25 Jahren nach Schmidt und Hunter (1998): Wo stehen wir heute in der beruflichen Eignungsdiagnostik?

Editorial zum Themenheft

Herausgeber: Prof. Dr. Stefan Höft (HdBA Mannheim) &
Prof. Dr. Klaus G. Melchers (Universität Ulm)

Generationen von Psychologinnen und Psychologen sind in ihrem Studium mit dem Überblicksartikel von Schmidt und Hunter aus dem Jahr 1998 vertraut gemacht worden, in dem der Stand aus (damals) 85 Jahren eignungsdiagnostischer Forschung gebündelt wurde. In dieser Übersicht wurde sehr eindrucksvoll belegt, dass gute Auswahlverfahren erstaunlich valide Vorhersagen von beruflicher Leistung bzw. von Trainings- und Ausbildungsleistung ermöglichen. Damit wurde auch deutlich, welchen finanziellen Mehrwert gute berufliche Eignungsdiagnostik haben kann. Die wissenschafts- und praxisbezogenen Schlussfolgerungen der Autoren waren prägend für viele nachfolgende Forschungsbestrebungen und Gegenstand vielfältiger Wissenschafts-Praxis-Kontroversen.

Aber was bleibt nach 25 Jahren in der Rückschau auf diesen Artikel bestehen? Welche Befunde und Argumente haben aus heutiger Sicht weiterhin Relevanz, welche Darstellungen und Argumente werden im Rückblick anders gesehen oder relativeren sich angesichts aktueller Befunde? Speziell die Publikation von Sackett et al. (2022) zur Reanalyse und Aktualisierung der Validitätsbefundlage der unterschiedlichen eignungsdiagnostischen Verfahren hat hier neue Impulse und Anlass gegeben, scheinbar etabliertes Wissen neu zu bewerten.

Im vorliegenden Themenheft der Zeitschrift „Wirtschaftspsychologie“ werden die Publikationen von Schmidt und Hunter (1998) sowie von Sackett et al. (2022) selbst und ihre Wirkung auf Wissenschaft und Praxis in insgesamt elf Einzelbeiträgen aus ganz unterschiedlichen Perspektiven beleuchtet.

- Als Bezugsrahmen für die Einzelbeiträge des Themenhefts beschreibt Stefan Höft zunächst die wesentlichen Befunde von Sackett und Hunter (1998) und fasst die relevanten Kritikpunkte und Aktualisierungen von Sackett et al. (2022) zusammen.

In den nachfolgenden fünf Beiträgen wird das Themenfeld aus unterschiedlichen wissenschaftlichen Blickwinkeln betrachtet:

- Klaus Melchers diskutiert fünf teils methodische, teils inhaltliche Aspekte, die Einfluss nehmen auf den Aussagegehalt der Übersichtsstudie von Schmidt und Hunter (1998) und die zum Teil auch für die Aktualisierung von Sackett et al. (2022) noch relevant sind.
- Martin Kleinmann betrachtet zunächst die Resonanz auf die Schmidt&Hunter-Übersicht in der Fachwelt und geht noch einmal genauer auf die neueren Ergebnisse der Arbeit von Sackett et al. ein. Zusätzlich beschreibt er bestehende metaanalytische Befunde und relevante Einzelstudien, die nicht bei Sackett et al. berücksichtigt wurden.
- Klaus Moser und George Gunnesch-Luca nehmen das Jubiläum der Schmidt&Hunter-Übersicht zum Anlass, unterschiedliche Unzulänglichkeiten der eignungsdiagnostischen Forschung zu diskutieren, die ihres Erachtens (un-)mittelbar auf Setzungen und Aussagen in dieser Studie zurückzuführen sind.
- Marvin Neumann, Susan Niessen und Rob Meijer erweitern die Befunde von Sackett et al., indem sie ergänzend zu Einzelverfahrensvaliditäten die voraussichtliche Validität von Verfahrenskombinationen unter unterschiedlichen Rahmenbedingungen berechnen. Zusätzlich diskutieren sie Möglichkeiten, wie die Akzeptanz für die aus wissenschaftlicher Sicht überlegenen mechanischen Verfahrenskombinationen bei betrieblichen Entscheidungsträgern gesteigert werden kann.
- Uwe Kanning schlägt eine direkte Brücke zum eignungsdiagnostischen Praxiseinsatz: Er diskutiert bestehende Barrieren für die Umsetzung wissenschaftlicher Erkenntnisse in der betrieblichen Anwendung und leitet daraus mögliche Aufgaben für die Forschung ab.

