Qualität und Validität von Assessment-Centers zur Personalauswahl: Show oder echte Prognose von Berufserfolg?

Marcus Kuhnhardt

„Ohne Güteprüfung und Qualitätskontrolle wird ein Assessment-Center zu einem sinnlosen Ritual.“

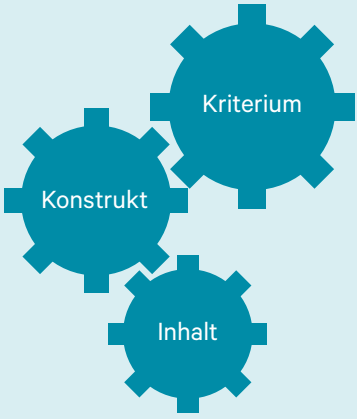
Wolf, B., Barell, G. & Hoenle, S.; S.5

Seit seiner Gründung unterstützt die dgp den öffentlichen Sektor und die Privatwirtschaft bei der Personalauswahl und wirbt dabei mit besonders hohen wissenschaftlichen Standards, also mit hoher Objektivität, Reliabilität und vor allem mit tatsächlicher Gültigkeit der Ergebnisse (Validität). Schließlich soll das Assessment-Center (AC) nicht einfach nur eine Auswahl erlauben, um die Bewerberzahl zu reduzieren, sondern bei dieser Auswahl auch tatsächlich jene Bewerber*innen selektieren, die später im Beruf auch die besseren Leistungen bringen. Die Problematik dabei: Die Überprüfung dieser Erfolgsprognose ist schwierig, ressourcenintensiv und kann erst mit enormer Zeitverzögerung nach einem jeweiligen Verfahren erfolgen. Daher sucht man bei vielen Anbietern vergeblich nach tatsächlich belastbaren Beweisen für die Aussagekraft der eigenen Auswahlverfahren (Schuler, 2014). Gleichzeitig erfahren ACs weiterhin eine stetig steigende Beliebtheit, während jedoch die Qualität nicht in gleichem Maße steigt, sondern einem deutlichen Abwärtstrend unterliegt (Fesefeldt, 2015). Dabei nehmen die qualitativen Unterschiede zwischen verschiedenen ACs stetig zu (Diagnostik- und Testkuratorium, 2018).

Es wird deutlich, dass eine konsequente und regelmäßige Überprüfung der Qualität der eigenen Verfahren (Evaluation) ein zwingend notwendiger Schritt ist. Die dgp stellt sich seit jeher dieser Selbstprüfung und veröffentlicht unter anderem jedes Jahr in den dgp-Informationen mindestens einen Artikel zur Evaluation von Teilen oder ganzen Verfahren (z.B.: Jobmann, 2017; Stadelmaier et al., 2018; Kleinmanns, Jobmann, 2019). In diesem Jahr erfolgte die Evaluation einer ganzen Verfahrensreihe von ACs zur Potentialanalyse für Aufstiegs- und Qualifikationsverfahren. Die Ergebnisse sollen hier sehr komprimiert zusammengefasst werden. Für die ausführliche Studie oder Nachfragen steht die unten aufgeführte Korrespondenz gern zur Verfügung.

Was macht Qualität von Assessment-Centers aus?

Investitionsentscheidungen werden oft nach dem Preis-Leistungs-Verhältnis getroffen. So müssen auch ACs beweisen, dass sie ihre hohen Kosten (im Vergleich zu Testverfahren wie Kognitive Leistungstests) auch mit ähnlich hohem Erkenntnisgewinn rechtfertigen. Bei ACs ist diese Einschätzung nicht so einfach, da eine schier unendliche Zahl an Faktoren in den Prozess einfließt. Dennoch hat sich in der psychologischen Forschung eine 3-Teilung der Validierung nach Inhalt, Konstrukt und Prognose (Kriterium) als sinnvoll etabliert. Wie Zahnräder greifen die einzelnen Formen ineinander und bedingen sich. Ist beispielsweise die Konstruktvalidität gering, leidet auch die Gesamtprognose.



