

Zur zweifelhaften Validität und Nützlichkeit von Anforderungsanalysen für die Interpretation eignungsdiagnostischer Daten

Martin Kersting und Matthias Birk

1 Zur Bedeutung und Validierung von Anforderungsanalysen

Am Anfang war die Anforderungsanalyse – so könnte ein Bericht über eine gelungene Eignungsdiagnostik beginnen, denn Anforderungsanalysen sind nach Schuler (2006a, S. 181) die Voraussetzung hochwertiger Eignungsdiagnostik. „Geeignet“ ist man immer nur „für“ etwas. Das Konzept der Eignung betrifft das „Ausmaß der Übereinstimmung der Anforderungen des Arbeitsplatzes und der weiteren Arbeitsumgebung mit den Leistungsvoraussetzungen der Person“ (Schuler & Funke 1995, S. 237). Die Untersuchung, ob eine Person für eine Ausbildung, ein Studium, eine Tätigkeit, ein Aufgabenfeld, einen Arbeitsplatz, eine Position usw. geeignet ist, kann erst in Angriff genommen werden, wenn Informationen zu den jeweils gestellten Anforderungen vorliegen. Entsprechend wird der Arbeits- und Anforderungsanalyse in der DIN 33430 eine zentrale Bedeutung für die Eignungsdiagnostik zugeschrieben (DIN, 2002; Kersting, 2008). Auch andere personalpsychologische Anwendungen, wie z. B. das Personalmarketing oder die Personalentwicklung sind auf valide Informationen über die Anforderungen angewiesen (für einen Überblick über die Verwendungsmöglichkeiten von Anforderungsanalysen siehe z. B. Brannick & Levine 2002; Schuler 2006b). Letztendlich sind jegliche Versuche, Arbeitsverhalten zu verstehen und zu beeinflussen, von der Beschreibung dieses Verhaltens abhängig (Ash, 1988).

Schuler, der sich – neben zahlreichen anderen Fragestellungen – auch um das Thema Anforderungsanalyse besonders verdient gemacht hat, beschreibt in verschiedenen Publikationen (z. B. Schuler 1989, 2006b) unterschiedliche Vorgehensweisen, Anforderungen zu analysieren und schildert wertvolle Erfahrungen aus eigenen Projekten (z. B. Schuler, Funke, Moser & Donat, 1995). Er betont die Notwendigkeit der Evaluation (z. B. Schuler, 2006b), da die Ergebnisse von Anforderungsanalysen zumeist auf subjektiven Daten gründen und folglich einer Überprüfung bedürfen. Morgeson und Campion (2000, S. 819) formulieren: “Assessing the accuracy or quality of job analysis data is critical because it forms the foundation upon which virtually every human resource system is built (e. g. selection systems, training programs, performance management systems)”.

De facto liegen aber nur wenige Studien zur Qualität von Anforderungsanalysen vor, in der Regel begnügt man sich mit Reliabilitätschätzungen in Form von Be-

urteilerübereinstimmungen¹ (Moser, Donat, Schuler und Funke, 1989), die dann gelegentlich als Inhaltsvalidität interpretiert werden. Während Harvey (1991) die Interrater-Reliabilität als das bedeutsamste Kriterium für die Beurteilung der Datenqualität ansieht, stellen Sanchez und Levine (2000) den Nutzen von Reliabilitätsuntersuchungen für die praktische Umsetzung von Arbeits- und Anforderungsanalysen in Frage und plädieren für mehr Untersuchungen bezüglich der Validität von Arbeits- und Anforderungsanalysen. Daten zur Reliabilität von Anforderungsanalysen liegen in einem Umfang vor, dass sie mittlerweile metaanalytisch zusammengefasst werden konnten (Dierdorff & Wilson, 2003; Voskuijl & von Sliedrecht, 2002). Demgegenüber stehen Metaanalysen zur Validität von Anforderungsanalysen aktuell mangels einer ausreichenden Zahl von Primärstudien noch aus, insbesondere in jüngerer Zeit wurde zu dieser Frage kaum publiziert. Dabei stellt sich allerdings auch die Frage nach einer geeigneten Validierungsstrategie. In der Regel können lediglich indirekte Validitätsnachweise geführt werden, indem z. B. die Treffsicherheit geprüft wird, mit der Testwerte von Stelleninhabern und/oder stellenspezifische Validitätskoeffizienten auf der Basis von Anforderungsanalysen geschätzt werden können (z. B. McCormick, DeNisi & Shaw, 1979; McCormick, Jeanneret & Mecham, 1972; Sparrow, 1989). Indizien für die Validität der Anforderungsanalyse lassen sich auch aus der Tatsache ableiten, dass sich beispielsweise die Validität von Assessment Centern verbessert, wenn das Verfahren auf einer Anforderungsanalyse basiert (z. B. Bobrow & Leonards, 1997; Schippmann, Hughes & Prien, 1987).