Die restlichen fünf Beiträge behandeln die Thematik aus einer stärker praxisorientierten Perspektive:

- Benedikt Hell, Katja Päßler und Miriam Nido untersuchen, inwieweit die alten (Schmidt & Hunter) und neuen (Sackett et al.) Validitätsbefunde mit entsprechenden Einschätzungen von Praktikerinnen und Praktikern korrespondieren. Sie diskutieren die möglichen Ursachen für die gefundenen Divergenzen und leiten ähnlich wie Uwe Kanning Handlungsbedarf für die Wissenschaft ab.
- Dieter Hasselmann arbeitet als Unternehmensberater im Bereich der Top-Management-Auswahl. Aus dieser Perspektive betrachtet er die metaanalytischen Validitätsbefunde und relativiert den daraus ableitbaren Erkenntnisgewinn für sein Arbeitsgebiet.
- Hella Klemmert, Nicolas Sander, Erik Sengewald und Dorothea Klinck zeigen in ihrer umfassenden berufspsychologischen Studie, dass kognitive Fähigkeiten trotz aller relativierender Kritik eine beeindruckende Prognosequalität für Ausbildungserfolg aufweisen und ein sinnvolles Strukturierungsmerkmal für berufliche Anforderungen sind.
- Hubert Annen steuert ein Schweizer Anwendungsbeispiel zur Qualitätssicherung von Eignungsdiagnostik bei. Die Non-Profit-Organisation „Swiss Assessment“ hat hier ein Zertifizierungsprogramm für Assessment-Center-Realisierungen entwickelt und bietet dieses seit einigen Jahren erfolgreich für ihre Mitglieder an.
- Susanne Schulte und Maren Hiltmann greifen noch einmal den zentralen Befund von Sackett et al. auf, wonach strukturierte Einstellungsgespräche im Verfahrensvergleich die höchste Validität aufweisen. Sie konkretisieren, was genau den Strukturierungsgrad eines Einstellungsgesprächs ausmacht und leiten konkrete Praxisempfehlungen ab.

Unser Ziel war es, im Themenheft eine anregende Mischung aus Theoriereflexionen und Diskussionsbeiträge, aber auch empirische Forschungsbeiträgen und Best-Practice-Praxisanwendungen zu bündeln. Das ist aus unserer Sicht sehr gut gelungen. Wir bedanken uns bei allen beteiligten Autorinnen und Autoren sowie den Gutachtenden für die sehr angenehme Zusammenarbeit.

Allen Leserinnen und Lesern wünschen wir eine spannende Lektüre!

Stefan Höft und Klaus Melchers

Literatur

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.

25 Jahre nach Schmidt und Hunter (1998): Aktuelle Forschungslage zur Validität eignungsdiagnostischer Verfahren

Stefan Höft

Hochschule der Bundesagentur für Arbeit (HdBA), Mannheim

Zusammenfassung

Nach einer kurzen Einführung zu metaanalytischen Grundprinzipien werden die wesentlichen Aussagen der bekannten Übersichts zur kriteriumsbezogenen Validität von eignungsdiagnostischen Verfahren von Schmidt und Hunter (1998) zusammengefasst. Sehr deutlich wurde von den Autoren auch in anderen Publikationen die hohe Validität von kognitiven Fähigkeitstests betont. Im Anschluss werden die Kritikpunkte von Sackett, Zhang, Berry und Lievens (2022) zur Studie von Schmidt und Hunter dargestellt. Im Mittelpunkt steht dabei die aus Sicht von Sackett et al. zu großzügig angewandte Korrektur von Varianzeinschränkungen im Prädiktor. Quintessenz ist die Präsentation einer aktualisierten Validitätsübersicht, die auf angepassten Reanalysen und neuen Befunden basiert. Hier zeigt sich eine deutlich homogenere Validitätslage mit einer Mehrheit der Koeffizienten zwischen .20 und .40 über alle Verfahrensklassen hinweg. Kognitive Fähigkeitstests rangieren nur noch im Mittelfeld, während strukturierte Einstellungsgespräche die höchsten Werte aufweisen. Im Fazit werden die Konsequenzen der Aktualisierung für Forschung und Praxis kurz reflektiert.