Beim Inhalt stellt sich die Frage, ob das AC einen Realitätsbezug aufweist, also ob die Aufgaben und Stationen im AC zumindest teilweise etwas simulieren, das im späteren Arbeitsalltag relevant ist. Bei den Kon-

strukten dreht sich alles um die eigentliche Diagnostik im AC, also die Problematik, ob die festgelegten Anforderungen, z.B. Motivation oder Soziale Kompetenz, im AC auch tatsächlich erfasst werden oder ob lediglich ein Gesamturteil gefällt wird. Ein beliebter Einwand hierbei ist, dass festgelegte Dimensionen und Anforderungen irrelevant, fernab der Realität und überhaupt viel zu starr seien. Eine klassische Aussage in der Diskussion in Kommissionen wäre „Der Bewerber ist einfach gut!“. Aber was heißt „gut“? Es gibt zahlreiche Artikel und Forschungen dazu, dass vor allem extrovertierte Persönlichkeiten in ACs gut abschneiden, weil sie ein selbstbewusstes und lebhaftes Gesamterscheinungsbild zeigen (Cain, 2011). Allerdings ist auch hinreichend bewiesen, dass dies nicht einhergeht mit tatsächlicher Leistungsfähigkeit (Klimesch, 2009). Schein ist eben nicht Sein. Daher ist der Fokus auf die einzelnen, tatsächlichen Teilleistungen von Bewerber*innen so wichtig, da die Dimensionen so der Vielzahl an Beobachtungsfehlern wie Halo- und Primacy-Effekt, selbsterfüllende Prophezeiung und dergleichen entgegenwirken können. Für die Ermittlung der wichtigsten Leistungskriterien ist nach wie vor die umfassende Anforderungsanalyse alternativlos. Die Konstruktvalidität stellt zusammenfassend also ein Maß für die Methodik dar und liefert daher oft einen sehr guten Anhaltspunkt für Verbesserungsmöglichkeiten. Sie stellt auch seit jeher das Sorgenkind der Qualität von ACs dar und wird in der Forschung meist sehr kritisch bewertet (ausführliche Betrachtung in der Meta-Analyse von Bowler & Woehr, 2006).

Das eigentliche Kernanliegen von ACs ist die prognostische Validität, also die Vorhersagekraft von ACs, potentiell leistungsfähige Bewerber*innen aus der Vielzahl von Mitbewerber*innen auszuwählen. Da das AC eine Prognose für den späteren Berufserfolg liefern soll, muss dieser ermittelt und auf Zusammenhänge mit dem AC untersucht werden. In der hier untersuchten Verfahrensreihe diente als Vergleichskriterium die Studienleistung der Bewerber*innen.

Was wurde evaluiert?

Die untersuchten Verfahren wurden im Zeitraum von 2012 bis 2019 durchgeführt und bestanden aus den Stationen der Gruppendiskussion, Präsentation und Strukturiertes Interview. Ohne Beteiligung der dgp wurde zur Vorselektion ein bekannter Intelligenztest und ein vom Kunden entworfener Wissenstest durchgeführt. Jene Bewerber*innen, die das AC erfolgreich bestanden und im Ranking auf vorderen Plätzen landeten, erhielten einen Studienplatz, der nach erfolgreichem Bestehen für den Aufstieg zur oder die Weiterqualifikation innerhalb der Laufbahngruppe 2 (gehobener Dienst) befähigt. Für die Evaluation wurden die Ergebnisse im AC mit den Prüfungsnoten im Studium verglichen sowie innerhalb der ACs die Umsetzung der Bewertungsdimensionen ausgewertet.

Ergebnisse

Die erfreuchliche Nachricht vorweg – die über viele Jahre durchgeführten ACs der dgp können eine gute Vorhersage für den späteren Studienerfolg liefern (siehe Tabelle 1). Bewerber*innen, die im AC bessere Leistungen gezeigt haben, erzielen tendenziell im anschließenden Studium ebenfalls bessere Noten.

	Studien-leistung	AC	Wissens-test	Intelli-genztest
Studien-leistung	1.00	.728	.462	.599
AC	.28**	1.00		
Wissens-test	.03	-.07	1.00	
Intelligenz-test	-.04	-.21**	.28**	1.00

Anmerkung. ** = Korrelationen sind bei α ≤ .01 signifikant (zweiseitig)

Tabelle Korrelationen zwischen Verfahrens-Teilergebnissen und Studienleistung

Zusatzbemerkung: Die Korrelation zwischen AC und Studienleistung wurde um Varianzeinschränkungen durch Vorselektion korrigiert. Nach Korrektur der Varianzeinschränkung berechnet sich eine korrigierte Korrelation von T = .42 und ist somit als gut bis sehr gut einzustufen (Schuler, 2014).