Eine andere Form der indirekten Validitätsanalyse besteht in der Prüfung der Unabhängigkeit der anforderungsanalytischen Einschätzungen von Person- oder Situationsmerkmalen. Dabei zeigte sich beispielsweise, dass Stelleninhaber die an sie gestellten Anforderungen, zumindest beim Einsatz bestimmter anforderungsanalytischer Verfahren, übertrieben darstellen oder – umgekehrt formuliert – eine Mildetendenz aufweisen (z. B. Gibson, Harvey & Harris, 2007; Smith & Hakel, 1979). In den Kontext dieser Studien gehören auch Untersuchungen, die im Rahmen der Anforderungsanalyse nicht nur nach den tatsächlich vorkommenden Tätigkeiten der Stelleninhaber, sondern mit wenigen Items auch nach sogenannten Scheintätigkeiten (bogus tasks) fragen: Tätigkeiten, die von Inhabern bestimmter Stellen tatsächlich nicht ausgeübt werden. Pine (1995) stellte fest, dass je nach Art der Befragung 39 bis 50 % der Befragten kontrafaktisch angaben, mindestens eine der „bogus tasks“ auszuüben. Dies stellt die Validität derartiger Befragungen und somit bestimmter Anforderungsanalysen in Frage. Zweifel an der Validität von Arbeitsanalysen (für Anforderungsanalysen liegen bislang keine entsprechenden Studien vor) werden auch durch die Tatsache geweckt, dass in einigen Untersuchungen Laien zu weitgehend gleichen arbeitsanalytischen Ergebnissen gelangten wie Stelleninhaber und deren Vorgesetzte. Dieser Befund stellt zumindest die verbreitete Annahme in Frage, dass Stelleninhaber und Vorgesetzte über Expertenwissen verfügen und dementsprechend die bevorzugten Ansprechpartner der Arbeits- und Anforderungsanalytiker sein sollten. Smith und Hakel (1979)

¹ Um schwer verständliche Satzkonstruktionen zu vermeiden, wird im Folgenden das generische Maskulinum verwendet, gleichwohl männliche und weibliche Personen in gleicher Weise gemeint sind.

behaupten demgegenüber, dass sich Stellenexperten ebenso wie Laien in ihren anforderungsanalytischen Urteilen von Stereotypen beeinflussen lassen. Zumindest seien auch Laien recht gut dazu in der Lage, allgemeine Beurteilung von Stellen abzugeben (vgl. auch Friedman & Harvey, 1988; Jones, Main, Butler & Johnson, 1982). Die von Smith und Hakel (1979) durchgeführte Studie stützt ihre Auffassung, wurde aber kritisiert, da die Übereinstimmung zwischen Stellenexperten und Laien zu einem beträchtlichen Teil auf solche Items zurückzuführen war, die mit dem einzustufenden Arbeitsplatz nichts zu tun hatten – was auch für Laien leicht zu erkennen war (Cornelius, DeNisi & Blencoe, 1984; Friedman & Harvey, 1988; Harvey & Hayes, 1986). Die Bedeutung der Berufserfahrung für die Ergebnisse der Anforderungsanalyse ist insgesamt umstritten, während z. B. Schmitt und Cohen (1989) keine derartigen Effekte aufzeigen konnten, kommen Personen mit höherer Berufserfahrung der Studie von Borman, Dorsey und Ackerman (1992) zufolge zu anderen anforderungsanalytischen Einschätzungen als Personen mit geringerer Berufserfahrung.

2 Eine Studie zur Nützlichkeit von Anforderungsanalysen für die Interpretation eignungsdiagnostischer Daten

2.1 Theoretischer Ansatz der Studie

Schuler (2006b) unterscheidet drei verschiedene methodische Prinzipien der Anforderungsanalyse (erfahrungsgeleitet-intuitive Methode, personenbezogen-empirische Methode und arbeitsplatzanalytisch-empirische Methode) sowie drei Beschreibungsebenen von Anforderungen, nämlich die aufgabenbezogene, verhaltensbezogene und eigenschaftsbezogene Anforderungsanalyse. In der vorliegenden Studie, die ausführlich bei Birk (2004) dargestellt ist, wurden die Anforderungen der Ausbildung zum gehobenen Dienst in der allgemeinen inneren Verwaltung (siehe z. B. <http://berufenet.arbeitsagentur.de/berufe/start?dest=profession&prof-id=13915>, Zugriff: Januar 2010) auf der Eigenschaftsebene erfasst. Ein Vorteil dieser Methode besteht darin, dass die Eigenschaftsebene weniger von den beständigen Veränderungen der Arbeitswelt betroffen ist als die Aufgabenebene. Nachteilig ist allerdings der hohe Abstraktionsgrad von Eigenschaften.

Als relativ neuen Ansatz der Anforderungsanalyse nennt Schuler (2006b) unter Bezug auf Arthur, Doverspike und Barrett (1996, zitiert nach Schuler, ebd.) die Ermittlung des relativen Gewichts, welches einem Subtest in einer Testbatterie zukommt. Praktisch umgesetzt werden Aspekte dieser Idee beispielsweise in der revidierten Fassung des Wilde-Intelligenztests (WIT-2; Kersting, Althoff & Jäger, 2008). Bei der Bildung des Testgesamtwertes wird den WIT-2 Anwendern die Methode der anforderungsorientierten Integration von Verfahrensergebnissen (MAIV) zur Verfügung gestellt, mit der sie die anforderungsanalytische Bedeutung der einzelnen Dimensionen für den diagnostisch in Frage stehenden Ausbildungs- und/oder Berufserfolg durch eine entsprechende Gewichtung berücksichtigen können. Dabei wird davon ausgegangen, dass die Gewichte der Fähigkeitsdimensionen die Bedeutung dieser Fähigkeiten für die erfolgreiche Ausübung der Tätigkeit widerspiegeln. Das Verfah-

ren entspricht dem Grundprinzip der personenbezogen-empirischen Methode nach Schuler (2006b), wobei allerdings die Eigenschaften analysiert werden, die die Bewerber *vor* Berufseintritt, also zum Zeitpunkt der Personalauswahl, auszeichneten, während bei der „klassischen“ personenbezogenen-empirischen Methode die Anforderungen über die Bestimmung des statistischen Zusammenhangs zwischen Merkmalen von Stelleninhabern (also *nach* der Personalauswahl) und Berufs-/Ausbildungserfolgskriterien ermittelt werden.

