Konstruktvalidität im Fokus – wie sich Komponenten des Assessment Center Designs auf die Varianzaufklärung in Assessment Center Ratings auswirken

Um die Kompetenzen und Verhaltenstendenzen arbeitstätiger oder -suchender Personen einzuschätzen, stehen heute eine Vielzahl von Messinstrumenten und Verfahren zur Verfügung. Eines der beliebtesten und am häufigsten verwendeten diagnostischen Verfahren im Personalwesen ist dabei das Assessment Center (AC; Höft & Obermann, 2010; Schuler, Hell, Trapmann, Schaar & Boramir, 2007). Trotz der Beliebtheit und des häufigen Einsatzes von ACs wird seit Jahrzehnten diskutiert, ob ACs tatsächlich konstruktvalide sind: Forschungsbefunde werfen immer wieder die Frage auf, inwiefern ACs in der Lage sind, wie beabsichtigt situations- beziehungsweise übungsübergreifende und damit stabile Kompetenzen (d.h. Dimensionen) erfassen zu können oder ob die gewonnenen AC-Ratings lediglich Abbild situativer Gegebenheiten sind (Bowler & Woehr, 2009; Woehr & Arthur, 2003).

Eine Methode, um den Einfluss verschiedener varianzverursachender Komponenten von ACs und damit die Konstruktvalidität von ACs zu untersuchen, stellt dabei die Generalisierbarkeitstheorie (G-Theorie; Cronbach, Gleser, Nanda & Rajaratnam, 1972) dar. Ihr Vorteil ist, dass sie den Einfluss verschiedener varianzverursachender Komponenten von ACs (folgend Varianzkomponenten genannt) wie Dimensionen, Übungen, Assessoren und Teilnehmer identifizieren und deren Zusammenspiel systematisch untersuchen kann (Shavelson, Webb & Rowley, 1989). Empirisch zeigt sich die Problematik der Konstruktvalidität in den betreffenden Studien, indem die Dimensionen nicht wie gewünscht hauptverantwortlich für die Varianz in den AC-Ratings der Teilnehmer sind und sich die Anteile der Varianz, welche auf die Dimensionen zurückzuführen sind, erheblich voneinander unterscheiden (z.B. Bowler & Woehr, 2006; Jackson, Michaelides, Dewberry & Kim, 2016; Lance, Lambert, Gewin, Lievens & Conway, 2004). So können die Befunde bisheriger G-Theorie-Studien die Zweifel daran, ob ACs tatsächlich das messen, was sie zu messen beanspruchen, kaum entkräften.

Weshalb die Befunde bisheriger Studien derart heterogen ausfallen, bleibt bisher unklar. Verschiedene Studien weisen darauf hin, dass Komponenten des AC-Designs (folgend Designkomponenten genannt), welche sich auf die Dimensionen, Übungen, Assessoren und Teilnehmer beziehen (beispielsweise die Beobachtbarkeit der Dimensionen, die Komplexität der Übungen oder der berufliche Hintergrund der Assessoren), die Varianzaufklärung beziehungsweise die Varianz in den AC-Ratings beeinflussen und so möglicherweise die Variabilität der Befunde begünstigen (Lievens, 1998; Woehr & Arthur, 2003).

Die vorliegende Masterarbeit verfolgt deshalb zweierlei Ziele: Zum einen soll anhand mehrerer unterschiedlich gestalteter ACs – bisherige Studien analysierten jeweils nur ein einziges AC – geprüft werden, wie viel Varianz in den AC-Ratings auf die Dimensionen, Übungen, Assessoren sowie Teilnehmer zurückgeführt werden kann. Dabei bedient sich diese Studie ebenfalls der Methode der G-Theorie beziehungsweise der Mehrebenenanalyse. Zum anderen soll untersucht werden, welchen Einfluss dimensions-, übungs-, assessoren- und teilnehmerbezogene Designkomponenten auf die Varianzaufklärung in den AC-Ratings besitzen beziehungsweise welche Ausprägung der Designkomponenten eine möglichst akkurate Beurteilung der Dimensionen ermöglicht. In Bezug auf die dimensionsbezogenen Designkomponenten wird untersucht, welchen Einfluss die Anzahl der zu beurteilenden Dimensionen in einer Übung, die Breite, die Unterscheidbarkeit und die Beobachtbarkeit der Dimensionen besitzt. Bezüglich der übungsbezogenen Designkomponenten interessiert, inwiefern die Anzahl, die Komplexität, die Unterscheidbarkeit und die Strukturiertheit der Übungen Einfluss auf die Varianzaufklärung nimmt. Innerhalb der assessorenbezogenen Designkomponenten steht der Einfluss der Assessorenrotation sowie der des beruflichen Hintergrunds der Assessoren im Fokus und bezüglich der teilnehmerbezogenen Designkomponenten interessiert, inwiefern die Intelligenz der Teilnehmer sowie deren Fähigkeit, Anforderungskriterien in Beurteilungssituationen zu erkennen – die *ability to identify criteria* (ATIC) –, die Varianzaufklärung beeinflusst.

