewährt hat. In dem nachweislich sehr effektiven Bezugsrahmentraining geht es vor allem um die Einhaltung eines gemeinsamen Beurteilungsmaßstabes und um die Zuordnung von Verhalten zu Dimensionen. Die in AC-Trainings sonst oft gebetsmühlenartig aufgestellte Forderung nach einer Trennung von Beobachtung und Beurteilung spielt hingegen kaum eine Rolle.

Während das Bezugsrahmentraining nachweislich sehr effektiv ist, steigert das häufig praktizierte alleinige Fehlertraining (Vermeidung gängiger Beurteilungsfehler) kaum die Beurteilungsakkuratheit (siehe Höft / Funke 2006). Zur Unterstützung des Beobachtertrainings können videogestützte Kalibrierungsübungen eingesetzt werden, mit denen zum Beispiel die Einhaltung des einheitlichen Beurteilungsmaßstabes

und / oder die Zuordnung von Verhaltensweisen zu Dimensionen geübt werden können.

Weitere Qualitätsmerkmale bei der AC-Durchführung sind die Einhaltung ausreichender Beobachtungs- und Reflexionszeiten und vor allem die Nutzung von Materialien zur Unterstützung der Beobachtung und Bewertung. Häufig nährt bereits die Inspektion der AC-Unterlagen begründeten Zweifel an der Qualität des AC, beispielsweise wenn

- die Assessoren in kürzester Zeit simultan für mehrere Kandidaten Urteile über zahlreiche unterschiedliche Dimensionen treffen sollen,
- geeignete Hilfsmittel fehlen,
- für die Assessoren keine Zeit für Reflexionen und Pausen vorgesehen ist,
- die Arbeitszeit der Assessoren insgesamt so lang ist, dass die maximale Konzentrationsfähigkeit überschritten wird.

Die Studie des AK-AC hat gezeigt, dass den Beobachtern nur wenige Hilfsmittel zur Bewältigung ihrer schwierigen Aufgabe zur Verfügung gestellt werden. Nur knapp ein Drittel der Beobachter kann beispielsweise bei der Arbeit auf Verhaltensanker zurückgreifen. Anstelle von klassischen Verhaltensprotokollbogen (im Wesentlichen leere Blätter) sollten die Protokollbogen zumindest eine direkte Zuordnung der Notizen zu den Dimensionen vorsehen und übungsspezifisch ausgearbeitet sein. Darüber hinaus sollten den Assessoren Hilfsmittel wie Verhaltenschecklisten und / oder Verhaltensanker zur Verfügung gestellt werden (Höft / Lüth 2005).

Eine hohe Aufmerksamkeit sollte vor allem der Integration der Daten gewidmet werden, die häufig in sogenannten Beobachterkonferenzen vollzogen wird. Wenn einzelne Bestandteile des AC, zum Beispiel das Interview, allein eine höhere Treffsicherheit erzielen als alle AC-Bestandteile zusammen, sollte man vor al-

lem über die Zusammenfassung der Daten nachdenken, bei der offensichtlich wichtige Einzelinformationen verloren gehen. Die zu einer Dimension vorliegenden Urteile unterschiedlicher Assessoren müssen ebenso zusammengefasst werden wie die Urteile für ein und dieselbe Dimension in unterschiedlichen Übungen. Zumindest bei AC, die zur Personalauswahl eingesetzt werden, ist darüber hinaus die Bestimmung eines Gesamtwertes notwendig. Falls Tests oder Persönlichkeitsfragebogen eingesetzt werden, muss auch geklärt werden, wie diese Daten mit den AC-Daten zusammengebracht werden.

All diese Datenintegrationsprozesse sollten im Vorfeld eindeutig geregelt werden. De facto ist in der Praxis zwar häufig der Ablauf jeder Übung bis ins Detail geregelt, der Prozess der Datenintegration bleibt aber dem freien Spiel der Kräfte überlassen. Die Umfrage des AK-AC zeigt, dass hier die Konsensdiskussion überwiegt, nur in ungefähr jedem fünften AC wird das Ergebnis numerisch (z. B. durch Mittelwertbildung) ermittelt. Gegen die Konsensdiskussion wäre nichts einzuwenden, wenn zunächst jeder Assessor ein unabhängiges Urteil fällen würde und es für die Konsensdiskussion vorher festgelegte Spielregeln gäbe (z. B. die Regel: „Nach einer kurzen Diskussion im Umfang von x Minuten kann jeder Assessor auf Wunsch sein Urteil modifizieren. Sofern nach der Modifikation noch Urteilsdifferenzen bestehen, wird der Mittelwert über alle Urteile bestimmt.“). Oft ist der Konsensprozess aber nicht nachvollziehbar, das Ergebnis hängt weniger vom beobachteten Verhalten der beurteilten Person als vielmehr von Mehrheiten (Konformität), Min-

derheiten (Innovation) oder Autoritäten (Gehorsam) in der Assessorengruppe ab.

