Originalarbeit

Validitätsbefunde zum Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP)

Vorhersage des Berufserfolgs durch klassische und neuere Validierungsmethoden

Robin Merchel[1], Philip Frieg[2] und Rüdiger Hossiep[2]

[1]Lehrstuhl für Industrial Sales and Service Engineering, Fakultät für Maschinenbau, Ruhr-Universität Bochum
[2]Fakultät für Psychologie, Ruhr-Universität Bochum

Zusammenfassung: Das Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) erfasst berufsbezogene Persönlichkeitsmerkmale und kann in linearen Regressionen verschiedene Maße subjektiven und objektiven Berufserfolgs aufklären. Um zusätzliche Nachweise für die Kriteriumsvalidität zu erbringen, werden in der vorliegenden Arbeit Cluster- und Klassifikationsverfahren verwendet. Mithilfe von k-Means-Clusteranalysen können typische Persönlichkeitsstrukturen identifiziert werden: Personen, die sich durch Flexibilität und Gestaltungsmotivation auszeichnen, weisen einen bedeutsamen Zusammenhang zu höheren beruflichen Entgelten auf, während solche, die durch emotionale Instabilität und geringe Durchsetzungsstärke geprägt sind, häufig ein niedriges Entgelt erzielen. Klassische und neuere Klassifikationsverfahren (logistische Regressionen bzw. Random Forests) besitzen substantielle Trefferquoten in der Identifikation von Mitarbeitenden als Fach- oder Führungskraft. Die Ergebnisse sind als mittlere bis große Effekte einzustufen und liefern damit einen Nachweis über die Relevanz der Persönlichkeit für beruflichen Erfolg.

Schlüsselwörter: Persönlichkeitsfragebogen, Berufserfolg, Clusteranalyse, Klassifikation, Random Forests, BIP

Validity Findings on the Business-Focused Inventory of Personality (BIP). Prediction of Job Success by Classical and Newer Validation Methods

Abstract: The Business-Focused Inventory of Personality (BIP) measures job-related personality traits. It is able to explain variance in various measures of subjective and objective occupational success by means of linear regressions. In this paper, we used cluster and classification analyses to provide additional evidence for the criterion validity. Typical personality structures can be identified by k-means cluster analyses: Persons characterized by flexibility and power motivation receive higher wages than those characterized by emotional instability and low assertiveness, with a medium-high probability. Classical and newer classification methods (e. g., logistic regressions or random forests, respectively) have a significantly higher hit rate than random probability in identifying employees as specialists or managers using the BIP traits. The results can be classified as medium-to-large effects and thus provide evidence of the relevance of personality for professional success.

Keywords: personality questionnaire, occupational success, cluster analysis, classification, random forests, BIP

Ein Verfahren der Personalauswahl, das auf wissenschaftlich fundierten Methoden basiert, gereicht Arbeitnehmenden, Arbeitgebenden und weiteren Parteien innerhalb wie außerhalb einer Organisation zum Vorteil (Kanning, 2017). Die personalpsychologische Forschung hat bereits viele Erkenntnisse darüber gewonnen, aus welchen Elementen eine gute Personalauswahl bestehen sollte. Diese Befunde werden in der Zusammenstellung nicht-harmonisierter meta-analytischer Befunde von Schmidt und Hunter (1998) zusammengefasst. Jedoch zeigen die Untersuchungen von Ryan, McFarland, Baron und Page (1999) auf internationaler Ebene, von Armoneit, Schuler und Hell (2020) für Deutschland sowie von König, Klehe, Berchtold und Kleinmann (2010) für die Deutschschweiz, dass Verfahren mit guter Vorhersagefähigkeit für berufliche Leistung nicht so verbreitet sind, wie zu erwarten wäre. Der vorliegende Artikel versucht, mithilfe klassischer und neuerer, aus dem maschinellen Lernen stammender Validierungsmethoden Erkenntnisse zur Vorhersagegüte eines berufsbezogenen Persönlich-

keitsfragebogens zu generieren, u. a. auch über demographische Angaben hinaus. Dies könnte unterstreichen, dass Persönlichkeit ein validerer Prädiktor für Führungsqualitäten ist als z. B. Alter oder Geschlecht.

Persönlichkeit als Prädiktor im Berufsleben

Das Testgütekriterium der Validität beantwortet die Frage, ob ein psychometrischer Test tatsächlich das misst, was er zu messen vorgibt, und baut auf den Testgütekriterien der Objektivität (Unverfälschbarkeit u. a. durch Testleiter_innen und Testauswerter_innen) und der Reliabilität (geringe Beeinflussung durch Messfehler) auf (Moosbrugger & Kelava, 2020). Kriteriumsvalidität als Fähigkeit zur Vorhersage von Außenkriterien (z. B. Leistungsbeurteilung oder Gehalt) ist in der beruflichen Eignungsdiagnostik von besonderer Wichtigkeit (Kauffeld & Grohmann, 2019). In der Regel wird die Kriteriumsvalidität dadurch berechnet, dass die Varianz in Kriterien von Berufsleistung oder Berufserfolg durch die Ergebnisse des Testverfahrens vorhergesagt werden. Hierzu wird in der Regel die multiple Korrelation R zwischen den durch die Testergebnisse vorhergesagten Kriteriumswerten und dem Kriterium oder dessen Quadrat, der Determinationskoeffizient R^2, berechnet (s. a. Bortz & Schuster, 2010).

Neben generellen und insbesondere spezifischen (z. B. visuospatialen, verbalen und mathematischen) mentalen Fähigkeiten (Lang & Kell, 2020) sind auch Persönlichkeitsmerkmale valide Prädiktoren beruflicher Erfolgskriterien: Die „Big Five“ (Neurotizismus, Extraversion, Offenheit, Verträglichkeit und Gewissenhaftigkeit) weisen eine multiple Korrelation zur individuellen Arbeitsleistung von $R = .27$ auf (Ones, Dilchert, Viswesvaran & Judge, 2007). Der valideste Prädiktor ist hierbei die Gewissenhaftigkeit (Barrick, Mount & Judge, 2001), die zusammen mit Intelligenz eine gemeinsame Validität von $R = .60$ aufweist (Mount & Barrick, 1995) und damit einer der stärksten Prädiktoren über Intelligenz hinaus ist (Schmidt & Hunter, 1998).

Zur Forschungsversion des berufsbezogenen Persönlichkeitsfragebogens BIP (Hossiep & Paschen, 2019) berichten Weiß und Hossiep (2013) weiterhin für das Kriterium Entgelt eine korrigierte Varianzaufklärung von $R^2_{adj} = .17$ und für das Kriterium hierarchische Position $R^2_{adj} = .20$, was als mittlere bis mittelhohe Varianzaufklärung interpretiert werden kann. Zu beiden Kriterien wies die Führungsmotivation hierbei den stärken Zusammenhang auf.

Ryan et al. (1999) ließen die Häufigkeit des Einsatzes von Methoden der Personalauswahl auf einer Skala von 1 (nie) bis 5 (fast immer oder immer) bewerten und fanden, dass der Einsatz von Persönlichkeitstests in Deutschland besonders selten stattfindet ($M = 1.70$); im benachbarten Belgien liegt dieser Wert deutlich höher ($M = 3.75$). Armoneit et al. (2020) berichten, dass lediglich 19 % der 140 befragten Unternehmen Persönlichkeitstests und 24 % Online-Persönlichkeitstests als Auswahlverfahren einsetzen. In deutschen Großunternehmen/-konzernen werden hierbei der MBTI (28 %), das DISG/persolog-Persönlichkeitsprofil (27 %) und der BIP-Fragebogen (23 %) am häufigsten eingesetzt (Hossiep, Schecke & Weiß, 2015). Für den MBTI (Pittenger, 1993) und das DISG/persolog-Persönlichkeitsprofil (König & Marcus, 2013) muss eine Validitätsbewertung jedoch kritisch ausfallen – für den BIP-Fragebogen positiver (Marcus, 2004). Dies bedeutet, dass die Validität eines Fragebogens in der Praxis offensichtlich nur selten als Kriterium für dessen Einsatz herangezogen wird.

Wenngleich die multiplen Korrelations- und Regressionsanalysen (wie z. B. $R = .27$ oder $R^2_{adj} = 17\,\%$) in der Arbeits- und Organisationspsychologie die meistgewählten statistischen Methoden im Rahmen der Validierungsstrategie eines Testverfahrens mithilfe von Berufserfolgskriterien darstellen, ist auch der Einsatz alternativer Methoden denkbar. Die vorliegende Studie möchte dies am Beispiel des BIP (Hossiep & Paschen, 2019) demonstrieren, da dieses bislang lediglich mit Hilfe multipler Korrelationen und linearer Regressionen validiert wurde. Weitere Validitätsbefunde liefern stärkere Argumente für den Einsatz von Persönlichkeitsinventaren in der Personalarbeit und könnten eine Möglichkeit darstellen, mehr Personalverantwortliche hiervon zu überzeugen.