Methode

Die Grundlage für die Studie stellten die Daten von ACs von insgesamt zehn deutschsprachigen Organisationen dar, welche ACs für externe Kunden, den Eigenbedarf oder zu Forschungszwecken durchführen. Fünf Organisationen stellten dabei eine Stichprobe zur Verfügung, fünf Organisationen zwei Stichproben. Die einem ähnlichen oder identischen Kompetenzmodell zugehörigen Daten wurden dabei grundsätzlich als eine Stichprobe definiert. Insgesamt konnten die Daten von 15 Stichproben (N = 2 000) ausgewertet werden.

Der Ablauf der Studie gliederte sich in drei übergeordnete Schritte. In einem ersten Schritt wurden die zur Verfügung gestellten Daten der Organisationen digital erfasst und in einheitlicher Form aufbereitet. Es wurde festgehalten, welcher Teilnehmer in welcher Dimension und in welcher Übung von welchen Assessoren welches AC-Rating erhielt, wie die Teilnehmer in einem Intelligenztest abschnitten und welche Ausprägung der ATIC die Teilnehmer besaßen. Diese Datenaufbereitung stellte die Grundlage dar, um die Mehrebenenanalysen durchzuführen.

In einem zweiten Schritt wurden die Designkomponenten anhand eines Kodierschemas sowie Fragebogens vor Ort bei den Organisationen erfasst. Zur Einschätzung sichteten und verglichen die Rater systematisch relevante Unterlagen der ACs und befragten Verantwortliche des jeweiligen ACs. Jede Stichprobe und jede Dimension beziehungsweise Übung wurde separat von den Ratern hinsichtlich der Ausprägung der jeweiligen Designkomponente eingeschätzt (z.B. jede Dimension hinsichtlich ihrer Breite). Die Einschätzung anhand des Kodierschemas erfolgte dabei immer aus Basis einer Frage (z.B. *wie breit ist die jeweilige Dimension?*), welche je nach Frage offen (z.B. durch Eintragen einer Zahl), anhand einer fünfstufigen Ratingskala (z.B. von *schmal* bis *breit*) oder durch das Ankreuzen bereits vorgegebener Antwortalternativen (z.B. *ja* oder *nein*) beantwortet werden konnte. Waren die Designkomponenten anhand einer fünfstufigen Ratingskala einzuschätzen, standen den Ratern jeweils Beispiele für eine geringe, mittlere sowie hohe Ausprägung der jeweiligen Designkomponente zur Verfügung.

In einem dritten Schritt wurden für jede Stichprobe Mehrebenenanalysen durchgeführt. Für jede Stichprobe wurde berechnet, wie viel Prozent der gesamten Varianz auf die Dimensionen, Übungen, Assessoren und Teilnehmer sowie deren Zusammenspiel zurückgeführt werden kann und wie sich die Varianzaufklärung bei unterschiedlicher Ausprägung der Designkomponenten verhält beziehungsweise ändert.

Ergebnisse und Diskussion

Ein zentraler Befund der vorliegenden Studie bezüglich der Varianzaufklärung ist, dass die Anteile der Varianz, die auf die Dimensionen, Übungen, Assessoren, Teilnehmer sowie die Interaktion dieser Varianzkomponenten zurückgeführt werden können, zwischen den ACs deutlich variieren. In Bezug auf bisherige Studien finden sich damit sowohl bestätigende als auch widersprüchliche Befunde, abhängig davon, welches der insgesamt 15 ACs im Fokus steht. So variieren die Varianzanteile in der vorliegenden Studie von AC zu AC im Extremfall um über 50% – obwohl sie sich auf dieselbe Varianzkomponente beziehen. Die Studie zeigt damit, dass die Konstruktvalidität der ACs weitaus facettenreicher ausfällt, als es bisherige G-Theorie-Studien durch die Analyse von nur einem AC zeigen konnten und vor allem eines ist: schwierig, pauschal zu beurteilen. Basierend auf den 15 Stichproben findet die Studie dennoch einige saliente Muster in der Varianzaufklärung, welche sich in den *meisten* ACs der Studie zeigen beziehungsweise besonders relevant sind, um Antworten auf die Frage zu generieren, wie viel Varianz in den AC-Ratings auf die Dimensionen, Übungen, Assessoren und Teilnehmer zurückgeführt werden kann. Die Ergebnisse zeigen, dass vor allem die Übungen sowie die Teilnehmer, aber auch die Dimensionen entscheidende Anteile

der Varianz in den AC-Ratings erklären können. Die Ergebnisse weisen darauf hin, dass Dimensionen eine zentralere Rolle für das Funktionieren von ACs einnehmen, (d.h., einen größeren Varianzanteil erklären) als in bisherigen G-Theorie-Studien angenommen. Bezüglich des Teilnehmers zeigt sich, dass in den meisten ACs zu einem gewissen Ausmaß ein genereller Leistungsfaktor zum Tragen kommt (Lance et al., 2004). So scheinen einige Teilnehmer gänzlich unabhängig der Dimensionen, Übungen und Assessoren besser beurteilt zu werden als andere (siehe auch Putka & Hoffman, 2013). Dieser Effekt wird verstärkt, wenn die Teilnehmer besonders intelligent sind beziehungsweise eine hohe ATIC aufweisen.