Konkret leiten wir in der Studie im ersten Schritt aus den Ergebnissen einer Anforderungsanalyse die Gewichte ab, die einzelnen Subtests bzw. Testdimensionen (im Folgenden: „Dimensionen“ genannt) bei der Bildung des Gesamtwertes zukommen. In einem zweiten Schritt prüfen wir die Kriteriumsvalidität der derart gewichteten Testbatterie. Das heißt, die Ergebnisse der Anforderungsanalyse werden als diagnostische Indikatoren benutzt, deren Validität geprüft werden kann (siehe auch Tergan, 1988). Die Kriteriumsvalidität des anforderungsanalytisch konstruierten Auswahlverfahrens lässt dann Rückschlüsse auf die Validität der Anforderungsanalyse selbst zu. Die Analyse stellt, wie auch andere Ansätze (siehe oben), allerdings nur eine indirekte Prüfung der Validität der Anforderungsanalyse dar (siehe Abschnitt 4).

Die Gewichtung der Dimensionen haben wir anhand der Ergebnisse der Anforderungsanalysen vorgenommen, die wir bei drei verschiedenen Gruppen durchgeführt haben:

1. Als interne Stellenexperten wurden 121 Stelleninhaber und Vorgesetzte von Stelleninhabern befragt, dabei handelte es sich um Beamte des gehobenen und höheren Verwaltungsdiensts. Die Stelleninhaber waren im Durchschnitt 38.5 Jahre alt (Standardabweichung (SD) = 9.3) und verfügten im Mittelwert über 13.7 Jahre Erfahrung mit dem zu beurteilenden Berufsbild ($SD = 8.8$)². Der Zusammenhang der Anforderungsbeurteilungen zwischen der Gruppe der Stelleninhaber und der Gruppe der Vorgesetzten war mit $ICC_{unjust} = .938$ sehr hoch, so dass die beiden Gruppen zur Gruppe der internen Stelleninhaber zusammengefasst wurden.
2. Befragt wurden außerdem 10 externe Experten. Die im Weiteren als „Subject Matter Experts“ (SMEs) bezeichneten Personen verfügten über langjährige Erfahrung im Bereich der Personalauswahl für die Zielgruppe, es handelte sich um Diplompsychologen der Deutschen Gesellschaft für Personalwesen e. V.
3. Zusätzlich wurde der anforderungsanalytische Fragebogen auch 71 Laien vorgelegt. Als Laien-Beurteiler wurden Studierende der Psychologie befragt. Sie erhielten eine dem Berufenet (<http://berufenet.arbeitsagentur.de>) entnommene kurze Beschreibung des in Frage stehenden Berufsbildes sowie der in Frage stehenden Ausbildung. Als Manipulationscheck wurden die Teilnehmer vorab gefragt, wie bekannt ihnen das in Frage stehende Berufsbild ist. Personen, die angaben über Kenntnisse des Berufsbildes zu verfügen (z. B. durch eine vor dem Studium absolvierte Ausbildung in diesem Bereich), wurden von der weiteren Analyse ausgeschlossen ($N = 5$ der ursprünglich 76 Teilnehmer). Die 71 Laien waren im Mittel 24.5 Jahre alt ($SD = 4.16$).

² Aufgrund mangelnder Angaben beträgt das N für die zuletzt genannten Mittelwerte $N = 110$ (Alter) sowie $N = 116$ (Erfahrung).

Zusätzlich wurde als eine Art „baseline“ eine Gewichtung der Dimensionen vorgenommen, die nicht auf einer Anforderungsanalyse basiert. Hierzu wählten wir 4. die häufig in der Literatur anzufindende gleichmäßige Gewichtung aller Dimensionen, im englischsprachigen Raum als „unit-weighting“ bezeichnet. Bobko, Roth und Buster (2007) führten eine Metaanalyse zur Evaluation des „unit-weighting“-Ansatzes durch. Ihrer Studie zufolge korrelieren die durch Experten generierten Gewichte zu .98 mit den „unit weight scores“.

2.2 Annahmen

Sofern die Anforderungsanalyse valide Ergebnisse erbringt, sollten die auf der Basis von Anforderungsanalysen gewichteten Testergebnisse (1 bis 3) zu einer höheren Kriteriumsvalidität führen als die Gewichtung, die auf eine Anforderungsanalyse verzichtet (4).

Darüber hinaus sollten die auf der Basis der Befragung von Laien abgeleiteten Gewichte (3) zu einer geringeren Kriteriumsvalidität der Testbatterie führen, als die Gewichte, die aufgrund von Expertenbefragungen (1 und 2) gewonnen wurden. Mit dem Vergleich von Experten und Laien wenden wir den Ansatz von z. B. Smith und Hakel (1979) erstmalig auf Anforderungsanalysen an (die bisherigen Untersuchungen bezogen sich auf Arbeitsanalysen).