Schlüsselwörter: kriteriumsbezogene Validität, eignungsdiagnostische Verfahren, Metaanalyse Korrektur von Varianzeinschränkungen, Personalauswahl

25 years after Schmidt and Hunter (1998): Current research findings concerning the validity of personnel selection instruments

Abstract

After a brief introduction to basic meta-analytic principles, the main statements of the well-known review by Schmidt and Hunter (1998) on the criterion-related validity of personnel selection instruments are summarized. The high validity of cognitive ability tests was also very clearly emphasized by the authors in other publications at that time. In the following, the current criticisms of Sackett, Zhang, Berry, and Lievens (2022) regarding this study are presented. The focus is on the correction for variance restrictions in the predictor, which Sackett et al. consider too generously applied. The quintessence is the presentation of an updated validity overview based on adjusted reanalyses and new findings. Here, a considerably more homogeneous validity situation is shown with a majority of the coefficients between .20 and .40 across all types of selection procedure. Cognitive ability tests rank only in the midfield, while structured job interviews show the highest values. The conclusion briefly reflects on the consequences of the update for research and practice.

Keywords: criterion-related validity, aptitude testing, meta-analysis, corrections of restriction of variance, personnel selection

1. Referenzpunkt für die eignungsdiagnostische Forschung

Wenn in einem fachlichen Austausch oder einer einführenden Publikation zum Thema ein Überblick zur Validität von eignungsdiagnostischen Verfahren gegeben werden soll, stellt die Arbeit von Schmidt und Hunter (1998) seit 25 Jahren eine feste Orientierungsgröße dar. Generationen von Psychologinnen und Psychologen sind in ihrem Studium mit dieser Studie vertraut gemacht worden, in der der Stand aus (damals) 80 Jahren eignungsdiagnostischer Forschung gebündelt wurde. Die wissenschafts- und praxisbezogenen Schlussfolgerungen der Autoren waren prägend für viele nachfolgende Forschungsbestrebungen und Gegenstand vielfältiger Praxiskontroversen.

Insoweit stellt bereits das silberne Jubiläum der Publikation nach 25 Jahren einen sinnvollen Anlass dar, eine Reflexion zu dieser Studie vorzunehmen. Gleichzeitig wurde letztes Jahr eine metaanalytische Aktualisierung von Sackett, Zhang, Berry und Lievens (2022) vorgelegt. Hier werden wesentliche Vorgehensweisen und Aussagen von Schmidt und Hunter (1998) hinterfragt, und aufbauend auf aktualisierten Befunden wird eine Neuberechnung und -bewertung zur generellen Validität unterschiedlicher eignungsdiagnostischer Verfahrensansätze präsentiert.

In diesem Artikel werden nach einer kurzen Einführung in die metaanalytischen Grundprinzipien die wesentlichen Aussagen der Originalstudie von Schmidt und Hunter (1998) zusammengefasst. Danach wird die Kritik von Sackett et al. beschrieben und ein Überblick zu ihren Aktualisierungen gegeben. Ziel ist es, die geänderte Befundlage in einem kompakten Überblick darzustellen. In den nachfolgenden Beiträgen des Themenhefts werden dann unterschiedliche Aspekte der wissenschaftlichen Debatte aus verschiedenen Perspektiven diskutiert.

2. Metaanalysen als wichtige Bezugspunkte für systematische Forschung

Vor einer Detailanalyse der Studien soll noch einmal kurz das Grundprinzip von Metaanalysen in vereinfachter Form beschrieben werden. Eine ausführlichere, weitgehend nichttechnische Einführung gibt Höft (2014). An dieser Stelle werden nur die im Weiteren relevanten Aspekte knapp dargestellt.