Eine Möglichkeit zur Generierung weiterer, anschaulicher Validitätsbefunde liefern Clusteranalysen, die bereits in der Big-Five-Forschung zur Identifikation typischer Profilstrukturen genutzt wurden. Die dort lange Zeit anerkannte Clusterlösung eines resilienten, eines unterkontrollierten und eines überkontrollierten Clusters (Robins, John, Caspi, Moffitt & Stouthamer-Loeber, 1996) wurde jedoch von Gerlach, Farb, Revelle und Amaral (2018) als instabil kritisiert, weswegen sie eine alternative, in mehreren großen Datensätzen stabile Lösung mit vier Clustern („Durchschnitt“, „Selbstzentriert“, „Reserviert“ und „Rollenmodell“) empfahlen.

Auch Frieg und Schulz (2014) identifizierten fünf typische BIP-Profile, die sie der dichotomen Variable „Beschäftigungsstatus“ (berufstätig vs. arbeitssuchend) gegenüberstellten. Sie fanden beispielsweise, dass durch hohe Führungsmotivation und Kontaktfähigkeit, aber geringe Gewissenhaftigkeit gekennzeichnete Profile in der Gruppe der Arbeitstätigen häufiger vertreten sind, wäh-

rend Profile, die durch geringe Ausprägungen in der psychischen Konstitution und den sozialen Kompetenzen gekennzeichnet sind, zumeist unter den Arbeitssuchenden auftreten. Diese Vorgehensweise möchte die vorliegende Arbeit fortführen und dabei das berufliche Entgelt als objektives Kriterium des Berufserfolgs verwenden. Es ergibt sich die Hypothese:

H1. Personen, die sich in der Gesamtstruktur ihres Profils im BIP unterscheiden, weisen ebenfalls Unterschiede im beruflichen Entgelt auf.

Weiterhin fanden bereits Hossiep, Mutwill und Schulz (2012), dass für Fach- und Führungskräfte teilweise unterschiedliche Merkmale erfolgsentscheidend sind: Die Korrelationskoeffizienten fallen beispielsweise für Führungskräfte bei Begeisterungsfähigkeit (einer Zusatzskala der BIP-Forschungsversion) und für Fachkräfte bei Gewissenhaftigkeit (jeweils mit unterschiedlichen Erfolgskriterien) größer aus. Diese Ergebnisse sind jedoch nicht peer-reviewed publiziert worden, sondern lediglich in einem Kongress-Band veröffentlicht worden. Sind nun in der Gruppe der Fachkräfte im Mittel überzufällig häufig andere Persönlichkeitsausprägungen als bei den Führungskräften vorhanden, so sollte anhand eines BIP-Profils mit Klassifikationsanalysen eine Prognose darüber möglich sein, ob eine Person sich innerhalb einer Fach- oder einer Führungskarriere befindet. Diese Prognosegüte sollte weiterhin über demografische Faktoren wie Alter und Geschlecht oder auch über den aktuellen Einsatzbereich im Unternehmen hinaus inkrementelle Validität aufweisen, um den Einsatz von Persönlichkeitstests rechtfertigen zu können. Es folgen die Hypothesen:

H2a. Wird mithilfe der Skalen eines BIP-Profils eine Prognose darüber getroffen, ob eine Person eine Fach- oder Führungskraft ist, so ist die Trefferquote dieser Prognose größer als die Zufallswahrscheinlichkeit.

H2b. Werden zu einem Modell, das mit den demografischen Variablen Alter, Geschlecht und Einsatzbereich eine Prognose darüber trifft, ob eine Person eine Fach- oder Führungskraft ist, zusätzlich die Skalen des BIP-Profils als Prädiktoren hinzugefügt, so verbessert sich die Trefferquote der Prognose.

Methode

Stichprobe und Berufserfolgskriterien

Der für die Analyse verwendete Datensatz besteht aus n = 20 560 internetgestützt durchgeführten Persönlichkeitstestungen in der aktuellen dritten Revision des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP). Die Testungen wurden zum Zwecke praktisch relevanter Fragestellungen im Rahmen von z. B. Personalauswahl oder Personalentwicklung erhoben. Weiterhin sind in dem Datensatz freiwillige Selbstauskünfte der Teilnehmenden zu demografischen Variablen und beruflichen Erfolgskriterien wie der hierarchischen Position und dem Entgelt enthalten.

Die Spannweite des Alters reicht in der Gesamtstichprobe von *Min* = 21 bis *Max* = 65 mit einem Mittelwert von M = 36.91 bei einer Standardabweichung von SD = 9.04. Die Stichprobe besteht zu 64 % aus Männern und zu 36 % aus Frauen. Die meisten Beschäftigten sind im Bereich des Verkaufs, des Vertriebs und des Kundenservice tätig (3 155 Nennungen), gefolgt vom Personalwesen, der Weiterbildung und Beratung (2 365 Nennungen) sowie Management, Strategie und Führung (1 287 Nennungen). Mehrheitlich wurde das BIP in der Personalentwicklung (3 500 Nennungen) sowie der Personalauswahl (2 649 Nennungen) eingesetzt

Die Variable des Entgeltes liegt in fünf Kategorien vor, wobei Jahresbruttoentgelte von jeweils 40 000 €, 50 000 €, 60 000 € und 80 000 € die Kategoriengrenzen bilden. Teilnehmende ohne Angabe des Entgelts wurden entfernt, wodurch die Analysestichprobe für Hypothese 1 von 20 560 auf n = 10 638 Fälle reduziert wurde.

Für die Hypothesen 2a und 2b wurde die Position als dichotome Variable (Fach- oder Führungskraft) benötigt. Alle Personen, die die Frage nach Führungsverantwortung verneinten und sich selbst als „Sachbearbeiter_in bzw. Fachkraft“ einstuften, erhielten die Ausprägung einer Fachkraft. Alle Personen, die eine andere Positionsstufe (von „Gruppenleiter_in“ bis „Vorstand“) angaben und die Frage nach der Führungsaufgabe bejahten, wurden als „Führungskraft“ eingestuft. Alle anderen Fälle wurden bei der zweiten Hypothese nicht beachtet. Hierdurch verringerte sich die zweite Analysestichprobe von 20 560 auf 8 267 Fälle.

Für die Untersuchung der Hypothese 2b ist der Bereich relevant, in dem die Person arbeitet. Daher konnten hier nur Fälle betrachtet werden, die einen aus 12 Arbeitsbereichen (z. B. „Controlling/Rechnungswesen“ oder „Steuern/Recht“) auswählten. Hierdurch reduzierte sich die Analysestichprobe für die Hypothesen 2a und 2b auf 6 920 Fälle, bestehend zu 42.6 % aus Fach- und zu 57.4 % aus Führungskräften.

Der zentrale Kennwert der zweiten Hypothese ist die Klassifikationsgenauigkeit, mit der BIP-Profile der korrekten Position zugeordnet werden können. Die Ergebnisse lassen sich besonders leicht interpretieren, wenn die dichotomen Gruppen gleich groß sind und die Klassifikationsgüte gegen die Zufallswahrscheinlichkeit von μ_0 = 50 % verglichen werden kann. Die Technik des sog. Downsampling (Kuhn, 2019) wurde angewendet, die hierfür aus der Analysestichprobe zufällig 1 022 Füh-

rungskräfte entfernte, sodass die letztlich für die Klassifikation verwendete Stichprobe von n = 5898 zu jeweils 50 % (entsprechend je 2949 Fällen) aus Fach- und Führungskräften besteht.

Instrumente

In dieser Studie kam die aktuelle dritte Revision des Fragebogens „Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung“ (BIP; Hossiep & Paschen, 2019) zum Einsatz. Das BIP ist ein differenzierter berufsbezogener Persönlichkeitsfragebogen, der vorwiegend im Rahmen der Personalarbeit von Organisationen zu Fragestellungen der Personalentwicklung und Personalauswahl eingesetzt wird. Die hier verwendete dritte Revision des BIP (Hossiep & Paschen, 2019) erfasst 14 Persönlichkeitsskalen mithilfe von 12 bis 16 Items pro Skala. In der Instruktion werden die Teilnehmer_innen darauf hingewiesen, dass sich alle Aussagen bzw. Items auf das Berufsleben beziehen. Beispiele für Skalen sind Leistungsmotivation, Gewissenhaftigkeit, Sensitivität und emotionale Stabilität. Ein Beispiel-Item (für die Skala Leistungsmotivation) lautet „Ich bin mit mir erst dann zufrieden, wenn ich außergewöhnliche Leistungen vollbringe“. Die Teilnehmer_innen beantworten die Items auf einer sechsstufigen Likert-Skala von „trifft voll zu“ bis „trifft überhaupt nicht zu“.