2.3 Anforderungsanalyse

Basis der Studie war ein Anforderungsanalyseverfahren, bei dem die Bedeutung von Fähigkeiten, Fertigkeiten und Kenntnissen für eine erfolgreiche Bewältigung der Ausbildung direkt beurteilt wird. Der in der Untersuchung verwendete anforderungsanalytische Fragebogen wurde von Kersting (1997) entwickelt, er umfasst zehn Items. Diese beschreiben die sieben hochgradig generellen intellektuellen Fähigkeiten des Berliner Intelligenzstrukturmodells von Jäger (1984). Darüber hinaus werden mit drei weiteren Items die Bedeutsamkeiten der Merkmale Arbeitsverhalten (AV), gegenwartskundliche Kenntnisse (GK) und Rechtschreibleistungen (R) erfragt. Zunächst wird, zur Sicherung eines einheitlichen Verständnisses, jede Dimension kurz beschrieben. Ein Beispiel für eine solche Dimensionsbeschreibung lautet: „Merkfähigkeit: Aktives Einprägen und kurz- oder mittelfristiges Wiedererkennen von Informationen“.

Anschließend werden die einschätzenden Personen gebeten, auf einer fünfstufigen Likert-Skala (von „unwichtig“ bis „sehr wichtig“) die Bedeutung der Dimension für die erfolgreiche Bewältigung der Ausbildung in dem entsprechenden Berufsbild anzugeben. Diese Art von Befragung stellt ein übliches Vorgehen im Bereich der Anforderungsanalyse dar (vgl. z. B. Harvey, 1991). Prominente, wenngleich ungleich umfangreichere und ausführlicher operationalisierte, Verfahrensvertreter sind die Ability Requirement Scales (ARS; Fleishman & Quaintance, 1984).

2.4 Zum Zusammenhang der Bedeutsamkeit der Anforderungen und deren sozialer Erwünschtheit

Um zu prüfen, ob es einen Zusammenhang zwischen der eingeschätzten Bedeutsamkeit der Anforderungen einerseits und deren sozialer Erwünschtheit andererseits gibt, wurden die Laienbeurteiler (Gruppe 3) im Anschluss an die eigentliche Anforderungsanalyse gefragt, für wie wünschenswert sie die bislang beurteilten Eigenschaften allgemein einschätzen. Die Instruktion lautete: „Die folgende Beurteilung nehmen Sie bitte unabhängig von der Position des/r Beamten/in in der Kommunalverwaltung vor. Beurteilen Sie hier ganz allgemein, für wie wünschenswert Sie persönlich folgende Eigenschaften halten. Bitte beurteilen Sie auf einer Skala von 1 bis 5, mit 1 = unerwünscht und 5 = sehr erwünscht.“ Trotz der Instruktion, die Einstufung der sozialen Erwünschtheit unabhängig von der Position des Beamten vorzunehmen, kann nicht ausgeschlossen werden, dass die Beurteilung der Erwünschtheit von der vorher erfolgten Beurteilung der Bedeutsamkeit der gleichen Merkmale beeinflusst wurde.

2.5 Testverfahren

Zur Auswahl von Anwärtern für die Laufbahn zum gehobenen Dienst (allgemeine innere Verwaltung) wurde der Berliner Intelligenzstruktur-Test der Deutschen Gesellschaft für Personalwesen (BIS-r-DGP Test) eingesetzt. Der BIS-r-DGP Test wurde von Kersting und Beauducel (1997, 2004) auf Basis des Berliner Intelligenzstruktur Tests (BIS-4 Test) von Jäger, Süß und Beauducel (1997) entwickelt. Die Faktorwerte des BIS-r-DGP Tests sind innerhalb der Operations- und Inhaltsklassen des BIS-Modells gering korreliert (siehe Kersting & Beauducel, 1997), was eine trennscharfe Interpretation dieser Werte ermöglicht und Voraussetzung der hier durchgeführten Analysen ist. Zusätzlich zu den mit dem BIS-r-DGP Test erfassten Dimensionen werden die Kenntnisse der Bewerber in Gegenwartskunde und Rechtschreibung erfasst und das Arbeitsverhalten wird geprüft. Insgesamt werden somit zehn Dimensionen erhoben (siehe Tabelle 1).

2.6 Kriterium

Für die Studie wurden Reanalysen einer von Thielepape und Kersting (2005) durchgeführten Untersuchung zur Kriteriumsvalidität des BIS-r-DGP Tests vorgenommen. Die Treffsicherheit der Testbatterie wurde anhand der Vorhersage der Laufbahnprüfungsergebnisse (Abschluss der theoretischen Ausbildung) von 140 Personen bestimmt. Weitere demografische Angaben (z. B. Alter) liegen aus Gründen der Anonymisierung nicht vor.

3 Ergebnisse

Bezüglich der Interkorrelation der Bedeutsamkeitsurteile der unterschiedlichen Beurteilergruppen (Mittelwerte über alle Beurteiler einer Gruppe) wurden die folgenden Werte ermittelt: Die Beurteilungen von internen Stelleninhabern korrelieren zu $r = .76$ mit den Beurteilungen der externen Stellenexperten und zu $r = .75$ mit den Urteilen der Laien. Mit $r = .85$ fällt die Korrelation der Urteile der Personalexperten mit den Urteilen der Laien nominell höher aus.