Überraschend ist, dass der Intelligenztest kaum Vorhersagekraft zu haben scheint, denn die Prognose-Gültigkeit von Intelligenztests auf späteren Studienerfolg ist so gut erwiesen wie kaum ein anderes Konstrukt der Psychologie (Diagnostik- und Testkuratorium, 2018). Möglicherweise liegt dies daran, dass der externe Test (nicht von der dgp) teilveröffentlicht im Internet zu finden ist und daher durch Auswendiglernen seiner Grundlage beraubt ist. In den letzten Jahrgängen wurde der Intelligenztest ausgetauscht und tatsächlich zeigte sich anschließend eine bessere Gültigkeit, die jedoch weiterhin hinter den Erwartungen zurückblieb. Eventuell liegt dies daran, dass der neue Intelligenztest der gleichen Intelligenztheorie angehört wie der vorangegangene, entsprechend viele Aufgabentypen teilt und sich der Lerneffekt dadurch übertrug. Auch der Wissenstest, der nicht von der dgp konzipiert wurde, kann keine Vorhersage treffen. Es wird also deutlich, dass die reine Durchführung von verschiedenen Verfahren keinen Mehrgewinn an Prognosegüte liefert, sofern die Teilverfahren nicht hohen Standards gerecht werden. Dies zeigt sich auch noch mal nachdrücklich in Tabelle 2, die mit dem b-Gewicht ein Maß dafür gibt, wie groß der Einfluss der Teilergebnisse auf den späteren Studienerfolg ist.

	b	b standardisiert	a
AC	1.44	0.61	.00
Intelligenztest	0.28	0.55	.01
Wissenstest	0.21	0.44	.04

Anmerkung. Abhängige Variable ist Studienleistung

Tabelle 2 Multiple lineare Regression mit Einschluss aller Prädiktoren

Gute Prognose mit Luft nach oben

... aber es gibt auch Verbesserungspotential, und dies vor allem bei der Verfahrensumsetzung. Wie sich aus Abbildung 1 und Tabelle 3 zeigt, wurde in den Verfahren vor allem ein Gesamturteil gebildet und die einzelnen festgelegten Dimensionen, die sich aus den Anforderungen ergeben sollten, kaum getrennt bewertet.

Dies hat mehrere Gründe:

1. Es wurden schlicht zu viele Dimensionen in das Verfahren gepackt. Die DIN 33430 gibt hierfür mit 3, maximal 5 Dimensionen pro Aufgabe klare Empfehlungen, die in den Verfahren nicht umgesetzt wurden. Bei zu vielen Dimensionen kann das tatsächlich beobachtbare Verhalten in der Aufgabe gar nicht genug Stoff für differenzierte Bewertungen liefern und somit ist die beurteilende Person gezwungen, ein Gesamturteil vage auf die einzelnen Dimensionen aufzuteilen. Das widerspricht aber dem eigentlichen Ziel. Hierbei braucht es mehr Mut, verwandte Dimensionen zusammenzufassen und sich auf die wirklich wichtigen Kernanforderungen zu konzentrieren. Die Tendenz, möglichst viele verschiedene Kompetenzen erfassen zu wollen, ist verständlich. Wie sich jedoch in dieser Studie, aber auch in der breiten Forschungsbasis zeigt (Bowler, 2006), erzeugt dieser gutgemeinte Wille schnell genau das Gegenteil, nämlich, dass bei zu vielen Dimension auf einmal effektiv gar keine einzelnen Kompetenzen mehr gemessen werden. Bei den Dimensionen im AC gilt also nachdrücklich das bekannte Motto: Weniger ist oft mehr. Dafür ist eine regelmäßige, ausführliche Anforderungsanalyse nötig, die in diesen Verfahren aufgrund des Wiederholungscharakters nicht mehr erfolgte.