Studien zeigen, dass einfache statistische Hilfsmittel (z. B. einfache oder gewichtete Mittelwertbildung) die Qualität der Datenintegration erhöhen. Dieser Befund passt zu Studien aus der allgemeinen Forschung zu Beurteilungsprozessen, die zeigen, dass ein formalisiertes Urteils geschehen (statistische Urteilsbildung) der freien Eindrucksbildung (klinische Urteilsbildung) zumeist überlegen ist (Grove et al. 2000). Diskussionen innerhalb der Abschlussrunden haben ihre Berechtigung, um das Feedback und / oder Personalentwicklungsmaßnahmen vorzubereiten und um sicherzustellen, dass die Ergebnisse von allen Beteiligten getragen werden. Allein aus der Perspektive der Treffsicherheit sind sie aber zumindest verzichtbar, wenn nicht schädlich. Entsprechend kann sich die Konsensdiskussion auf die relevanten Aspekte konzentrieren und kürzer gestaltet werden. Die Einführung von klaren Regeln für die Integration der Beobachterdaten führt somit zur gleichzeitigen Steigerung von Qualität und Effizienz.

Evaluation, Qualitätssicherung und -optimierung In der Befragung des AK-AC gaben 76 Prozent der Anwender an, ihre AC zu evaluieren. Dies wäre eine gute Nachricht, würde auf Nachfrage nicht zutage treten, dass sich die vermeintliche Evaluation in den meisten Fällen auf eine Einschätzung der Zufriedenheit der Beteiligten reduziert. Natürlich ist es schon aufgrund der in der Regel geringen Fallzahlen nur selten möglich, eine wirkliche Bewährungskontrolle zu rechnen und die AC-Ergebnisse zum Beispiel systematisch

dem späteren Berufserfolg gegenüberzustellen. Dies bedeutet aber nicht, dass man seine Hände in den Schoß legen kann.

Ausgangspunkt der Evaluation und Qualitätsoptimierung ist die Dokumentation. Nur was dokumentiert und gemessen wird, kann auch ausgewertet und optimiert werden. Häufig wird die Dokumentation aber nur sehr nachlässig vorgenommen, es wird nur das AC-Gesamtergebnis festgehalten, die Urteile der einzelnen Assessoren (z. B. vor der Konsensbildung) werden häufig nicht dokumentiert.

Ganz wichtig für die Qualitätssicherung und -optimierung ist es darüber hinaus, das AC so zu gestalten, dass die Assessoren ein unmittelbares Feedback über die Qualität ihrer Arbeit (und nicht nur darüber, wie diese „ankommt“) erhalten. Als Beispiel kann die Bestimmung der Beurteilerübereinstimmung angeführt werden, die zum Beispiel nach DIN 33430 vorgeschrieben ist (siehe Checklistenpunkt 212 im DIN-Screen, Kersting 2008a). Was zunächst wie Schikane anmutet, erweist sich in der Praxis schnell als wichtiges Qualitätssicherungswerkzeug. Wendet man das Werkzeug in der Praxis an, so stellt man zum Beispiel fest, dass Assessoren sich über drei Beurteilungsdimensionen recht einig sind, in einer vierten aber nicht. Durch dieses unmittelbare Feedback wird nun ein Optimierungsprozess ausgelöst. Die Assessoren tauschen sich zeitnah über ihr Verständnis der betroffenen Dimension und des Beurteilungsmaßstabes aus und legen zum Beispiel gemeinsam neue Operationalisierungen fest. Ohne das Feedback über die Beurteilerübereinstimmung wäre der Optimierungsbereich hingegen überhaupt nicht identifiziert worden.