Die Umrechnung der Mittelwerte der anforderungsanalytischen Beurteilungen zu den Gewichten für die Testdimensionen erfolgte durch eine Lineartransformation der Mittelwerte in Prozentwerte. Tabelle 1 zeigt die Ergebnisse für die drei Gruppen.

Der größte Bedeutsamkeitsunterschied zwischen den Einschätzungen der Stelleninhaber/Vorgesetzten einerseits und den Laien andererseits ergibt sich bei der Einfallsmenge: Laien schätzen diese Teilkomponente der Kreativität als weniger bedeutsam ein als Stelleninhaber und Vorgesetzte. Dafür sprechen die Laien den Kenntnissen in der Gegenwartskunde eine höhere Bedeutung zu.

Unterschiede zwischen den Einschätzungen der internen Stellenexperten sowie Laien einerseits und den externen Stellenexperten (Subject Matter Experts, SMEs) zeigen sich u. a. bezüglich der Verarbeitungskapazität (so bezeichnet Jäger, 1984, die Dimension, die häufig auch schlussfolgerndes Denken genannt wird). Die SMEs sprechen dieser Dimension eine wesentliche größere Bedeutung zu als die Laien und internen Stellenexperten. Darüber hinaus sehen die SMEs das sprachliche Denken als bedeutsamer an. Dem gegenüber ist die Bedeutung der Rechtschreibkenntnisse nach Ihrer Ansicht zu vernachlässigen, auch die Einfallsmenge, das figural-bildhafte Denken und das Arbeitsverhalten werden in ihrer Bedeutsamkeit von den SMEs deutlich schwächer eingeschätzt als von den anderen beiden Gruppen.

Die in Tabelle 1 berichteten Gewichte wurden verwendet, um die Ergebnisse, welche die Teilnehmer der Studie zur Kriteriumsvalidität (Thielepape & Kersting, 2005) in den unterschiedlichen Testdimensionen erzielt haben, zu einem Gesamttestwert zusammenzufassen. Insgesamt liegen die entsprechend dieser drei Gewichtungen gebildeten Gesamttestwerte eng beieinander. Dies ist zum Teil auf die Interkorrelationen der Skalen zurückzuführen. Um diesen Effekt zu kontrollieren wurde eine zusätzliche Analyse nur auf Basis der vier, nur gering interkorrelierten, operativen Faktoren des BIS-Modells gerechnet. Zusätzlich zu den Aggregaten auf der Basis der anforderungsanalytischen Gewichtungen wurde für beide Varianten (zehn sowie vier Dimensionen) ein Testgesamttestwert nach der Methode des unit-weighting gebildet (d. h. bei der Variante der zehn Dimensionen erhält jede Dimension das Gewicht 10, bei der Vier-Dimensionen-Variante das Gewicht 25).

Trotz der Ähnlichkeit der resultierenden Gesamttestwerte unterscheiden sich diese in ihrer Kriteriumsvalidität (vgl. Tabelle 2), wobei die Gewichtungen der Testkomponenten entsprechend den Bedeutsamkeitsaussagen der SMEs zu den nominell besten Ergebnissen führen. Die Transformation der anforderungsanalytischen Aussagen der internen Stellenexperten (Stelleninhaber und Vorgesetzte) zu Gewichten für die Testbatterie führt nicht zu besseren Ergebnissen als die Nutzung der anforderungsanalytischen Aussagen der Laien. Die Ergebnisse halten einer Kreuzvalidierung stand.

Tabelle 1: Aus den Ergebnissen der Anforderungsanalyse abgeleitete Gewichte für die Testdimensionen (in Prozent) für Variante 1 mit zehn Dimensionen und für Variante 2 (in Klammern) für die vier Operationen des Denkens

Testdimensionen	Interne Stellenexperten	Externe Stellenexperten (SMEs)	Laien
Bearbeitungsgeschwindigkeit (B)	9.1 (21.4)	10 (18.9)	10 (24.5)
Merkfähigkeit (M)	10.7 (25)	8 (15.1)	10.9 (26.8)
Einfallsmenge (sprachlich) (E)	10.9 (25.6)	5 (9.4)	8.9 (21.7)
Verarbeitungskapazität (K)	12 (28.1)	30 (56.6)	11 (27)
Figural-bildhaftes Denken (F)	7.3	4	6.8
Sprachgebundenes Denken (V)	11.7	15	11.3
Zahlgebundenes Denken (N)	9.9	10	10.4
Arbeitsverhalten, -tempo u. -sorgfalt (AV)	7.9	6	8.8
Gegenwartskundliche Kenntnisse (GK)	9.4	10	10.9
Rechtsschreibkenntnisse (R)	11.2	2	10.9

Anmerkung: SMEs: Subject Matter Experts.

Um zu prüfen, ob es einen Zusammenhang zwischen der eingeschätzten Bedeutsamkeit der Anforderungen einerseits und deren sozialer Erwünschtheit andererseits gibt, wurden die Bedeutsamkeitseinschätzungen der internen Stelleninhaber mit den Einschätzungen der sozialen Erwünschtheit der Dimensionen korreliert. Der jeweilige Grad der sozialen Erwünschtheit der Eigenschaften wurde durch die Gruppe der Laien festgesetzt, die Dimensionen Arbeitsverhalten und Rechtsschreibkenntnisse wurden beispielsweise allgemein, unabhängig von einer konkreten Stelle, als besonders wünschenswert angesehen, während das figural-bildhafte Denken der Einschätzung der Laien zufolge eine allgemein nur wenig erwünschte Fähigkeit ist. Die Korrelation der Mittelwerte der Anforderungsanalyse von internen Stellenexperten mit den Mittelwerten der Beurteilung der Erwünschtheit beträgt $r = .80$.

