Bestandsaufnahme und Erweiterung um digitale Verfahren bei der Personalauswahl

Von **Prof. Dr. Torsten Biemann** (Universität Mannheim), **Christoph Vogel** (Fachhochschule Nordwestschweiz) und **Prof. Dr. Heiko Weckmüller** (Hochschule Koblenz)

Bei der Greyston-Bäckerei gibt es keine Personalauswahl: keine Lebensläufe, keine Interviews, keine Zeugnisse, stattdessen eine Warteliste. Werden neue Mitarbeitende gesucht, erfolgen Angebote an die obersten Listenplätze. Die kleine Greyston-Bäckerei ist ein Sonderfall, denn auch wenn zurzeit über einen Mangel an Bewerberinnen und Bewerbern geklagt wird, ganz ohne Personalauswahl geht es in den meisten Unternehmen dann doch nicht. Und das ist auch gut so, wie über einhundert Jahre Forschung zur prädiktiven Validität von Auswahlverfahren zeigen. Wir haben zu diesem Thema bereits im Jahr 2012 einen Beitrag verfasst (Biemann/Weckmüller, 2012), Zeit also für ein Update und eine Erweiterung, schließlich handelt es sich bei der Personalauswahl um das „supreme problem" (Ployhart et al., 2017) im HR.

Wir haben zwei Schwerpunkte ausgewählt, auf die wir in diesem Beitrag genauer eingehen möchten: Erstens ist in den letzten Jahren eine neue Diskussion zur Schätzung der prädiktiven Validität[1] aufgekommen, die einen neuen Blick auf die Personalauswahl ermöglicht. Zweitens verändert die Digitalisierung die Personalauswahl, da neue Methoden zur Durchführung (z. B. digitale Interviews) und zur Auswertung (z. B. KI-gestützte Videoanalysen) entstanden sind. Auch wenn die Forschung hierzu noch eher in den Anfängen ist, lohnt sich ein erster Blick auf die Nützlichkeit dieser neuen Möglichkeiten.

Wie valide sind verschiedene Verfahren zur Personalauswahl?

Unser erster Beitrag zur Personalauswahl (Biemann/Weckmüller, 2012) basierte stark auf dem klassischen Artikel von Schmidt und Hunter (1998), der mit über 7.000 Zitationen bei Google Scholar auch heute noch eine wichtige Grundlage für die Schätzung der prädiktiven Validität von Personalauswahlverfahren darstellt und als ein, wenn nicht *der* Klassiker der evidenzbasierten Personalforschung gilt. In den letzten Jahren wurden die dort vorgestellten Ergebnisse aktualisiert, indem neuere empirische Studien in Metaanalysen berücksichtigt wurden. Parallel werden auch die methodischen Grundlagen vergangener Metaanalysen kritisiert. Wir möchten hier nicht auf statistische Details der noch andauernden Diskussion eingehen. Im Wesentlichen ergibt sich gerade bei der Personalauswahl aber das Problem, dass nur diejenigen mit den besten Ergebnissen in den Testverfahren eingestellt werden und man entsprechend gar nicht sagen kann, wie gut die Leistung der Bewerber mit niedrigen Testergebnissen tatsächlich gewesen wäre, weil sie gar nicht die Chance erhalten haben, sich im Unternehmen zu beweisen. Diese Einschränkung der Prädiktorwerte („range restriction") kann korrigiert werden, allerdings müssen dazu Annahmen getroffen werden. Im Ergebnis sind die so korrigierten und ausgewiesenen Korrelationen immer größer als die tatsächlich gemessenen. Da das gleiche Verfahren angewendet wird, bleiben die Ergebnisse innerhalb einer Metastudie zwischen den Auswahlverfahren vergleichbar. Einige neuere Studien wählen dazu vorsichtigere Ansätze und kommen entsprechend zu niedrigeren Schätzungen der prädiktiven Validität. Sackett et al. (2022) sprechen von einer „systematic overcorrection", und die im Folgenden aus dieser Studie berichteten Koeffizienten sind vor diesem Hintergrund zu interpretieren. Die Koeffizienten sind tendenziell geringer, können aber weiterhin genutzt werden, um die nützlichsten Auswahlverfahren zu identifizieren.

Basierend auf Sackett et al. (2022) haben wir die prädiktive Validität wichtiger Auswahlverfahren in Abbildung 1 dargestellt. Wir stellen nicht alle Verfahren dar, sondern beschränken uns auf diejenigen Verfahren, für die eine neue Schätzung mit einer Validität von $r > 0{,}10$ vorliegt.