Eingesetzte Clusteranalysetechniken

Das Ziel der Clusteranalyse besteht in der Identifikation von Gruppen, die jeweils in sich selbst homogen in Bezug auf die relevanten Merkmale sind, sich untereinander jedoch in möglichst starkem Maße unterscheiden (Backhaus, Erichson, Plinke & Weiber, 2018). Zur Clusteridentifikation wurde beim BIP bereits die k-Means-Clusteranalyse angewandt (Frieg, 2012), bei der die gewünschte Clusteranzahl (k) im Vorfeld anzugeben ist. Daraufhin werden die beiden Schritte der Identifikation der Clusterzentren und der Zuordnung der Datenpunkte zu diesen so lange wiederholt, bis die Clusterlösung konvergiert (Jain, 2010).

Bereits Frieg (2012) sowie Frieg und Schulz (2014) konnten an unterschiedlichen Stichproben und mit unterschiedlichen Versionen des BIP einander ähnliche, inhaltlich plausible Lösungen mit fünf Clustern identifizieren. Es konnten bereits Zusammenhänge zu den externen Kriterien Alter (Frieg, 2012) sowie aktueller Beschäftigungsstatus (Frieg & Schulz, 2014) gefunden werden. Daher verwendet auch die vorliegende Arbeit die Clusteranzahl von k = 5, zumal auch verschiedene berechnete Indizes zur Bestimmung der Clusteranzahl (z. B. die Statistik nach Hartigan, 1975) diese Clusteranzahl gegenüber k = 4 oder k = 6 Clustern empfehlen. Ebenso wie Gerlach et al. (2018), die vier in unabhängigen Datensätzen stabile Cluster fanden, versucht auch die vorliegende Arbeit, die Clusterlösung von Frieg (2012) mit der dortigen 5-Means-Clusteranalyse in einem anderen Datensatz zu replizieren.

In der vorliegenden Arbeit soll der Clusterlösung das objektive und ordinal skalierte Berufserfolgskriterium des Entgeltes mit fünf Ausprägungen gegenübergestellt werden. Es ist der H-Test von Kruskal und Wallis (1952) zu verwenden, um Unterschiede in der interessierenden Variablen zwischen mehr als zwei Gruppen zu finden (Chan & Walmsley, 1997). Mit den Rängen der Beobachtungen in der Kriteriumsvariablen kann eine H-Statistik berechnet werden, deren Verteilung der Chi-Quadrat-Verteilung ähnelt und daher auf Signifikanz überprüft werden kann (Chan & Walmsley, 1997; Kruskal & Wallis, 1952). Erreicht der Kruskal-Wallis-H-Test Signifikanz, so kann mit einem nonparametrischen post-hoc-Test geprüft werden, welche Gruppenpaare für die signifikanten Unterschiede verantwortlich sind. Es wird die Alphafehlerkorrektur nach Sidak verwendet, welche die Einhaltung des Alphaniveaus sicherstellt (Ruxton & Beauchamp, 2008).

Zur Berechnung der Effektstärke bietet sich die Verwendung des A von Vargha und Delaney (2000) an. Sie bezeichnet die Wahrscheinlichkeit, mit der ein Wert aus einer Gruppe höher ausgeprägt sein wird als ein Wert der anderen Gruppe (Mangiafico, 2016). Die Effektstärke A beträgt bei einem kleinen Effekt $0.56 \leq A < 0.64$, bei einem mittleren Effekt $0.64 \leq A < 0.71$ und bei einem großen Effekt $0.71 \leq A$ (Mangiafico, 2016). Diese zentrale Effektstärke wird verdeutlichen, wie gut sich beruflicher Erfolg, gemessen am Kriterium des Entgeltes, mit Persönlichkeit vorhersagen lässt.

Klassifikationsanalysen

Für binäre Klassifikationsaufgaben existieren verschiedene Algorithmen, die von Lim, Loh und Shin (2000) in drei Gruppen eingeteilt werden: statistische Verfahren, neuronale Netze sowie Verfahren auf Basis von Entscheidungsbäumen. Für die vorliegende Analyse wurden das statistische Verfahren der logistischen Regression sowie das entscheidungsbaumbasierte Verfahren der *Random Forests* gewählt. Es finden sich wissenschaftliche Studien aus verschiedensten Fachbereichen, die die Klassifikationsgüte der logistischen Regression mit Random Forests vergleichen, um Hinweise dafür zu finden, in welchen Kontexten klassische statistische Verfahren bzw. Verfahren des Machine Learning zu bevorzugen sind, sodass

neben den fachlichen Studienergebnissen auch das Wissen über die Methoden der Datenanalyse wächst. Beispiele stammen aus der Politologie (Bürgerkriegsvorhersagen; Muchlinski, Siroky, He & Kocher, 2016), den Geowissenschaften (Erdrutschvorhersage; Chen, Sun & Han, 2019) oder der Medizin (Krankheitsvorhersage; Nusinovici et al., 2020). Eines von wenigen Beispielen zu Themen der Arbeits- und Organisationspsychologie stammt von Gao, Wen und Zhang (2019) und befasst sich mit der Vorhersage von Mitarbeitendenfluktuation. Mutmaßlich aufgrund häufig kleiner Datensätze bleibt die Verwendung von Machine Learning in dieser Fachdisziplin hinter ihrem Potential zurück. Die vorliegende Studie verwendet daher beide Verfahren parallel, um die Literatur im Bereich des Data Mining um ein Beispiel aus der Arbeits- und Organisationspsychologie zu ergänzen.

Für Hypothese 2a werden die metrischen und standardisierten Skalen des BIP als Prädiktoren herangezogen, während für Hypothese 2b zusätzlich die Kontrollvariablen Alter (metrisch und normalisiert), Geschlecht (dichotom nominal) und Bereich (nominal/dummykodiert) verwendet werden.

Logistische Regression

Die binäre logistische Regression liefert Wahrscheinlichkeiten der Zuordnung zu einer der beiden Gruppen mit einem logistischen anstelle eines linearen Regressionsmodells (Backhaus et al., 2018).

Die Regressionskoeffizienten werden hierbei mit der Maximum-Likelihood-Methode so geschätzt, dass die tatsächlichen Daten zu dem Modell maximale Plausibilität erlangen (Backhaus et al., 2018). An die Stichprobe wird die Anforderung gestellt, mindestens das Zehnfache der geschätzten Parameter zu betragen und in jeder Gruppe mit mindestens 25 Fällen vertreten zu sein (Backhaus et al., 2018). Die Variablenanzahl J wird vorliegend maximal 27 (13 demografische Variablen bei Dummykodierung der 12 Arbeitsbereiche und 14 BIP-Skalen) betragen. Daher sind die Empfehlungen mit $n = 5\,898 > J * 10$ bei $J \leq 27$ und $n_1 = n_2 = 2\,949 > 25$ erfüllt.

Random Forests

Grundlage der Random Forests sind Klassifikationsbäume, die auf Basis kategorialer oder metrischer Prädiktorvariablen eine binäre Klassenvariable vorhersagen können (Loh, 2011). Hierzu partitionieren sie dem Autor zufolge den Datensatz dahingehend, dass alle Fälle, die einen Trennwert in metrischen Variablen überschreiten oder bestimmte Ausprägungen in kategorialen Variablen aufweisen, einer Klasse zugeordnet werden und alle anderen Fälle der Alternativklasse. Diese Partitionierung des Datensatzes wird schrittweise vorgenommen, und die Regeln, nach denen dies geschieht, werden durch eine Baumstruktur visualisiert, die aufgrund ihrer leichten Interpretierbarkeit Beliebtheit in der Anwendung erlangt hat (Loh, 2014).

Je mehr Schritte durchgeführt werden, desto wahrscheinlicher wird ein Phänomen namens *Overfitting:* Die Trefferquote im Trainingsdatensatz steigt, doch unabhängige Daten können mit dem Modell nicht mehr vorhergesagt werden (Witten, Frank, Hall & Pal, 2017), da der Klassifikationsbaum spezielle Charakteristika des Datensatzes abbildet, mit dem er erzeugt wurde. Für eine Verbesserung der Klassifikationsgüte bei fremden Daten existieren Random Forests (Breiman, 2001), die viele verschiedene einzelne Bäume entwerfen (sog. *Ensemble Learning*). Random Forests lassen die gebildeten Klassifikationsmodelle über die Klassifikation „abstimmen" und klassifizieren mithilfe der „Mehrheitsentscheidung" (Han et al., 2012). Sie verwenden in jeder Durchführung einen zuvor zufällig gewählten Teil der Variablen, hiervon jedoch stets alle (Han et al., 2012).