Tabelle 2: Kriteriumsvalidität der Testbatterie in Abhängigkeit der unterschiedlichen Gewichtungsprozeduren für die einzelnen Dimensionen

Gewichtungen	Variante 1: zehn Dimensionen	Variante 2: nur vier Operationsfaktoren
Externe Stellenexperten	.256**	.243**
unit-weighting	.239**	.165
Laien	.244**	.172*
Interne Stellenexperten	.234**	.170*

Anmerkung: Korrelationen nach Pearson, ** = signifikant ($p < .01$, zweiseitig), * = signifikant ($p < .05$, zweiseitig).

4 Einschränkungen der Aussagekraft der Studie

Die Kriteriumsvalidität der auf der Basis der Anforderungsanalyse gewichteten Testverfahren gibt Aufschluss darüber, wie nützlich die Anforderungsanalyse bei der Interpretation von Testergebnissen ist. Sie stellt aber strenggenommen keine Validitätsprüfung für die Anforderungsanalyse dar. Dies liegt u. a. daran, dass die Einstufung der Bedeutsamkeit der Eigenschaften auf der Konstruktebene erfolgt, die Zuordnung von Gewichten aber auf der Ebene der Operationalisierung der Konstrukte durch Testverfahren. So ist es denkbar, dass die Stellenexperten und Laien der Dimension Einfallsmenge als Aspekte der Kreativität zu Recht ein hohe Bedeutsamkeit zuerkennen, eine hohe Gewichtung der entsprechenden Dimension bei der Bildung des Gesamtwertes sich aber nicht günstig auf die Kriteriumsvalidität auswirkt, weil z. B. die entsprechenden Testverfahren nur über eine vergleichsweise geringe Reliabilität verfügen. Weitere Aspekte sind die Interkorrelationen der Dimensionen sowie Stichprobenfehler. Dieser methodischen Gesichtspunkte könnte die Überlegenheit der anforderungsanalytischen Aussagen der Subject Matter Experts erklären, falls diese zusätzlich zu den Konstrukten auch Informationen über die Testverfahren in ihrem Urteil berücksichtigen.

Darüber hinaus kann in Frage gestellt werden, ob die Studenten zu Recht als Laien charakterisiert sind. Zwar ist es üblich, Studenten als Laienbeurteiler zu nutzen, Morgeson und Campion (1997, S. 643) bezeichnen Studierende gar als „truly naive raters“, da diese häufig nicht nur die zu beurteilende, sondern noch gar keinen Beruf ausgeführt haben. Als Studenten der Psychologie könnten die „Laien“ aber über eine gewisse Expertise hinsichtlich der Eigenschaften (Konstrukte) verfügen.

Diese Überlegungen stellen allerdings nur den Vergleich zwischen den internen Stellenexperten und den Laien in Frage, nicht aber den Vergleich zwischen den internen und externen Stellenexperten sowie dem Vergleich zwischen Stellenexperten und der Methode der Gleichgewichtung der Testdimensionen, die auf jegliche Anforderungsanalyse verzichtet. Betrachtet man den Fall, in dem nur die Bedeutung der vier Operationen des Denkens berücksichtigt wird, erweist sich die Anforderungsanalyse mit internen Stelleninhabern als Gewichtungsbasis aus dem Blickwinkel der Kriteriumsvalidität als schlechter als die Variante mit externen Stellenexperten. Mithilfe der Aussagen der Stellenexperten erzielt man kein besseres Ergebnis als mit einer Gleichgewichtung, die auf jede Anforderungsanalyse verzichtet. Angesichts des Aufwands von Anforderungsanalysen würde man erwarten, dass die Befragung von internen Stelleninhabern in einen deutlichen Validitätsvorteil mündet.

5 Zusammenfassung und Ausblick

Die Studie erbrachte vor allem die folgenden Befunde:

- Laien, die wenig bis keine Kenntnisse über eine Ausbildung haben, beurteilen die auf der Eigenschaftsebene erhobenen Anforderungen der in Frage stehenden Ausbildung sehr ähnlich wie interne Stellenexperten (Stelleninhaber und der Vorgesetzte), die über umfassende Erfahrung mit der Ausbildung verfügen.

- Die Befragung von internen Stellenexperten erbringt bei der Ermittlung des relativen Gewichts, das einer Testdimension in einer Verfahrensbatterie zukommt, unter dem Gesichtspunkt der Kriteriumsvalidität keinen Vorteil gegenüber einer Gleichgewichtung der Dimensionen.
- Die allgemeine Beurteilung der Erwünschtheit von Personmerkmalen korreliert hoch mit den Aussagen zur Bedeutsamkeit dieser Merkmale für eine spezifische Ausbildung.