2. Die Beurteilung in ACs ist ein komplexer und fordernder Prozess, der mit entsprechendem Training und Schulungen geübt werden sollte. Auch hat jedes AC seine eigenen Schwerpunkte und Dimensionen, die in einem Training ausreichend erläutert werden müssen. Dieses Training wird in der Praxis jedoch oft sehr reduziert gehalten oder gar nicht erst in Auftrag gegeben, da es einen zusätzlichen Zeit- und Kostenfaktor darstellt. Die Kommissionen dieser Verfahrensreihe erhielten durch den Kunden umfassende Schulungen zur Bewertung und Durchführung von ACs, was als sehr positiv zu werten ist. Dennoch sollten an diesen Schulungen, wenn möglich, auch die jeweils verantwortlichen Eignungsdiagnostiker*innen anwesend sein (DIN 33430), um gezielt die Charakteristika der Bewertungsdimensionen und ihre Umsetzung in den AC-Stationen vermitteln zu können. Durch die allgemeinen Schulungen dürfte die Diagnostik im Gesamtbild hochwertig, jedoch die Beurteilung der verwendeten Dimensionen teilweise eingeschränkt sein.

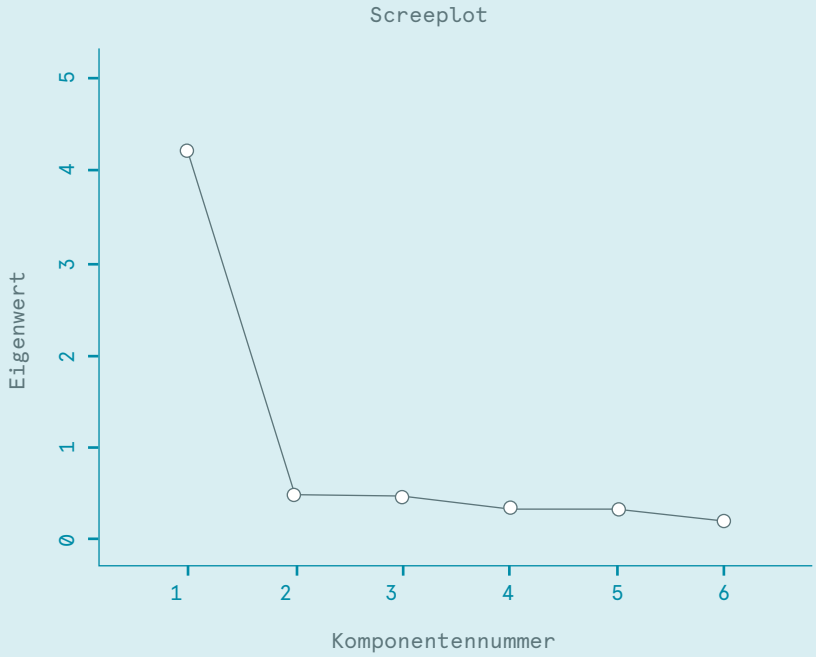


Abbildung 1. Screeplot Faktorenanalyse der Dimensionen.

Statistisch gesprochen:

Die explorative Faktorenanalyse extrahiert eine Hauptkomponente, auf die alle sechs Dimensionen sehr hoch laden. Die Kommunalitäten liegen im Bereich von .61 für ,,Engagement“ und .79 für ,,Kommunikationskompetenz“. Die ermittelte Hauptkomponente erklärt 70% der Gesamtvarianz.

Bedeutung:

Die einzelnen Bewertungen der Dimensionen sind so sehr voneinander abhängig, dass mit hoher Wahrscheinlichkeit ein Gesamturteil gebildet wurde, welches 70% der Bewertung ausmacht. Dies wird bestätigt durch Tabelle 3, die zeigt, dass die verschiedenen Dimensionen innerhalb einer Aufgabe mehr zusammenhängen als die Bewertungen einer Dimension über die Aufgaben hinweg. Im diagnostisch „korrekten“ Fall müsste es genau andersherum sein.

Anmerkung: Die Ergebnisse beziehen sich auf eine mtmm-Matrix der Ergebnisse von zwei Verfahren von 2018 und 2019, bei denen bereits die Dimensionen reduziert wurden. In älteren Verfahren kamen bis zu 8 Dimensionen zum Einsatz.

Gleiche Dimensionen in verschiedenen Aufgaben		Unterschiedliche Dimensionen in der gleichen Aufgabe	
Analytisch-logische Kompetenz	t Ø = .43	Gruppendiskussion	t Ø = .54
Soziale Kompetenz	t Ø = .31	Interview	t Ø = .62
Kommunikations-kompetenz	t Ø = .41	Präsentation	t Ø = .55
Emotionale Stabilität	t Ø = .39		
Gesamtdurchschnitt	t Ø = .38		t Ø = .57

Anmerkung. Grundlage für die Werte der Tabelle ist eine ausführliche multitraid-multimethod-Matrix (MTMM).