Vom Anwender sind für Random Forests lediglich zwei Parameter im Vorfeld festzulegen: Die Anzahl der Bäume sowie die Zahl der verwendeten Variablen für jeden einzelnen Baum (Liaw & Wiener, 2002). Boulesteix, Janitza, Kruppa und König (2012) definierten für das R-Paket randomForest eine Anzahl von $n_{Tree} = 500$ als Anzahl der Bäume und gaben für die Anzahl der Variablen pro Baum die Empfehlung, die Wurzel der Anzahl der insgesamt vorliegenden Prädiktorvariablen zu verwenden, also $m = \sqrt{(p)}$. Gleichwohl besitzt das R-Paket caret (Kuhn, 2019) zusätzlich die Fähigkeit des *Parameter Tuning*, d. h. es kann automatisiert mehrere Zahlen für die Variablenanzahl testen und letztlich die optimale Anzahl, die mit optimierter Klassifikationsgüte einhergeht, auswählen.

Kreuzvalidierung zur Ergebnisabsicherung

Schätzt man die Klassifikationsgüte am selben Datensatz, an dem das statistische Modell gebildet wird, so kann diese Schätzung zu optimistisch sein, wenn das Modell Besonderheiten des Datensatzes abbildet, an dem es gebildet wurde (Witten et al., 2017).

Als Goldstandard zur Evaluation von Klassifikationsmodellen empfehlen Witten et al. (2017) daher die zehnmalig wiederholte zehnfache Kreuzvalidierung. Der Datensatz ist bei der zehnfachen Kreuzvalidierung in zehn etwa gleich große Teile partitioniert, von denen stets neun gemeinsam das Modell generieren und das jeweils übrige Zehntel als Testdatensatz zur Evaluation der Modellgüte genutzt wird. Um die Schätzung der Güte weiter zu verbessern, wird diese Prozedur nicht einmal, sondern insgesamt zehnmal wiederholt (Witten et al., 2017). Damit kann letztlich die Klassifikationsgüte von hundert Modellen gemittelt werden, wodurch die Güte der Vorhersage steigt.

Zur Evaluation der Prognostizität des BIP wird die bereits dargestellte Technik des Downsampling verwendet. Da hierdurch die Trefferquote der beiden statistischen Ansätze gegen die Zufallswahrscheinlichkeit von $\mu_0 = 50\,\%$ verglichen werden kann, können die standardisierten Effektgrößenmaße von Cohen (1992) verwendet werden. Differenzen zwischen Anteilswerten (vorliegend der Anteil korrekter Klassifikation) und $P = .50$ von $g = .05$ spiegeln einen kleinen, von $g = .15$ einen mittleren und von $g = .25$ einen großen Effekt wider (Cohen, 1992; Döring & Bortz, 2016). Für Hypothese 2a wird der höchste Wert der beiden Klassifikationsmodelle ausgewählt, die mithilfe der 14 Skalen des BIP die Position vorhersagen, und mit einem Einstichproben-t-Test gegen $\mu_0 = .50$ verglichen. Eine Klassifikationsgüte von $P = 55\,\%$ wird folglich als kleine, von $P = 65\,\%$ als mittlere und von $P = 75\,\%$ als große Verbesserung der Vorhersage durch Einbezug der Skalen des BIP verstanden.

Für Hypothese 2b werden zwei Modelle pro statistischem Ansatz aufgestellt: ein Modell, das lediglich die Variablen des Alters, des Geschlechts und des Arbeitsbereiches verwendet, sowie ein kombiniertes Modell, das diese Variablen um die BIP-Skalen ergänzt. Das Modell mit der höchsten Klassifikationsgüte im kombinierten Modell wird gegen die Trefferquote des entsprechenden rein demografischen Modells per Zweistichproben-t-Test verglichen.

Ergebnisse

Skalenanalyse

Tabelle 1 präsentiert die internen Konsistenzen (Cronbachs α) der Skalen. Ebenso werden die Interkorrelationen der Skalen sowie die Korrelationen zwischen den Skalen bzw. demografischen Kriterien mit den beiden Berufserfolgskriterien in Tabelle 1 dargestellt. Der höchste Zusammenhang der Skalen zu den Erfolgskriterien liegt zwischen Führungsmotivation und Position vor und beträgt $r_{pb} = .40$.

Persönlichkeitscluster und Entgelt

Die Clusterzentren in den z-standardisierten Persönlichkeitsmerkmalen der fünf identifizierten Cluster sind in Abbildung 1 dargestellt. Kennzeichnend für das erste Cluster sind eine hohe Gestaltungsmotivation sowie eine ausgeprägte Flexibilität, weswegen es als „gestaltend-flexibel“ bezeichnet wird. Aufgrund der jeweils niedrigen Ausprägungen in den Skalen Sozialibilität und emotionale Stabilität wird das zweite Cluster als „streitbar-instabil“ ausgewiesen. Im dritten Cluster („unsicher-nachgiebig“) sind Führungsmotivation und Selbstbewusstsein gering ausgeprägt. Das vierte Cluster liegt fast vollständig im Intervall einer halben Standardabweichung um den Mittelwert. Eine Ausnahme ist die erhöhte Sozialibilität; demnach wird dieses Cluster als „soziabel-unauffällig“ bezeichnet. Ein durchgängig positives Selbstbild bei deutlich ausgeprägtem Selbstbewusstsein ist charakteristisch für das fünfte Cluster, das demnach mit „unkritisch-selbstbewusst“ benannt wird.

Abbildung 2 visualisiert die Clusteranteile in jeder Entgeltgruppe. Die fünf Cluster unterscheiden sich in Bezug auf die Verteilung des Entgelts, $\chi^2(4) = 755.85$, $p < .001$. Alle Dunn-Sidak-post-hoc-Tests erreichen Signifikanz; die z-Werte und Effektstärken für das A nach Vargha und Delaney sind Tabelle 2 zu entnehmen. Das Cluster „gestaltend-flexibel“ (1) erzielt in allen paarweisen Vergleichen stets die höchsten Entgeltwerte, gefolgt von den Clustern „unkritisch-selbstbewusst“ (5), „streitbar-instabil“ (2) und „soziabel-unauffällig“ (4). Das Cluster „unsicher-nachgiebig“ (3) erzielt in jedem Falle die geringsten Werte. Die größte Effektstärke liegt zwischen den Clustern 1 und 3 und beträgt $A = .709$.

Klassifikationsgüte

Tabelle 3 gibt als Trefferquoten die gemittelten Werte aller Wiederholungen der Kreuzvalidierung an. Für die Random Forests wurden die Werte desjenigen Modells genannt, das als Resultat des *Parameter Tunings* die beste Trefferquote ausgab. Es wurde für das Modell der BIP-Skalen als ideale Anzahl der Variablen pro Baum 7 (von 14), für das demografische Modell 3 (von 3) und für das kombinierte Modell 4 (von 17) ausgewählt.

Die höchste Güte bei Verwendung der 14 Skalen des BIP als Prädiktorvariablen und der dichotomen Prognosevariablen der Position (Fach- oder Führungskraft) erreicht die logistische Regression mit 69.3 %. Unter Verwendung der drei Variablen Alter, Geschlecht und Arbeitsbereich erzielen die Random Forests die beste Klassifikationsgüte von 70.0 %. Durch die Hinzunahme der 14 Skalen des BIP kann insbesondere bei der logistischen Regression eine erhebliche Verbesserung um 7.2 % auf 74.7 % festgestellt werden, wobei die Random Forests weiterhin mit 75.4 % die beste Klassifikationsgüte im kombinierten Modell von Demografie und Skalen erzielen.

Zur Testung von Hypothese 2a wird das Modell der logistischen Regression mit den 14 Skalen als Prädiktoren ausgewählt. Die Trefferquote von $M = 69.3\,\%$ (gemittelt aus 100 einzelnen Trefferquoten, $SD = 1.7\,\%$) ist größer als die Zufallswahrscheinlichkeit von $\mu_0 = 50\,\%$, $t(99) = 112.96$, $p < .001$.