Diese Befunde können unterschiedlich erklärt werden. Möglicherweise verfügen interne Stellenexperten über kaum mehr als ein stereotypes Wissen über die Anforderungen ihrer Stelle. Oder die die Anforderungen der Stelle sind so offensichtlich, dass diese auch von Außenstehenden korrekt eingeschätzt werden können. Schließlich ist es auch möglich, dass interne Stellenexperten zwar über nützliches Wissen über die Stellenanforderungen verfügen, die Testbatterie aber nicht in der Lage ist, die mit diesen Anforderungen korrespondierenden Information reliabel und valide zu erfassen. Darüber hinaus können Methodeneffekte (z. B. Befragung auf der Eigenschaftsebene, Besonderheiten der Testbatterie) sowie Stichprobeneffekte (Befragung von Studenten der Psychologie) zur Erklärung herangezogen werden.

Die mit der Studie aufgeworfenen Fragen müssen in weiteren Studien geklärt werden. Als praktisches Ergebnis lässt sich allerdings bereits ableiten, dass eine solide Anforderungsanalyse nicht nur einer Quelle (z. B. den internen Stellenexperten) trauen sollte. Im Sinne des am Lehrstuhl für Psychologie der Universität Hohenheim von Schuler für die Personalpsychologie erarbeiteten Prinzips der Multimodalität gilt es, unterschiedliche Informationen kontrolliert zu kombinieren. Der vorliegende Beitrag stellt insbesondere die Bedeutung der externen Experten, der SMEs heraus, die nicht nur über Expertise bezüglich der Anforderungen, sondern auch bezüglich der Möglichkeiten und Grenzen der Messinstrumente verfügen. Insgesamt zeigt die Studie, dass zum Thema Validität der Anforderungsanalyse noch umfassende theoretische und empirische Arbeiten ausstehen. Der Rückzug auf die Analyse der Beurteilerübereinstimmung ist unbefriedigend. Die in der vorliegenden Studie aufgezeigte eingeschränkte Nützlichkeit der anforderungsanalytischen Ergebnisse der Stelleninhaber für die Gewichtung der Testdimensionen wäre bei einer alleinigen Betrachtung der (unauffällig bis guten) Beurteilerübereinstimmung nicht erkannt worden.

„Nur wer sein Ziel kennt, findet den Weg.“ – Dieses unverbürgte Zitat wird u. a. Lao-Tse zugeschrieben. Anforderungsanalysen sind und bleiben unverzichtbar, zahlreiche Fragen nach der Vorgehensweise und Nützlichkeit sind aber noch ungeklärt. Sie müssen produktiv gestellt werden, um der Praxis möglichst bald gute Argumente und umsetzbare Einsichten zu bieten. Aktuell wird die Anforderungsanalyse häufig als eigenständiges, gesondertes Thema behandelt. Produktivitätsfortschritte sind zu erwarten, wenn die Anforderungsanalyse theoretisch und empirisch mit der geplanten Intervention (z. B. Eignungsdiagnostik, Personalentwicklung, Personalmarketing usw.) verknüpft wird. Dies würde bedeuten, nicht nur zu beschreiben, wie eine Anforderungsanalyse durchgeführt wird, sondern auch darzustellen, wie die Ergebnisse der Anforderungsanalyse genutzt werden sollen – wobei die dabei formulierten Annahmen empirisch zu validieren sind. Diesbezüglich stehen noch umfassende Ar-

beiten aus. Aber vielleicht ist nicht das Schlechteste über ein Thema gesagt, wenn man sich dessen intensivere Erforschung wünscht.

Literatur

- Ash, R. A. (1988). Job analysis in the world of work. In S. Gael (Ed.), *The Job Analysis Handbook for Business, Industry, and Government* (pp. 3-13). New York: John Wiley and Sons.
- Birk, M. (2004). *Zur Validität der Anforderungsanalyse. Eine theoretische und empirische Exploration* (Unveröffentlichte Diplomarbeit). Aachen: RWTH Aachen.
- Bobko, P., Roth, P. L. & Buster, M. A. (2007). The usefulness of unit weights in creating composite scores. A literature review, application to content validity, and meta-analysis. *Organizational Research Methods, 10*, 689-709.
- Bobrow, W. S. & Leonards, J. S. (1997). Development and validation of an assessment center during organizational change. *Journal of Social Behavior and Personality, 12*, 217-236.
- Borman, W. C., Dorsey, D. & Ackerman, L. (1992). Time-spent responses as time allocation strategies: Relations with sales performance in a stockbroker sample. *Personnel Psychology, 45*, 763-777.
- Brannick, M. T. & Levine, E. L. (2002). *Job analysis: Methods, research, and applications for human resource management in the new millenium*. Thousand Oaks: Sage Publications.
- Cornelius, E. T., Denisi, A. S. & Blencoe, A. G. (1984). Expert and naive raters using the PAQ: Does it matter? *Personnel Psychology, 37*, 453-464.
- Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571-582.
- Dierdorff, E. C. & Wilson, M. A. (2003). A meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635-646.
- DIN (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- Fleishman, E. A. & Quaintance, M. K. (1984). *Taxonomies of human performance: The description of human tasks*. Orlando, FL: Academic Press.
- Friedman, L. & Harvey, R. J. (1988). Can raters with reduced job descriptive information provide accurate Position Analysis Questionnaire (PAQ) Ratings? *Personnel Psychology, 39*, 779-789.
- Gibson, S. G., Harvey, R. J. & Harris, M. L. (2007). Holistic versus decomposed ratings of general dimensions of work activity. *Management Research News, 30*, 724-734.
- Harvey, R. J. (1991). Job Analysis. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 71-163). Palo Alto, CA: Consulting Psychologists Press.
- Harvey, R. J. & Hayes, T. L. (1986). Monte Carlo baselines for interrater reliability correlations using the Position Analysis Questionnaire. *Personnel Psychology, 39*, 345-357.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau, 35*, 21-35.