Tabelle 3 Gesamtergebnis einer Aufgabe versus Bewertung der einzelnen Dimensionen

Zusammenfassende Erkenntnisse aus den Ergebnissen (Auswahl)

1. In der aktuellen Zusammenstellung des Verfahrens ist das AC für eine zielführende Auswahl unverzichtbar. Es liefert eine gute bis sehr gute Prognose für späteren Studienerfolg.

2. Bei der Konzipierung von ACs ist eine umfassende Anforderungsanalyse und die Erstellung passender Dimensionen die wichtigste Basis für durchgehend hohe Qualität. Bei der Zahl der Dimensionen gilt: So wenig wie möglich, so viel wie nötig.

3. Selbst ein perfekt konzipiertes Verfahren kann nur vollständig funktionieren, wenn alle Beurteiler*innen ausreichend verfahrensspezifisch geschult und trainiert werden und dies immer wieder aufgefrischt wird.

4. Intelligenztests sind nur nützlich, wenn sie den Qualitätsstandards genügen und nicht teilveröffentlicht „auswendig lernbar“ sind. Eine Bedingung, welche die dgp mit ihren konsequenten, jährlichen Überprüfungen und Überarbeitungen der hauseigenen Tests erfüllen kann.

Mit der ausführlichen Evaluation konnte die dgp zeigen, dass eine von ihr durchgeführte AC-Reihe einen unverzichtbaren Anteil zur Auswahl von Potentialen liefert. Jedoch wird auch deutlich, dass neben der reinen Durchführung vor allem die Vorbereitung durch umfangreiche Anforderungsanalyse und Beobachterschulungen von sehr hoher Bedeutung sind und bei vergleichbaren Verfahren entsprechend eingeplant werden sollten.

Quellen

Arbeitskreis Assessment Center e.V. (2016). AC-Standards. Buxtehude: Arbeitskreis Assessment Center e.V.

Bowler, M. C., & Woehr, D. J. (2006). A Meta-Analytic Evaluation of the Impact of Dimension and Exercise Factors on Assessment Center Ratings. Journal of Applied Psychology 2006, Vol. 91, No. 5, S. 1114–1124.

Cain, S. (2011). Quiet. New York: The Crown Publishing Group.

Diagnostik- und Testkuratorium. (2018). Personalauswahl kompetent gestalten. Gießen: Springer-Verlag GmbH.

DIN Deutsches Institut für Normung e. V. (Juli 2016). DIN 33430. Anforderungen an berufsbezogene Eignungsdiagnostik. Berlin: Beuth Verlag GmbH.

Fesefeldt, J. (2015). Qualität und Validität des Assessment-Centers. dgp Informationen 2015, S. 38-49.

Hornke, L., & Winterfeld, U. (2004). Eignungsbeurteilung auf dem Prüfstand: DIN 33430 zur Qualitätssicherung. München: Elsevier GmbH.

International Taskforce on Assessment Center Guidelines. (2015). Guidelines and Ethical Considerations for Assessment Center Operations. Journal of Management, Vol. 41 No. 4, May 2015, 1244–1273.

Kanning, U. P., Pöttker, J., & Gelléri, P. (2007). Assessment Center-Praxis in deutschen Großunternehmen. Zeitschrift für Arbeits- u. Organisationspsychologie 51, S. 155-167.

Klimesch, S. (2009). Kompetenz, Persönlichkeit und Berufserfolg. Wuppertal: Bergische Universität Wuppertal.

Sackett, P., & Harris, M. (1988). A further Examination of the constructs underlying assessment center ratings. Journal of Business and Psychology Vol 3, Nr. 2, S. 214-236.

Schuler, H. (2014). Psychologische Personalauswahl. Göttingen: Hogrefe Verlag GmbH & Co. KG.

Wolf, B., Barell, G. & Hoenle, S. (1995). Assessment Center auf dem Prüfstand. Hamburg: Windmühle GmbH.

Kontakt:

M. Sc. Psychologie Marcus Kuhnhardt
kuhnhardt@dgp.de

Deutsche Gesellschaft für Personalwesen e. V.
Grassistraße 12, 04107 Leipzig