Bezüglich des AC-Designs zeigt sich, dass sich die ACs erheblich in ihrer Gestaltung unterscheiden und kaum miteinander zu vergleichen sind. Die verschiedenen Designkomponenten zeigen dabei auf AC-übergreifender Ebene einen Einfluss auf die Varianzaufklärung. So können die Dimensionen beispielsweise akkurater beurteilt werden, wenn die Anzahl der zu beurteilenden Dimensionen in einer Übung nicht zu hoch ausfällt, die Dimensionen möglichst breit – also viele verschiedene Verhaltensweisen umfassen – und gut voneinander zu unterscheiden sind. Die AC-Ratings werden zudem akkurater, wenn eine gewisse Strukturiertheit der Übungen gegeben ist – diese also in einem gewissen Masse standardisiert sind –, wenn die Übungen gut voneinander unterscheidbar sind und wenn der soziale Interaktionsgrad in Übungen zwar gegeben, aber nicht in jeder Übung gleich hoch ist. In Bezug auf die Assessorenrotation empfiehlt es sich gemäß den Befunden der Studie, auf eine solche zu verzichten und dieselben Assessoren die Teilnehmer beurteilen zu lassen. Daran anknüpfend zeigt sich, dass sich ein alleiniger Einsatz von Psychologen als Assessoren lohnen kann und die Assessorenteams nicht zu sehr mit verschiedenen beruflichen Hintergründen durchmischt sein sollten. Unklare beziehungsweise keine einheitlichen Einflüsse zeigten sich zur Beobachtbarkeit der Dimensionen, der Anzahl der Übungen und zur Komplexität der Übungen. Es gilt, diese Einflüsse in weiteren Studien zu untersuchen. Insbesondere das Zusammenspiel verschiedener Komponenten des AC-Designs sollte zudem vermehrt in den Fokus gerückt werden.

Generell weist diese Studie darauf hin, dass bei der Konstruktion von ACs sorgfältig darauf geachtet werden sollte, welche Dimensionen beurteilt, welche Übungen eingesetzt und auf welche Art und Weise die AC-Ratings erstellt werden. So kann die Qualität von ACs verbessert und deren Stellung als eines der beliebtesten diagnostischen Verfahren weiter ausgebaut werden.

Literaturverzeichnis

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, 91,* 1114–1124. doi:10.1037/0021-9010.91.5.1114

Bowler, M. C. & Woehr, D. J. (2009). Assessment center construct-related validity: Stepping beyond the MTMM matrix. *Journal of Vocational Behavior, 75,* 173–182. doi:10.1016/j.jvb.2009.03.008

Cronbach, L. J., Gleser, G. C., Nanda, H. & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York, NY: Wiley.

Höft, S. & Obermann, C. (2010). Der Praxiseinsatz von Assessment Centern im deutschsprachigen Raum: Eine zeitliche Verlaufsanalyse basierend auf den Anwenderbefragungen des Arbeitskreises Assessment Center e.V. von 2001 und 2008. *Wirtschaftspsychologie, 12*(2), 5–16.

Jackson, D. J. R., Michaelides, G., Dewberry, C. & Kim, Y.-J. (2016). Everything that you have ever been told about assessment center ratings is confounded. *Journal of Applied Psychology, 101,* 976–994. doi:10.1037/apl0000102

Lance, C. E. (2008). Why assessment centers do not work the way they are supposed to. *Industrial and Organizational Psychology, 1,* 87–100. doi:10.1111/j.1754-9434.2007.00017.x

Lance, C. E., Lambert, T. A., Gewin, A. G., Lievens, F. & Conway, J. M. (2004). Revised estimates of dimension and exercise variance components in assessment center postexercise dimension ratings. *Journal of Applied Psychology, 89,* 377–385. doi:10.1037/0021-9010.89.2.377

Lievens, F. (1998). Factors which improve the construct validity of assessment centers: A review. *International Journal of Selection and Assessment, 6,* 141–152. doi:10.1111/1468-2389.00085

Putka, D. J. & Hoffman, B. J. (2013). Clarifying the contribution of assessee-, dimension-, exercise-, and assessor-related effects to reliable and unreliable variance in assessment center ratings. *Journal of Applied Psychology, 98,* 114–133. doi:10.1037/a0030887

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,* 60–70. doi:10.1026/1617-6391.6.2.60

Shavelson, R. J., Webb, N. M. & Rowley, G. L. (1989). Generalizability theory. *American Psychologist, 44,* 922–932. doi:10.1037/0003-066X.44.6.922

Woehr, D. J. & Arthur, W., Jr. (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. doi:10.1177/014920630302900206