Die Bestimmung der Beurteilerübereinstimmung erfolgt PC-gestützt. Der Einsatz des Computers im AC sollte heutzutage eine Selbstverständlichkeit sein, was nicht bedeutet, dass die Assessoren bei der Beobachtung mit Rechnern arbeiten müssen. Die zunächst auf Papier fixierten Werte müssen aber schnellstmöglich (z. B. von einer Assistentenkraft) auf den Rechner übertragen werden und zumindest zum Ende des AC für Entscheidungs-, Rückmelde- und Dokumentationszwecke zur Verfügung stehen. Dabei spielt es keine Rolle, ob man sich selbst ein kleines Programm (z. B. auf Excel-Basis) erstellt oder vorhandene Programme² nutzt. Wichtig ist, dass das selbst entwickelte oder eingekaufte EDV-Programm Protokolle für die Dokumentation erstellt, die Auswertungsprozesse unterstützt und gegebenenfalls den Assessoren ein umfassendes Feedback³ zu ihren Beurteilungen liefert. Es kommt immer noch vor, dass Personen jahrelang als Assessoren arbeiten, ohne jemals ein Feedback über ihre Arbeit erhalten zu haben.

Ein Feedback für die Assessoren könnte zum Beispiel umfassen:

- ein vergleichendes Urteil über die Dimensionen und Übungen eines Assessors (z. B. „Im Vergleich aller Beurteilungen, die Frau Mustermann abgegeben hat, hat sie die Dimension Teamfähigkeit am strengsten beurteilt.“),
- ein Vergleich mit den anderen Assessoren über die Dimensionen und Übungen (z. B. „Im Vergleich mit den anderen Assessoren hat Frau Mustermann die Dimension Motivation am mildesten beurteilt.“),
- eine absolute Betrachtung der Dimensionen und Übungen (z. B. „Über alle Assessoren hinweg wurde die Dimension Kommunikationsfähigkeit im Durchschnitt über alle Teilnehmer als besonders kritisch betrachtet.“),
- den Wert für die Beurteilungsübereinstimmung der Assessoren.

Wertet man die Ergebnisse einer ganzen Serie von AC aus, kann man darüber hinaus unter anderem die Schwierigkeit und Differenzierungsfähigkeit der einzelnen Übungen und Anforderungsdimensionen bestimmen.

Eine andere (kombinierbare) Möglichkeit, die Qualität von AC zu kontrollieren, schildert Greco (2007). Er schlägt vor, die Assessoren bei jedem Urteil zusätzlich angeben zu lassen, wie sicher oder unsicher sie sich in ihrem Urteil sind. Auch wenn es keinen direkten Zusammenhang zwischen der Beurteilungsgenauigkeit und der Unsicherheit gibt, so liefert das Unsicherheitsgefühl gegebenenfalls doch Hinweise auf die mangelhafte Passung zwischen Dimension und Übung und somit auf Optimierungsmöglichkeiten.

Eine Qualitätskontrolle darf in letzter Hinsicht auch nicht davor zurückschrecken, die Eignung der Assessoren infrage zu stellen. Hinweise erhält man zum Beispiel über die Auswertung der Beobachtungsbogen. Wenn sich auch nach intensivem Training keine Notizen zu konkretem Verhalten finden lassen und / oder Verhaltensweisen den Dimensionen nachweislich falsch zugeordnet werden, muss über eine Nachschulung und letztlich auch über einen Ersatz nachgedacht werden.

Akzeptanz von AC

Die Akzeptanz von AC ist eigentlich ein Nebenaspekt der Qualität. Diesem Aspekt wird in der Praxis eine hohe Aufmerksamkeit gewidmet. Spätestens seitdem (wieder) ein Mangel an qualifiziertem Personal besteht, ist man besorgt, wie die eingesetzten Verfahren und damit auch die AC bei den internen oder externen Bewerbern ‚ankommen‘. Durch den aktuellen Fachkräftemangel gewinnt die Nachfragerseite an Verhandlungsmacht, sodass man aus strategischen Gründen die Perspektive der Bewerber stärker berücksichtigt. Das Erleben, Entscheiden und Handeln der Bewerber wird

unter dem Begriff der sozialen Validität (Schuler 1990) oder sozialen Akzeptanz zusammengefasst. Organisationen suchen Verfahren, in denen der Bewerber der Organisation gleichsam auf Augenhöhe begegnen kann. Die AC-Anwender setzen aus Sorge vor einer mangelhaften Akzeptanz keine Tests ein, und die Interviewanwender scheuen aus den gleichen Gründen vor dem AC zurück. Akzeptiert man die bloße Behauptung sowie die unsystematischen Berichte über ‚gute Erfahrungen‘ und ‚positive Rückmeldungen‘ aber nicht als Datenersatz, muss man sich fragen, worauf die jeweilige Sorge beruht, da sich kaum aussagekräftige empirische Untersuchungen zu dieser Fragestellung finden lassen.