Das strukturierte Interview hat von allen Verfahren mit einer korrigierten Korrelation von $r = 0{,}42$ die höchste prädiktive Validität, womit sich anhand des bisherigen Forschungsstands klar abzuzeichnen scheint, dass strukturierte Interviews gegenüber unstrukturierten Bewerbungsgesprächen ($r = 0{,}19$) einen Mehrwert bieten und damit insgesamt gute Schlüsse auf die künftige Leistung der Bewerbenden möglich sind. Ebenso weisen Job-Wissenstests ($r = 0{,}40$) und biographische Informationen wie zum Beispiel Ausbildung und vorherige Erfahrungen ($r = 0{,}38$) eine gute prädiktive Validität auf. Interessant ist das im Vergleich zu vorherigen Studien schlechtere Abschneiden von Intelligenztests ($r = 0{,}31$),

1 Die prädiktive Validität ist vereinfacht gesagt die Korrelation zwischen Ergebnissen eines Auswahlverfahrens und dem späteren Berufserfolg. Je höher diese Korrelation, desto besser ist das Auswahlverfahren.

Abb. 1: Validität von Personalauswahlverfahren

Auswahlverfahren	Prädiktive Validität (korrigierter Korrelationskoeffizient)*
Strukturiertes Interview	0,42
Job-Wissenstest	0,40
Biographische Informationen	0,38
Arbeitsprobe	0,33
Intelligenztest	0,31
Integritätstest	0,31
Persönlichkeitsbasierte emotionale Intelligenz	0,30
Assessment Center	0,29
Situational Judgement Test	0,26
Gewissenhaftigkeit**	0,25
Interessen	0,24
Emotionale Stabilität**	0,23
Extraversion**	0,21
Gewissenhaftigkeit	0,19
Unstrukturiertes Interview	0,19
Verträglichkeit**	0,19
Offenheit für Erfahrungen**	0,12

* Gemessen wird die prädiktive Validität als Korrelation des Auswahlkriteriums mit der Arbeitsleistung.

** Auf Basis kontextabhängiger Skalen zur Messung der Persönlichkeitseigenschaft.

Quelle: Eigene reduzierte Darstellung auf Basis Sackett et al. (2022)

da diese zwar immer noch klar zu den geeigneten Verfahren, aber nicht mehr zu den Spitzenreitern gehören, die Schmidt und Hunter (1998) noch in diesem Verfahren sahen. Die prädiktive Validität von Assessment Centern bleibt im guten Bereich (r = 0,29), ebenso wie einige der Persönlichkeitsdimensionen aus den Big 5. Interessant ist bei der Persönlichkeitsmessung die bessere Eignung von kontextspezifischen Skalen, die den jeweiligen Arbeitskontext bei der Erfassung mit berücksichtigen. So liegt die prädiktive Validität der kontextabhängigen Skalen zwischen r = 0,12 (Offenheit für Erfahrungen) und r = 0,25 (Gewissenhaftigkeit), bei den allgemeinen Skalen dagegen ist die prädiktive Validität nur bei Gewissenhaftigkeit mit r = 0,19 über r = 0,10, weswegen sie auch nicht in Abbildung 1 gelistet sind.

Auf zwei Verfahren möchten wir etwas genauer eingehen, weil sie in der bisherigen wissenschaftlichen Diskussion eine untergeordnete Rolle gespielt haben und deshalb in unserem ersten Beitrag keine Erwähnung fanden. Erstens wurde in den letzten Jahren viel zu Situational Judgement Tests (SJTs) geforscht. In diesem Verfahren werden den Kandidaten hypothetische Situationen skizziert, wie sie im Arbeitsalltag auftreten können, und die Kandidaten müssen aus verschiedenen Handlungsmöglichkeiten ihre Reaktion auswählen. Basis für die Fragen eines SJTs sind die Herausforderungen der ausgeschriebenen Position. Die prädiktive Validität (r = 0,26) ist im mittleren Bereich. Bei Sackett et al. (2022) wird innerhalb dieser Testverfahren noch unterschieden zwischen persönlichkeitsbasierten („personality-based“) und fähigkeitsbasierten („ability-based“) Verfahren zur Erfassung der emotionalen Intelligenz, deren prädiktive Validität mit r = 0,30 beziehungsweise r = 0,22 im guten Bereich liegt. Auf die anhaltende Diskussion zur Abgrenzung emotionaler Intelligenz von Persönlichkeit und dem traditionellen Intelligenzbegriff möchten wir lediglich hinweisen, diese hier aber nicht vertiefen. Wichtigstes Kriterium für die Unternehmenspraxis bleibt, dass ein Verfahren eine hohe Vorhersagekraft haben sollte, und nicht ganz so sehr, warum es (nicht) funktioniert. In der Praxis zeigt sich jedenfalls eine gute prädiktive Validität und auch eine ähnlich hohe Vorhersagekraft, wenn emotionale Intelligenz bei Mitarbeitern erhoben wird, um deren Arbeitsleistung vorherzusagen.