Tabelle 1. Cronbachs Alpha und Korrelationen

Variablen	α	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 Alter		–	−.14**	.32**	.31**	−.07**	.14**	.10**	−.12**	.11**	.04**	.01	.00	−.05**	.06**	.07**	.05**	.04**	.11**
2 Geschlecht			–	−.21**	−.26**	−.16**	−.15**	−.22**	.04**	−.10**	−.04**	.06**	.02**	.07**	−.10**	−.14**	−.24**	−.18**	−.25**
3 Position				–	.40**	.13**	.24**	.40**	−.11**	.17**	.04**	.03**	.07**	−.13**	.18**	.25**	.11**	.13**	.20**
4 Entgelt					–	.10**	.17**	.22**	−.12**	.17**	.05**	.02**	.05**	−.09**	.12**	.17**	.11**	.11**	.17**
5 Leistungsmotivation (LM)	.79					–	.49**	.50**	.31**	.43**	.40**	.27**	.33**	−.04**	.10**	.49**	.29**	.55**	.41**
6 Gestaltungsmotivation (GM)	.71						–	.56**	−.06**	.49**	.20**	.12**	.28**	−.34**	.07**	.60**	.20**	.29**	.46**
7 Führungsmotivation (FM)	.88							–	.02**	.53**	.40**	.36**	.52**	−.13**	.31**	.74**	.44**	.50**	.67**
8 Gewissenhaftigkeit (Ge)	.85								–	−.17**	.47**	.20**	.10**	.26**	.02**	.09**	.09**	.28**	.07**
9 Flexibilität (Fl)	.85									–	.38**	.32**	.42**	.01	.30**	.45**	.48**	.55**	.53**
10 Handlungsorientierung (HO)	.87										–	.48**	.43**	.28**	.28**	.41**	.54**	.67**	.50**
11 Sensitivität (Sen)	.83											–	.58**	.47**	.28**	.33**	.40**	.47**	.42**
12 Kontaktfähigkeit (Ko)	.90												–	.23**	.37**	.48**	.43**	.43**	.58**
13 Soziabilität (Soz)	.77													–	.31**	−.25**	.23**	.26**	−.05**
14 Teamorientierung (TO)	.87														–	.16**	.26**	.27**	.22**
15 Durchsetzungsstärke (Du)	.82															–	.43**	.46**	.67**
16 Emotionale Stabilität (ESt)	.88																–	.71**	.68**
17 Belastbarkeit (Bel)	.90																	–	.59**
18 Selbstbewusstsein (SB)	.85																		–

Anmerkungen: Korrelationen zum Entgelt wurden als τ_B, weitere Korrelationen zum Geschlecht und zur Position als r_{pb} berechnet, positive Werte stehen hier für eine höhere Ausprägung bei Frauen bzw. Führungskräften, negative für eine höhere Ausprägung bei Männern bzw. Fachkräften. Alle weiteren Korrelationen wurden als Pearsons *r* berechnet.
* $p < .05$, ** $p < .01$; $n = 20\,560$, außer für die Korrelation zwischen Entgelt und Position $n = 7\,791$, alle weiteren mit Entgelt $n = 10\,638$ sowie mit Position $n = 8\,267$.

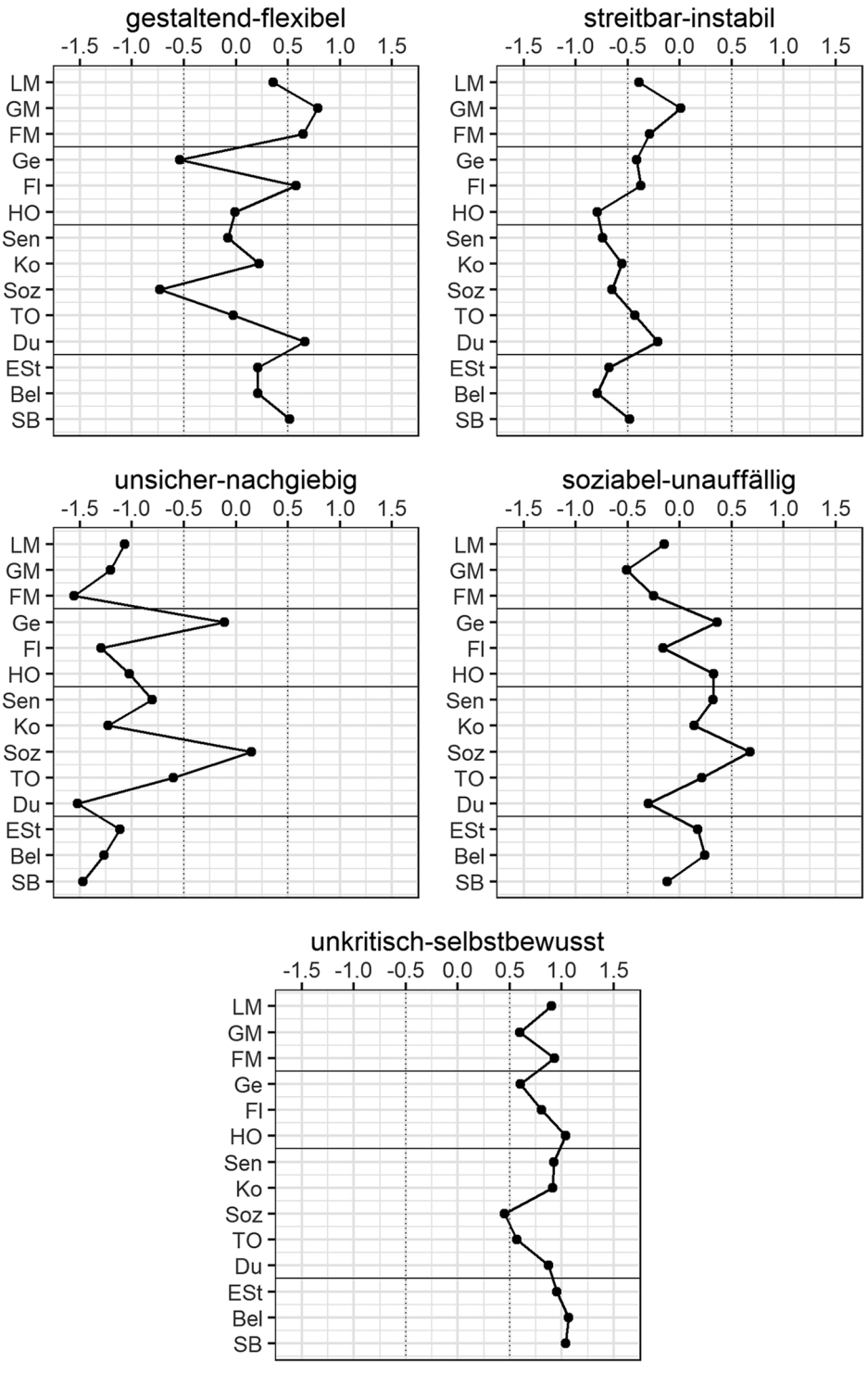

Abbildung 1. Zentren der fünf Cluster

Zur Testung der Hypothese 2b werden die Modelle der Random Forests verwendet. Dabei werden einerseits nur die demografischen Variablen und andererseits die demografischen Variablen sowie die BIP-Skalen berücksichtigt und miteinander verglichen. Die Trefferquote des kombinierten Modells (M = 75.4 %, SD = 1.7 %) liegt im Vergleich zur Trefferquote des rein demografischen Modells (M = 70.0 %, SD = 1.8 %) höher, $t(197.63) = 22.06$, $p < .001$.

Tabelle 2. Unterschiede der Cluster im Entgelt

	z	*A*	A_{max}		*z*	*A*	A_{max}
1 – 2	14.66	.62	.62	2 – 4	5.38	.54	.54
1 – 3	21.78	.71	.71	2 – 5	-9.90	.41	.59
1 – 4	20.60	.66	.66	3 – 4	-5.24	.45	.55
1 – 5	3.71	.54	.54	3 – 5	-17.48	.32	.68
2 – 3	9.47	.59	.59	4 – 5	-15.20	.37	.63

Anmerkungen: Bei negativen z-Werten erhält das zweitgenannte Cluster ein höheres Entgelt. A_{max} gibt dort die Effektgröße des umgekehrten Vergleichs an. Für den Clustervergleich 1 – 5 p = .002, ansonsten alle $p < .001$.

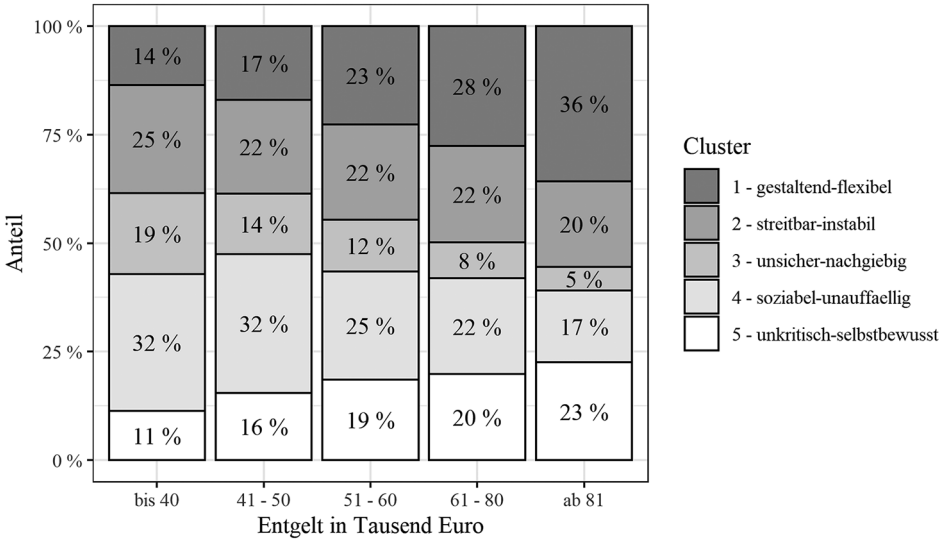

Abbildung 2. Clusteranteile nach Entgeltgruppe

Tabelle 3. Durchschnittliche Trefferquote der Klassifikationsmodelle

	Logistische Regression	Random Forests
Demografie	67.5 %	70.0 %
BIP-Skalen	69.3 %	68.6 %
Demografie + Skalen des BIP	74.7 % (+7.2 %)	75.4 % (+5.4 %)