Darüber hinaus wird die Akzeptanz durch zahlreiche Faktoren beeinflusst, die außerhalb des Verfahrens liegen, zum Beispiel die Attraktivität einer Position. Verfahrensteilnehmer erwarten, dass der Aufwand für das Verfahren in einer ‚vernünftigen‘ Relation zu der zu besetzenden Stelle steht. Für eine unattraktive Stelle ‚lohnt‘ es sich nicht, an einem mehrtägigen AC teilzunehmen, egal wie gut das AC gestaltet sein mag. Das Akzeptanzurteil variiert außerdem in Abhängigkeit von persönlichen Faktoren der Teilnehmer. So werden beispielsweise Personen, die in einem AC schlecht abschneiden oder auch nur vermuten, dass sie schlecht abgeschnitten haben, ihr positives Selbstwertgefühl schützen, indem sie das AC abwerten. Dennoch hat die Gestaltung eines Verfahrens wie zum Beispiel eines AC, erheblichen Einfluss auf das Akzeptanzurteil, sodass eine Erhebung der Akzeptanz als ein zusätzliches Qualitätskriterium aufschlussreich sein kann.⁴

In bislang zwei Untersuchungen (Kersting, im Druck) zeigte sich, dass insbesondere die Urteile über die Messqualität und Augenscheinvalidität das Gesamturteil über die Akzeptanz bestimmen, während die Belastung zwar zum Teil kritisch gesehen wird, sich aber nicht auf das Akzeptanzgesamturteil niederschlägt.

Die sieben Akzeptanz-Dimensionen von AC sind

- Messqualität,
- Augenscheinvalidität,
- Kontrollierbarkeit,
- Belastungsfreiheit,
- Intention zum unverfälschten Verhalten,
- positive Atmosphäre,
- gute Organisation.

AC sind bewährte Verfahren und werden so häufig eingesetzt wie nie zuvor. Allerdings sind AC nicht unumstritten. AC-Anwender sollten sich unvoreingenommen fragen, welche Kritik an- und aufzunehmen ist, um sie durch die Weiterentwicklung der AC-Technik langfristig gegenstandslos zu machen. Bis dahin ist es aber ein weiter Weg. Der vorliegende Beitrag betont Möglichkeiten zur Qualitätskontrolle und -optimierung, denn auch eine weite Reise beginnt bekanntlich mit dem ersten Schritt.

Summary

Future of Assessment Centers

“Playground for Laymen’s Diagnostics, “Blabbermouths Preferred”, “Job Applicants in the Monkey Circus” – these are some of the headlines of current articles discussing assessment centers (AC) which highlight the poor accuracy of these processes. For many years, AC have been severely criticized by HR staffers. But which criticism is legitimate, and which is exaggerated? Martin Kersting investigated the current status of AC in practice, related new trends and the general validity of AC test results. The author then addresses the key question of how a high quality AC can be realized at reasonable cost and effort in individual cases, and how to verify the quality of its results.

Anmerkungen

- 1 Siehe z. B. www.arbeitskreis-ac.de; www.assessmentcenters.org/pdf/00guidelines.pdf; www.kersting-internet.de/DIN-Screen.html
- 2 Z. B. „AC und Interview Organizer (ACIO)“, siehe <http://personalwahl-internet.de>; „easy.pes“, siehe <http://ingentis.de>; „KI-PAT“, siehe <http://KI-management.de>, oder „pro facts AC Pilot“, siehe <http://profacts.de>
- 3 Solche Feedbacks liefert z. B. das Programm ACIO.
- 4 Kersting hat einen standardisierten Fragebogen zur Erfassung der Akzeptanz diagnostischer Verfahren entwickelt, der aktuell in drei Varianten vorliegt: für Intelligenz- und Leistungstests (AKZEPTI-L), für Persönlichkeitsfragebogen (AKZEPTI-P) und für AC (AKZEPTI-AC), siehe www.kersting-internet.de/Akzeptanzfragebogen.html.