Stellenweise wird in der Forschung zu Personalauswahlverfahren moniert, dass Äpfel mit Birnen verglichen werden, weil einige Verfahren wie zum Beispiel Intelligenztests explizit einzelne Konstrukte erfassen, bei anderen Methoden wie zum Beispiel dem Interview dagegen eine Bandbreite verschiedener Konstrukte in eine Gesamtbewertung einfließen. Auch ist die Bedeutung sozialer Erwünschtheit weiterhin ein wichtiges Forschungsthema, die fehlende einheitliche theoretische Fundierung erschwert aber Fortschritte dazu. Da für die Unternehmenspraxis diese Diskussionen eher nachrangig sind, vertiefen wir auch diese Themen nicht, sondern verweisen interessierte Leser auf Ployhart et al. (2017).

Wie gut sind digitale Verfahren und KI-basierte Auswertungen in der Personalauswahl?

Die Digitalisierung bietet für die Personalauswahl sowohl neue Verfahren als auch neue Auswertungsmöglichkeiten, gerade in den letzten Jahren durch Weiterentwicklungen KI-basierter Ansätze. Auf beide Bereiche möchten wir im Folgenden eingehen. Leider können wir dazu nicht auf umfangreiche Metaanalysen zurückgreifen, sondern müssen uns auf

Review-Artikel und einzelne empirische Studien beschränken. Erstens wurden in den letzten beiden Jahrzehnten digitale Alternativen und Erweiterungen traditioneller Verfahren entwickelt, beispielsweise digitale Interviews und Online-Tests mit Gamification-Elementen. Inwieweit lassen sich bestehende Forschungsergebnisse auf diese Verfahren übertragen? Studien zur prädiktiven Validität digitaler Verfahren sind überraschend rar und ein robuster Vergleich mit ihren traditionellen Pendants ist noch nicht möglich. Was sich aber recht durchgängig zeigt, ist die geringere Akzeptanz von digitalen Verfahren durch die Bewerbenden (Woods et al., 2020).

Zweitens wurden in den letzten Jahren KI-basierte Auswertungen entwickelt, die mit dem traditionellen Vorgehen verglichen werden, bei denen Menschen die Beurteilungen vornehmen. Diesem Thema hat sich ein Sonderheft der Zeitschrift Personnel Psychology im Jahr 2023 gewidmet (Campion/Campion, 2023). Für die Erfassung der prädiktiven Validität wird eine Korrelation zwischen den Empfehlungen des Algorithmus und einem später erfassten Kriterium berechnet, zum Beispiel der Leistung des neu eingestellten Mitarbeiters in den ersten Jahren. Leider liegt Studien zur Nützlichkeit KI-basierter Verfahren in der Regel eine andere Logik zugrunde, denn die allermeisten Studien in diesem Bereich vergleichen die Ähnlichkeit KI-basierter und menschlicher Empfehlungen. Anders gesagt: Es wird getestet, wie gut die Maschine eine menschliche Entscheidung replizieren kann. Verschiedene Studien zur Personalauswahl zeigen, dass dies durchaus gelingen kann. Dieser Nachweis ist ein interessanter erster Schritt, muss aber noch über Studien ergänzt werden, die einen Zusammenhang mit tatsächlicher Arbeitsleistung oder anderen relevanten Kriterien aufzeigen. Für diese Erfassung der prädiktiven Validität KI-basierter Auswahlentscheidungen gibt es nur wenige Studien, ein Beispiel ist Campion et al. (2024) mit einer Studie, bei der Natural Language Processing zur Analyse von Bewerbungsunterlagen (z. B. Lebenslauf und Referenzschreiben) eingesetzt wurde. Die Autoren können zeigen, dass der Algorithmus geeignet ist, spätere Trainingsleistung vorherzusagen. Noch sind die einzelnen Ergebnisse allerdings eher fragmentiert und eine systematische Analyse der prädiktiven Validität, so wie wir sie oben für traditionelle Verfahren dargestellt haben, ist noch nicht möglich.