Diskussion

Die vorliegende Untersuchung analysierte einen großen Datensatz realer Testungen des BIP, um zusätzliche Validitätsnachweise für das Verfahren zu finden. Sie generierte zwei zentrale Erkenntnisse:

1. Personen mit einem BIP-Profil, das von Flexibilität und Gestaltungsmotivation bei gleichzeitig relativ geringen Ausprägungen in Soziabilität und Gewissenhaftigkeit geprägt ist, erhalten mit einer Wahrscheinlichkeit von 70.9 % ein höheres Entgelt als solche, die sich als emotional instabil und wenig durchsetzungsstark beschreiben. Die Effektstärke (das *A* von Vargha & Delaney, 2000) zwischen diesen beiden Clustern ist von mittlerer bis hoher Größe (Mangiafico, 2016).
2. Gegenüber der Zufallswahrscheinlichkeit von 50 % können die Skalen des BIP 69.3 % der Personen korrekt als Fach- oder Führungskraft identifizieren. Die Verbesserung der Klassifikationsgüte entspricht einem mittleren bis hohen Effekt (Cohen, 1992). Weiterhin ist durch die Hinzunahme der BIP-Skalen zu den Variablen Alter, Geschlecht und Arbeitsbereich eine signifikante Verbesserung der Trefferquote um bis zu 7.2 % festzustellen.

Implikationen für die Forschung

Insbesondere in der verwendeten Methodik kann eine Chance liegen, die Qualität psychologischer Forschung zu erhöhen, damit zukünftig nicht mehr ein Großteil der psychologischen Forschung in der Replikation scheitert (Open Science Collaboration, 2015).

Bei der für die Klassifikationsanalysen verwendeten großen Stichprobe (n = 5 898) ist die Technik der wiederholten Kreuzvalidierung geeignet, um akkurate Schätzungen der Klassifikationsgüte hervorzubringen (Han et

al., 2012), da sie Variablen an neuen, dem Modell unbekannten Fällen vorhersagt. Dabei werden die generierten Modelle durch Stichprobenteilungen sozusagen in sich selbst repliziert. Die Güte des vorliegenden Ergebnisses zeigt sich ebenfalls darin, dass es verschiedene bestehende Erkenntnisse inhaltlicher und methodischer Art bestätigen kann. So konnte z. B. die Clusterlösung von Frieg (2012) für das BIP mit minimalen Abweichungen (maximal 0.1 Standardabweichungen Unterschied in den Clusterzentren) repliziert werden. Daher existiert mit der vorliegenden 5-Means-Clusterlösung für das BIP eine über unterschiedliche Stichproben stabile Entsprechung zur Big-Five-Clusterlösung von Gerlach et al. (2018).

Auch methodische Empfehlungen können mit der vorliegenden Arbeit bestätigt werden: Das komplexe Machine-Learning-Verfahren der Random Forests erreichte zwar, wie zu erwarten war, bei dem komplexesten Modell (der Kombination kategorialer und metrischer Prädiktoren) die beste Vorhersagegüte. Beim ausschließlichen Vorliegen der metrischen, standardisierten BIP-Skalen jedoch zeigte sich die logistische Regression überlegen. Dies ist ein Indiz dafür, in welcher Situation welches Klassifikationsverfahren anzuwenden ist: In Fällen mit komplexem Prädiktionsmodell und komplexen Zusammenhängen zur Klassenvariable stiftete das non-lineare Machine-Learning-Verfahren einen Zusatznutzen, in linearen Fällen mit gleichförmiger Prädiktorenstruktur hingegen können die klassischen statistischen Verfahren weiterhin überlegen sein. Damit leistet die vorliegende Untersuchung über die fachlichen Erkenntnisse zur BIP-Validierung hinaus auch eine methodische Erkenntnis: Je komplexer die Variablenstruktur im Datensatz, desto wahrscheinlicher ist ein Zusatznutzen durch Verwendung komplexerer Analyseverfahren.

Zusammenfassend sollte psychologische Forschung sich zur Steigerung der Güte der Erkenntnisse hoher Stichprobengrößen, der Technik der Kreuzvalidierung sowie geeigneter statistischer Methoden bedienen.

Implikationen für die Praxis

Die Ergebnisse unterstreichen, dass sich mit dem BIP Persönlichkeitsdimensionen identifizieren lassen, die mit Berufserfolg assoziiert sind. Die hohe Klassifikationsgenauigkeit und erheblichen Zusammenhänge zum Entgelt sind Nachweise für die Kriteriumsvalidität des BIP. Es liegt im Interesse von Organisationen, Arbeitnehmende zu identifizieren, deren Persönlichkeit ein Spiegel beruflichen Erfolges ist und die in ihren Merkmalsausprägungen zur jeweiligen Funktion in der Organisation passen.

Basis der vorliegenden Arbeit war ein personenbezogen-empirischer Ansatz der Arbeitsanalyse, bei der Persönlichkeitsmerkmale identifiziert wurden, die in der Realität vermehrt bei gut entlohnten Arbeitnehmenden bzw. Führungskräften auftreten. In der Personalarbeit wird gleichwohl zumeist die arbeitsplatzanalytisch-empirische Methode verfolgt werden, bei der zunächst Positionen systematisch z. B. mit dem Anforderungsmodul BIP-AM (Hossiep & Weiß, 2020) bezüglich ihrer Anforderungen eingeschätzt werden und die Passung zum Bewerber oder Stelleninhaber eine Zusatzinformation liefert, um Personalentscheidungen zu optimieren.

Weiterhin unterstreichen die Erkenntnisse der Clusteranalyse, dass in der Personalarbeit der Blick auf einzelne Skalen eines Persönlichkeitstests durch ein Verständnis der Gesamtprofilstruktur zu ergänzen ist. Die Clusterlösung unterstreicht, dass eine niedrige Soziabilität anders zu interpretieren ist, je nachdem, ob sie im streitbar-instabilen Cluster mit Problemen in der psychischen Konstitution oder im gestaltend-flexiblen Cluster mit hoher Durchsetzungsstärke und Gestaltungsmotivation einhergeht. Auch eine niedrige Gestaltungsmotivation kann einerseits im unsicher-nachgiebigen Cluster Spiegel motivationaler Problematiken oder andererseits im soziabel-unauffälligen Cluster ein Element einer unauffälligen „Spezialistenpersönlichkeit“ sein. Weiterhin bleibt das BIP gleichwohl ein dimensionales Verfahren; die identifizierten Cluster dienen zum theoretischen Mehrgewinn, jedoch nicht für eine Typisierung einzelner Teilnehmender, da Typen eine unzulässige Vereinfachung der vielfältigen Persönlichkeitsunterschiede darstellen würden (Kersting, 2013). Wenngleich bei der Aggregation einer großen Anzahl von Einzelstudien metaanalytisch der eindeutige Nachweis gelingt (vgl. Kuncel, Klieger, Connelly & Ones, 2013), dass eine statistische Urteilsbildung mit der höchsten Validität verknüpft ist, möchten wir trotzdem beim Einsatz des BIP keine Zuordnung einzelner Personen zu Typen auf rein statistischer Basis empfehlen, wie dies etwa beim Myers-Briggs-Typenindikator in nicht belastbarer Weise geschieht. Resultate des BIP sind von Personalverantwortlichen mit diagnostischer Expertise differenziert mit Blick auf sowohl die Einzelskalen wie auf die Gesamtstruktur (und letztlich auch mit Blick auf die Beantwortung der Fragen zu einzelnen Skalenfacetten) in den Entscheidungsprozess zu berücksichtigen und einzubringen. Ist eine Interpretation der detaillierten BIP-Skalen nicht möglich, könnte die Alternative anstelle einer Typologie in der Verwendung der sechs breiten Gesamtskalen des faktorenanalytisch entstandenen BIP-6F (Hossiep & Krüger, 2012) bestehen.

Zusammenfassend weist die Arbeit nach, dass die Bedenken von Morgeson et al. (2007) bezüglich einer scheinbar geringen Validität ungerechtfertigt sind. Es überrascht nicht, dass Morgeson et al. (2007) niedrige Validitätskoeffizienten von Persönlichkeitstests monie-

ren, wenn die dort zugrundeliegende Persönlichkeitstheorie kriteriumsrelevante Dimensionen wie die Führungsmotivation nicht beinhaltet. Der vergleichsweise seltene Einsatz von Persönlichkeitsdiagnostik in deutschen Unternehmen (Armoneit et al., 2020) ist daher aus wissenschaftlicher Sicht zu bemängeln.