- Jäger, A. O., Süß, H. M. & Beauducel, A. (1997). *Berliner Intelligenzstruktur-Test. Form 4. Handanweisung*. Göttingen: Hogrefe.
- Jones, A. P., Main, D. S., Butler, M. C. & Johnson, L. A. (1982). Narrative job description as potential sources of job analysis ratings. *Personnel Psychology*, 35, 813-827.
- Kersting, M. (1997). FAAV. Fragebogen Anforderungen in der allgemeinen Verwaltung. *Unveröffentlichtes Verfahren der Deutschen Gesellschaft für Personalwesen e.V.* Hannover: DGP.
- Kersting, M. (2008). *Qualität in der Diagnostik und Personalauswahl: Der DIN Ansatz*. Göttingen: Hogrefe.
- Kersting, M., Althoff, K. & Jäger, A. O. (2008). *Testmanual/Verfahrenshinweise zum WIT-2 (Wilde-Intelligenztest)*. Göttingen: Hogrefe.
- Kersting, M. & Beauducel, A. (1997). Der neue DGP-Leistungstest »BIS-r-DGP«: Informationen zu ausgewählten Testgütekriterien und zur Normierung. *DGP Informationen*, 46, 92-102.
- Kersting, M. & Beauducel, A. (2004). BIS-r-DGP (sowie die Kurzformen A-1 und K-1). Berliner Intelligenzstruktur-Test der Deutschen Gesellschaft für Personalwesen e. V. In W. Sarges & H. Wottawa (Hrsg.), *Handbuch wirtschaftspsychologischer Testverfahren* (2. Aufl., S. 149-157). Lengerich: Pabst Science Publishers.
- McCormick, E. J., Jeanneret, P. R. & Mecham, R. C. (1972). A study of job characteristics and job dimensions as based on the Position Analysis Questionnaire (PAQ). *Journal of Applied Psychology*, 56, 347-368.
- McCormick, E. J., DeNisi, A. S. & Shaw, J. B. (1979). Use of the Position Analysis Questionnaire for establishing the job component validity. *Journal of Applied Psychology*, 64, 51-56.
- Morgeson, F. P. & Campion, M. A. (1997). Social and cognitive sources of potential inaccuracy in job analysis. *Journal of Applied Psychology*, 82, 627-655.
- Morgeson, F. P. & Campion, M.A. (2000). Accuracy in job analysis: toward an inference-based model. *Journal of Applied Psychology*, 21, 819-827.
- Moser, K., Donat, M., Schuler, H. & Funke, U. (1989). Gütekriterien von Arbeitsanalysen. *Zeitschrift für Arbeitswissenschaften*, 43, 65-72.
- Pine, D. E. (1995). Assessing the validity of job ratings: An Empirica. *Public personnel management*, 24, 451-460.
- Sanchez J. I. & Levine E. L. (2001). Analysis of work in the 20th and 21st centuries. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Eds.), *Handbook of Industrial, Work and Organizational Psychology* (pp. 71-89). Thousand Oaks, CA: Sage Publications.
- Schippmann, J. S., Hughes, G.L. & Prien, E.P. (1987). The use of structured multidomain job analysis for the construction of assessment center methods and procedures. *Journal of Business & Psychology*, 1, 353-366.
- Schuler, H. (1989). Some advantages and problems of job analysis. In M. Smith & T. Robertson (Eds.), *Advances in selection and assessment* (pp. 31-42). New York: John Wiley and Sons.
- Schuler, H. (2006a). Stand und Perspektiven der Personalpsychologie. *Zeitschrift für Arbeits- und Organisationspsychologie*, 4, 176-188.
- Schuler, H. (2006b). Arbeits- und Anforderungsanalyse. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (2. Aufl., S. 45-68). Göttingen: Hogrefe.

- Schuler, H. & Funke, U. (1995). Diagnose beruflicher Eignung und Leistung. In H. Schuler (Hrsg.), *Organisationspsychologie* (S. 235-284). Bern: Hans Huber.
- Schuler, H., Funke, U., Moser, K. & Donat, M. (1995). *Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren*. Göttingen: Hogrefe.
- Smith, J. E. & Hakel, M. D. (1979). Convergence among data sources, response bias, and reliability and validity of a structured job analysis questionnaire. *Personnel Psychology*, 32, 677-692.
- Schmitt, N. & Cohen, S. A. (1989). Internal analysis of task ratings by job incumbents. *Journal of Applied Psychology*, 74, 96-104.
- Sparrow, J. (1989). The utility of PAQ in relating job behaviours to traits. *Journal of Occupational Psychology*, 62, 151-162.
- Tergan, S.O. (1988). Qualitative Wissensdiagnose. In H. Mandl & H. Spada (Hrsg.), *Wissenspsychologie* (S. 400-422). Weinheim: Psychologische Verlags Union.
- Thielepape, M. & Kersting, M. (2005). Evaluation und Qualitätsoptimierung in der Personalauswahl: Zwei Bewährungskontrollen. *DGP Informationen*, 49, 2-20.
- Voskuil, O. F. & van Sliedregt, T. (2002). Determinants of interrater reliability of job analysis: A meta-analysis. *European Journal of Psychological Assessment*, 18, 52-62.