Abgesehen von der meist noch ungenauen Erfassung der prädiktiven Validität sowie der geringeren Akzeptanz der Verfahren bestehen sowohl von Seiten der Bewerberinnen und Bewerber wie auch von Organisationsvertretern zusätzliche Vorbehalte. So geht der Einsatz digitaler und KI-basierter Verfahren in der Personalauswahl unter anderem mit einer negativeren Wahrnehmung der Arbeitgeberattraktivität sowie einer schlechteren Bewertung der Ethik einher als die herkömmliche Sichtung von Bewerbungsunterlagen und die Durchführung von Einstellungsinterviews vor Ort (Kanning/Kraul/Litz, 2019). Darüber hinaus haben einige Beispiele aus der jüngsten Vergangenheit aufgezeigt, dass durch den Einsatz von KI die Personalauswahl nicht automatisch objektiver wird, sondern dass KI-Algorithmen auch diskriminierende Ergebnisse produzieren können (Köchling/Wehner, 2020). Aus diesem Grund betont zum Beispiel Green (2022) die Wichtigkeit der menschlichen Aufsicht beim Einsatz von KI. Dadurch können Gefahren wie fehlerhafte Outputs oder das Verletzen von gesetzlichen Rahmenbedingungen minimiert werden.

Zusammenfassung und Schlussfolgerungen

- Der Einsatz valider Personalauswahlverfahren bietet nach wie vor einen Mehrwert.
- Die teils größeren Veränderungen der Werte der prädiktiven Validität der neueren Resultate im Vergleich zu älteren Arbeiten ermöglichen eine exaktere Einordnung der Güte verschiedener Verfahren der Personalauswahl.
- Zur Validität digitaler und KI-basierter Personalauswahlverfahren existieren bislang nur wenige Studien und der Vergleich mit bisherigen Ergebnissen zur prädiktiven Validität von herkömmlichen Auswahlverfahren ist nur schwer möglich.
- Rund um die Nutzung digitaler und KI-basierter Verfahren gibt es zusätzliche Vorbehalte wie zum Beispiel die Akzeptanz der Verfahren oder die (negativen) Auswirkungen auf die Außenwirkung als Arbeitgeber, die mit bedacht werden müssen.

PROF. DR. TORSTEN BIEMANN
Lehrstuhl für Personalmanagement und Führung
Universität Mannheim
E-Mail: biemann@bwl.uni-mannheim.de
http://biemann.bwl.uni-mannheim.de

CHRISTOPH VOGEL
Wissenschaftlicher Mitarbeiter
Institut für Personalmanagement und Organisation
Fachhochschule Nordwestschweiz
E-Mail: christoph.vogel@fhnw.ch
https://www.fhnw.ch/de

PROF. DR. HEIKO WECKMÜLLER
Professur Human Resources
Hochschule Koblenz
Rheinahrcampus Remagen
E-Mail: weckmueller@rheinahrcampus.de
www.hs-koblenz.de

LITERATURVERZEICHNIS

Biemann, T./Weckmüller, H. (2012): Methoden der Personalauswahl: Was nützt? PERSONALquarterly, 64(1), 46-49.

Campion, M. A./Campion, E. D. (2023): Machine learning applications to personnel selection: Current illustrations, lessons learned, and future research. Personnel Psychology, 76(4), 993-1009.

Campion, E. D./Campion, M. A./Johnson, J./Carretta, T. R./Romay, S./Dirr, B./Deregla, A./Mouton, A. (2024): Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. Journal of Applied Psychology, 109(3), 307-338.

Green, B. (2022): The flaws of policies requiring human oversight of government algorithms. Computer Law & Security Review, 45, 1-22.

Kanning, U. P./Kraul, L.-F./Litz, R. Z. (2019): Einstellungen zu digitalen Methoden der Personalauswahl. Journal of Business and Media Psychology 10(1), 57-71.

Köchling, A./Wehner, M. C. (2020): Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. Business Research 13, 795–848.

Ployhart, R. E./Schmitt, N./Tippins, N. T. (2017): Solving the Supreme Problem: 100 years of selection and recruitment at the Journal of Applied Psychology. Journal of Applied Psychology, 102(3), 291-304.

Sackett, P. R./Zhang, C./Berry, C. M./Lievens, F. (2022): Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. Journal of Applied Psychology, 107(11), 2040-2068.

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(2), 262-274.

Woods, S. A./Ahmed, S./Nikolaou, I./Costa, A. C./Anderson, N. R. (2020): Personnel selection in the digital age: a review of validity and applicant reactions, and future research challenges. European Journal of Work and Organizational Psychology, 29(1), 64-77.