Weiterhin zeigte sich, dass Persönlichkeitsprofile gegenüber Geschlecht, Alter und Arbeitsbereich inkrementelle Kriteriumsvalidität aufweisen. Ein Grund für die nachweisbare Validität demografischer Merkmale wie Geschlecht oder Alter kann in bewusster oder unbewusster Diskriminierung bestimmter Personengruppen (etwa junger Frauen) ausgemacht werden. Der Einsatz von Persönlichkeitsfragebogen kann somit im Sinne eines organisationskulturellen Wandels zu einer stärkeren Diversität innerhalb der Führungsriegen beitragen, indem weniger äußere Personenmerkmale wie Geschlecht oder Alter bei der Besetzung von Führungspositionen berücksichtigt werden, sondern tatsächlich arbeitsrelevante Dimensionen wie Ausprägungen in Persönlichkeitsmerkmalen.

Limitation und zukünftige Forschung

Eine Limitation der verwendeten Klassifikationsmodelle liegt in ihrer mangelnden inhaltlichen Interpretierbarkeit. Die Interpretation der Koeffizienten der logistischen Regression ist ohnehin bei Konstanthaltung der jeweils 13 übrigen Skalen des BIP schwierig. Die Random Forests erlauben lediglich eine Abschätzung der Wichtigkeit einzelner Variablen.

Wenngleich an vielen Stellen die Qualität der Stichprobe hervorgehoben wurde, ist natürlich auch diese eine selektive Auswahl: Es ist plausibel, dass ein komplexes Verfahren wie das BIP häufiger bei Führungskräften und spezialisierten Fachkräften angewendet wird, während weniger Ressourcen in die Personalentwicklung für Beschäftigte mit vorwiegend operativen bzw. weniger komplexen Tätigkeiten investiert werden. Daher sind Führungskräfte in der Stichprobe überrepräsentiert, was das ausgeglichene Verhältnis überhaupt erst ermöglicht. Folglich gelten die Befunde für die primäre Zielgruppe des BIP (Führungskräfte und spezialisierte Fachkräfte).

Eine weitere Limitation liegt darin, dass die Berufserfolgskriterien konkurrent erhoben wurden. Es wurden Zusammenhänge analysiert, keine Kausalität. Dass Arbeitnehmende deswegen in eine Führungsposition kommen, weil sie hoch führungsmotiviert sind, kann auf Basis der vorliegenden Ergebnisse nicht zwangsläufig gefolgert werden. Daher sollte eine Studie durchgeführt werden, die die prädiktive Validität verschiedener Personalauswahlverfahren bestimmt. Im ersten Schritt sollten Arbeitnehmende ein Verfahren der Eignungsdiagnostik durchlaufen, das neben einem Persönlichkeitsfragebogen auch bspw. einen Intelligenz-, einen Wissens- und einen Interessenstest umfasst. Für Persönlichkeit und Interessen sollten zusätzlich die Anforderungen der Vakanz eingeschätzt werden, um statt der Rohwerte einen Passungskoeffizienten zwischen Stelle und den Ausprägungen der Arbeitnehmenden berechnen zu können. Im zweiten Schritt sollten nach einer gewissen Zeitspanne subjektive und objektive Berufserfolgskriterien sowie Maße für die Arbeitsleistung erhoben werden, um die prädiktive Validität der Verfahren zu bestimmen.

Weiterhin könnte zukünftige Forschung weitere Variablen (z. B. den Bearbeitungskontext) oder nur bestimmte Führungspositionen (z. B. nur Vorstände) einbeziehen, um noch differenziertere Validierungsbefunde zu schaffen. Da die im Auswahlkontext auftretende (Birkeland, Manson, Kisamore, Brannick & Smith, 2006) soziale Erwünschtheit keine Bedrohung der Validität von Persönlichkeitstests darstellt (Ones, Viswesvaran & Reiss, 1996), wurde der Bearbeitungskontext vorliegend nicht gesondert betrachtet und die Robustheit der Ergebnisse angenommen. Zukünftige Forschung könnte dennoch anhand der Substichproben die Auswirkungen des Bearbeitungskontextes spezifisch für das BIP betrachten. Eine weitere Möglichkeit zur Ergänzung der Befunde wäre die hier nicht vorgenommene Betrachtung von Interaktionen der Persönlichkeitsdimensionen mit den demographischen Variablen sowie der Einbezug subjektiver Erfolgskriterien wie Arbeitszufriedenheit und Berufserfolg, für die bereits im Manual des BIP (Hossiep & Paschen, 2019) noch leicht höhere Zusammenhänge als für die vorliegenden objektiven Kriterien gefunden wurden. Weiterhin könnte die Variable des Arbeitsbereichs als Moderatorvariable einbezogenen werden, um eine differentielle Validierung für verschiedene Arbeitsbereiche vorzunehmen, da nach der Theorie des Person-Environment Fit für verschiedene Arbeitsinhalte verschiedene Persönlichkeitsmerkmale relevant werden (wie für die Big Five bereits bei Denissen et al., 2018, nachgewiesen).

Nicht zuletzt sollte zukünftige Forschung auch untersuchen, ob die hier generierten Erkenntnisse basierend auf Wahrscheinlichkeiten und Trefferquoten Personalverantwortliche ergänzend zu klassischen Maßen z. B. der Varianzaufklärung noch stärker von der wissenschaftlichen Qualität von Persönlichkeitsverfahren überzeugen.

Schlussfolgerung

In der vorliegenden Arbeit konnten durch die Anwendung von Cluster- und Klassifikationsanalysen Validitätsnachweise von mittlerer bis großer Effektstärke für das BIP

bezüglich objektiver Berufserfolgskriterien aufgezeigt werden. Dies stützt die Aussage, dass das BIP ein valides Verfahren der Persönlichkeitsdiagnostik ist und daher von Organisationen im Rahmen wissenschaftlich fundierter Personalarbeit eingesetzt werden kann.

Ein besonderes Gütemerkmal der Studie liegt in der Kreuzvalidierung in Verbindung mit einer großen Stichprobe. Daher konnte die Studie nicht nur ihre Ergebnisse „in sich selbst" bestätigen, sondern weiterhin verschiedene inhaltliche und methodische Erkenntnisse vorheriger Forschung bekräftigen. Die verwendete Methodik kann daher auch bei zukünftiger Forschung im Feld der Arbeits- und Organisationspsychologie verwendet werden, um die Güte von Ergebnissen zu erhöhen.

Gleichwohl ist angesichts der Theorie des Person-Job-Fit plausibel, dass die Zusammenhänge höher ausfallen, wenn die konkrete Passung zur jeweils ausgeübten oder angestrebten Stelle den Berufserfolgskriterien gegenübergestellt wird. Hohe Korrelationen des selbsteingeschätzten Person-Job-Fit mit den Kriterien Arbeitszufriedenheit und Berufserfolg (Weiß, Krumscheid & Frieg, 2014) stützen diese Vermutung. Eine Erhebung von Erfolgs- und Leistungskriterien zu mehreren Zeitpunkten im Anschluss an die psychologische Diagnostik, wie bereits bei Hossiep (1995) durchgeführt, könnte in Kombination mit den vorliegend verwendeten Methoden einen hochwertigen empirischen Beweis für die Wichtigkeit berufsbezogener Persönlichkeitsdiagnostik erbringen.

Literatur

Armoneit, C., Schuler, H., & Hell, B. (2020). Nutzung, Validität, Praktikabilität und Akzeptanz psychologischer Personalauswahlverfahren in Deutschland 1985, 1993, 2007, 2020. *Zeitschrift für Arbeits- und Organisationspsychologie, 64*, 67 – 82.

Backhaus, K., Erichson, B., Plinke, W. & Weiber, R. (2018). *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung* (15. Aufl.). Berlin: Springer.

Barrick, M. R., Mount, M. K., & Judge, T. A. (2001). Personality and performance at the beginning of the new millennium: What do we know and where do we go next? *International Journal of Selection and Assessment*, *9*(1-2), 9 – 30.

Birkeland, S. A., Manson, T. M., Kisamore, J. L., Brannick, M. T. & Smith, M. A. (2006). A meta-analytic investigation of job applicant faking on personality measures. *International Journal of Selection and Assessment*, *14*, 317 – 335.

Bortz, J. & Schuster, C. (2010). *Statistik für Human- und Sozialwissenschaftler* (7. Aufl.). Berlin: Springer.

Boulesteix, A.-L., Janitza, S., Kruppa, J. & König, I. R. (2012). Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 493 – 507.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5 – 32.

Chan, Y. & Walmsley, R. P. (1997). Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical Therapy*, *77*(12), 1755 – 1761.

Chen, W., Sun, Z. & Han, J. (2019). Landslide susceptibility modeling using integrated ensemble weights of evidence with logistic regression and random forest models. *Applied Sciences*, *9*, 1 – 26.