Literatur

- Arbeitskreis Assessment Center (AK-AC) (2004): Standards der Assessment Center Technik, Hamburg
- Bowler, M. C. / Woehr, D. J. (2006): A meta-analytic evaluation of the impact of dimension and exercise factors on assessment center ratings, in: *Journal of Applied Psychology*, 91, 1114–1124
- Bray, D. W. / Grant, D. L. (1966): The assessment center in the measurement of potential for business management, in: *Psychological Monographs: General and Applied*, 80, 1–27
- Greco, R. (2007): Unsicherheit bei der Verhaltensbeobachtung und -bewertung im Assessment Center. Mögliche Einflussfaktoren und ihr Zusammenhang zur Beurteilungsgenauigkeit, unveröffentlichte Diplomarbeit an der Universität Bielefeld, Bielefeld
- Grove, W. M. / Zald, D. H. / Lebow, B. S. / Snitz, B. E. / Nelson, C. (2000): Clinical versus mechanical prediction: A meta-analysis, in: *Psychological Assessment*, 12 (1), 19–30
- Hardison, C. M. / Sackett, P. R. (2007): Kriteriumsbezogene Validität des Assessment Centers: Lebendig und wohl-auf?, in: H. Schuler (Hg.): *Assessment Center zur Potenzialanalyse*, Göttingen, 192–202
- Hermelin, E. / Lievens, F. / Robertson, I. T. (2007): The validity of assessment centres for the prediction of supervisory performance ratings: A meta-analysis, in: *International Journal of Selection and Assessment*, 15, 405–411
- Höft, S. / Funke, U. (2006): Simulationsorientierte Verfahren der Personalauswahl, in: H. Schuler (Hg.): *Lehrbuch der Personalpsychologie*, 2. Aufl., Göttingen, 145–187
- Höft, S. / Lüth, N. (2005): Beobachtung und Bewertung im Assessment Center: Gestaltungsmerkmale eines AC-Beobachtungssystems, in: K. Sünderhauf / S. Stumpf / S. Höft (Hg.): *Assessment Center. Von der Auftragsklärung bis zur Qualitätssicherung*, Lengerich, 164–180
- Holzenkamp, M. / Spinath, F. / Höft, S. (im Druck): Wie valide sind AC im deutschsprachigen Raum?, in: *Arbeitskreis Assessment Center e.V. (Hg.): Diagnostische Kompetenz: Entwickeln und Anwenden. Dokumentation zum 7. Deutschen Assessment-Center-Kongress 2008*, Lengerich
- Hossiep, R. (2001): Psychologische Tests – die vernachlässigte Dimension in Assessment Centern, in: W. Sarges (Hg.): *Weiterentwicklungen der Assessment Center Methode*, 2. Aufl., Göttingen, 53–68
- Kersting, M. (2008a): Qualität in der Diagnostik und Personalauswahl: Der DIN-Ansatz, Göttingen
- Kersting, M. (2008b): Zur Akzeptanz von Intelligenz- und Leistungstests, in: *Report Psychologie*, 33, 420–433
- Kersting, M. (im Druck): Akzeptanz von Assessment Centern: Was kommt an und worauf kommt es an?, in: *Arbeitskreis Assessment Center e.V. (Hg.): Diagnostische Kompetenz: Entwickeln und Anwenden. Dokumentation zum 7. Deutschen Assessment-Center-Kongress 2008*, Lengerich
- McDaniel, M. A. / Whetzel, D. L. / Schmidt, F. L. / Maurer, S. D. (1994): The validity of employment interviews: A comprehensive review and meta-analysis, in: *Journal of Applied Psychology*, 79, 599–616
- Obermann, C. / Höft, S. (2008): *Assessment-Center: Entwicklung statt Auswahl*, in: *Wirtschaftspsychologie aktuell*, 15, 18–20
- Ones, D. S. / Viswesvaran, C. / Reiss, A. D. (1996): Role of social desirability in personality testing for personnel selection, in: *Journal of Applied Psychology*, 81, 660–679
- Salgado, J. F. / Anderson, N. / Mosco, S. / Bertua, C. / De Fruyt, F. (2003): International validity generalization of GMA and cognitive abilities: A European community meta-analysis, in: *Personnel Psychology*, 56, 573–605
- Schuler, H. (1990): Personalauswahl aus der Sicht der Bewerber: Zum Erleben eignungsdiagnostischer Situationen, in: *Zeitschrift für Arbeits- und Organisationspsychologie*, 34, 184–191
- Schuler, H. (2007): Spielwiese für Laien? Warum das Assessment-Center seinem Ruf nicht mehr gerecht wird, in: *Wirtschaftspsychologie aktuell*, 14, 27–30
- Task Force on AC Guidelines (2000): *Guidelines and ethical considerations for assessment center operations*, Bridgeville, PA
- Thornton III, G. C. / Gaugler, B. B. / Rosenthal, D. B. / Bentson, C. (1987): Die prädiktive Validität des Assessment Centers – eine Metaanalyse, in: H. Schuler / W. Stehle (Hg.): *Assessment Center als Methode der Personalentwicklung*, Göttingen, 36–60