Cohen, J. (1992). A power primer. Psychological Bulletin, 112, 155 – 159.

Denissen, J. J. A., Bleidorn, W., Hennecke, M., Luhmann, M., Orth, U., Specht, J. & Zimmermann, J. (2018). Uncovering the power of personality to shape income. *Psychological Science*, *29*(1), 3 – 13.

Döring, N. & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Berlin: Springer.

Frieg, P. (2012). *Alterseffekte bei berufsbezogenen Persönlichkeitsmerkmalen – Fakt oder Artefakt?* Dissertation, Ruhr-Universität Bochum.

Frieg, P. & Schulz, R. (2014). *Hartz-IV-Empfänger nicht „faul" – Eine Studie zur berufsbezogenen Persönlichkeit von Arbeit Suchenden und Berufstätigen* (Forschungsbericht). Bochum: Ruhr-Universität, Projektteam Testentwicklung.

Gao, X., Wen, J., & Zhang, C. (2019). An improved random forest algorithm for predicting employee turnover. *Mathematical Problems in Engineering*, *2019*, 1 – 12.

Gerlach, M., Farb, B., Revelle, W. & Amaral, L. A. N. (2018). A robust data-driven approach identifies four personality types across four large data sets. *Nature Human Behaviour*, *2*, 735 – 742.

Han, J., Kamber, M. & Pei, J. (2012). *Data mining: concepts and techniques.* Waltham: Elsevier.

Hartigan, J. A. (1975). *Clustering algorithms.* New York: Wiley.

Hossiep, R. (1995). *Berufseignungsdiagnostische Entscheidungen. Zur Bewährung eignungsdiagnostischer Ansätze.* Göttingen: Hogrefe.

Hossiep, R. & Krüger, C. (2012*). Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – 6 Faktoren (BIP-6F).* Göttingen: Hogrefe.

Hossiep, R., Mutwill, A. & Schulz, R. (2012). "Das Geheimnis meines Erfolgs..." – Berufserfolgsattributionen im Top-Management. In Arbeitskreis Assessment Center e.V. (Hrsg.), *Talent(e) entdecken und fördern: Personaldiagnostik als Wettbewerbsvorteil. Dokumentation zum 8. Deutschen Assessment-Center-Kongress 2012* (S. 426 – 444). Lengerich: Pabst.

Hossiep, R. & Paschen, M. (2019). *Bochumer Inventar zur berufsbezogenen Persönlichkeitsbeschreibung – BIP* (3. Aufl.). Göttingen: Hogrefe.

Hossiep, R., Schecke, J. & Weiß, S. (2015). Zum Einsatz von persönlichkeitsorientierten Fragebogen. Eine Erhebung unter den 580 größten deutschen Unternehmen. *Psychologische Rundschau, 66*, 127 – 129.

Hossiep, R. & Weiß, S. (2020). *Das Anforderungsmodul zum BIP (BIP-AM).* Göttingen: Hogrefe.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651 – 666.

Kanning, U. P. (2017). Fairness und Akzeptanz von Personalauswahlmethoden. In D. E. Krause (Hrsg.), *Personalauswahl* (S. 271 – 299). Wiesbaden: Springer Gabler.

Kauffeld, S. & Grohmann, A. (2019). Personalauswahl. In S. Kauffeld (Hrsg.), *Arbeits-, Organisations- und Personalpsychologie für Bachelor* (S. 139 – 165). Berlin: Springer.

Kersting, M. (2013). Persönlichkeit ist keine Typfrage. Grundlagen zu Persönlichkeitsfragebogen. *Personalmagazin, 12/13*, 26 – 29.

König, C. J. & Marcus, B. (2013). TBS-TK Rezension: „Persolog Persönlichkeits-Profil." *Psychologische Rundschau, 64*, 189–191.

König, C. J., Klehe, U. C., Berchtold, M., & Kleinmann, M. (2010). Reasons for being selective when choosing personnel selection procedures. *International Journal of Selection and Assessment*, *18*(1), 17 – 27.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, *47*, 583 – 621.

Kuhn, M. (2019). *The caret Package.* Verfügbar unter: http://topepo.github.io/caret/index.html

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*, 1060-1072.

Lang, J. W., & Kell, H. J. (2020). General mental ability and specific abilities: Their relative importance for extrinsic career success. *Journal of Applied Psychology, 105*, 1047 – 1061.

Liaw, A. & Wiener, M. (2002). Classification and regression by randomForest. *R News*, *2/3*, 18 – 22.

Lim, T.-S., Loh, W.-Y. & Shin, Y.-S. (2000). A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*, *40*(3), 203 – 228.

Loh, W.-Y. (2011). Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(1), 14 – 23.

Loh, W. Y. (2014). Fifty years of classification and regression Trees. *International Statistical Review*, *82*(3), 329 – 348.

Mangiafico, S. S. (2016). *Summary and analysis of extension program evaluation in R.* Verfügbar unter: http://rcompanion.org/documents/RHandbookProgramEvaluation.pdf

Marcus, B. (2004). Rezension der 2. Auflage des Bochumer Inventars zur berufsbezogenen Persönlichkeitsbeschreibung (BIP) von R. Hossiep und M. Paschen. *Zeitschrift für Arbeits- und Organisationspsychologie A&O*, *48*, 79 – 86.

Moosbrugger, H. & Kelava, A. (2020). *Testtheorie und Fragebogenkonstruktion* (3. Aufl.). Berlin: Springer.

Morgeson, F. P., Campion, M. A., Dipboye, R. L., Hollenbeck, J. R., Murphy, K. & Schmitt, N. (2007). Reconsidering the use of personality tests in personnel selection contexts. *Personnel Psychology*, *60*, 683 – 729.

Mount, M. K. & Barrick, M. R. (1995). The Big Five personality dimensions: Implications for research and practice in human resources management. In G. R. Ferris (Hrsg.), *Research in personnel and human resources management* (Vol. 13, S. 153 – 200). Bingley: Emerald.

Muchlinski, D., Siroky, D., He, J. & Kocher, M. (2016). Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. *Political Analysis*, *24*, 87 – 103.

Nusinovici, S., Tham, Y. C., Yan, M. Y. C., Ting, D. S. W., Li, J., Sabanayagam, C., Wong, T. Y. & Cheng, C. Y. (2020). Logistic regression was as good as machine learning for predicting major chronic diseases. *Journal of Clinical Epidemiology*, *122*, 56 – 69.

Ones, D. S., Dilchert, S., Viswesvaran, C. & Judge, T. A. (2007). In support of personality assessment in organizational settings. *Personnel Psychology*, *60*, 995 – 1027.

Ones, D. S., Viswesvaran, C. & Reiss, A. D. (1996). Role of social desirability in personality testing for personnel selection: The red herring. *Journal of Applied Psychology*, *81*, 660 – 679.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), 943 – 951.

Pittenger, D. J. (1993). Measuring the MBTI... and coming up short. *Journal of Career Planning and Employment*, *54*(1), 48 – 52.

Robins, R. W., John, O. P., Caspi, A., Moffitt, T. E. & Stouthamer-Loeber, M. (1996). Resilient, overcontrolled, and undercontrolled boys: Three replicable personality types. *Journal of Personality and Social Psychology*, *70*(1), 157 – 171.

Ruxton, G. D. & Beauchamp, G. (2008). Time for some a priori thinking about post hoc testing. *Behavioral Ecology*, *19*, 690 – 693.

Ryan, A. M., McFarland, L., Baron, H. & Page, R. (1999). An international look at selection practices: nation and culture as explanations for variability in practice. *Personnel Psychology, 52*, 359 – 391.

Schmidt, F. L. & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262 – 274.

Vargha, A. & Delaney, H. D. (2000). A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics*, *25*(2), 101 – 132.

Weiß, S. & Hossiep, R. (2013). *BIP-R6. Gütekriterien: Objektivität – Reliabilität – Validität* (Forschungsbericht). Bochum: Ruhr-Universität, Projektteam Testentwicklung.

Weiß, S., Krumscheid, M. & Frieg, P. (2014). *Survival of the fittest!? Befunde zum Person-Job-Fit im deutschsprachigen Raum (Forschungsbericht).* Bochum: Ruhr-Universität, Projektteam Testentwicklung.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). *Data mining: Practical machine learning tools and techniques* (4. Aufl.). Burlington: Morgan Kaufmann.

Historie
Eingegangen: 28. 02. 2020
Revision eingegangen: 14. 09. 2021
Onlineveröffentlichung: 08. 11. 2021

Förderung
Open Access-Veröffentlichung ermöglicht durch die Ruhr-Universität Bochum.

ORCID
Robin Merchel
https://orcid.org/0000-0002-4988-5204

Robin Merchel, M. Sc.
Lehrstuhl für Industrial Sales and Service Engineering
Fakultät für Maschinenbau
Ruhr-Universität Bochum
44780 Bochum
robin.merchel@rub.de