Als Ankerpunkte in der Forschungsdiskussion dienen häufig Übersichtsarbeiten, in denen der bestehende Forschungsstand zu einem Thema zusammengefasst wird. Bei traditionellen „narrativen“ Arbeiten bewerteten die jeweiligen Autoren nach selbst gewählten Kriterien die bestehenden Forschungsbefunde und formulierten dann eine subjektive Gesamtbewertung. Die Funktion dieser Arbeiten, den „state of the art“ wiederzugeben, konnte dadurch nur unzureichend erfüllt werden, da immer auch die jeweils gewählte Vorgehensweise kritisch hinterfragt werden kann (z. B. „Warum wurde die Einzelstudie X als weniger aussagekräftig bewertet als Einzelstudie Y? Wie ist die letztliche Bewertung der Forschungslage zustande gekommen?“). Bei einer Metaanalyse (auch Sekundärstudie genannt) werden hingegen empirischen Einzelstudien zum gleichen Thema (so genannte Primärstudien) systematisch kombiniert, um zu einem gut rekonstruierbaren Gesamturteil zu kommen, das den aktuellen Forschungsstand bestmöglich repräsentiert.

Es gibt unterschiedliche metaanalytische Techniken (vgl. z. B. Döring, 2023). Im Bereich der Eignungsdiagnostik hat sich allerdings die Validitätsgeneralisierung nach Schmidt und Hunter (2015) durchgesetzt, da sie maßgeschneidert ist für die Fragestellungen in der Eignungsdiagnostik. Das Grundprinzip des Vorgehens ist in Abbildung 1 dargestellt.

Im Mittelpunkt der im Weiteren relevanten Untersuchungen steht der empirisch gefundene Zusammenhang zwischen dem Ergebnis eines eignungsdiagnostischen Verfahrens (hier X genannt) und einem beruflichen Erfolgsmaß (z. B. ein Vor-

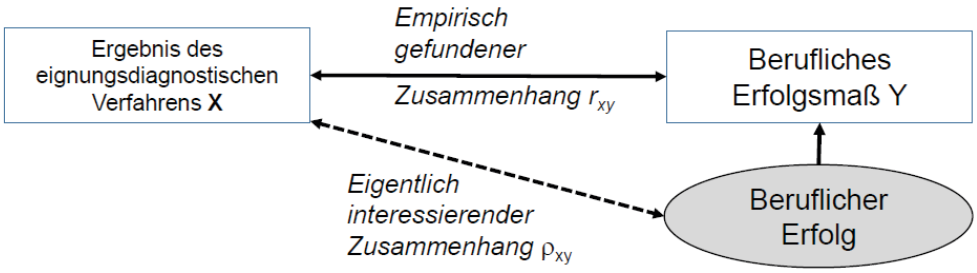

Abbildung 1
Das Grundmodell der metaanalytischen Validitätsgeneralisierung nach Schmidt und Hunter (2015).

gesetztenurteil, hier mit Y bezeichnet). Als Effektmaß wird der Korrelationskoeffizient *r* verwendet, der den linearen Zusammenhang zwischen dem eignungsdiagnostischen Ergebnis und dem beruflichen Erfolg als Zahl angibt, die zwischen -1 und 1 liegen kann. Je näher der Koeffizient an 1 liegt, desto stärker ist der gleichsinnige Zusammenhang zwischen X und Y: Inhaltlich bedeuten positive Werte von *r*, dass aus dem guten Abschneiden in dem eignungsdiagnostischen Verfahren mit einer gewissen Sicherheit auf den beruflichen Erfolg der betreffenden Person geschlossen werden kann. Das entsprechende Verfahren erweist sich somit als kriteriumsbezogen valide, da es seinen eignungsdiagnostischen Zweck erfüllt.

In einer Metaanalyse werden die in den Primärstudien zu gleichen Themenstellungen (z. B. zur kriteriumsbezogenen Validität des Verfahrens X) erhobenen Korrelationskoeffizienten gemittelt. Parallel werden unterschiedliche Korrekturen durchgeführt, durch die verzerrende Einflüsse auf den empirisch gefundenen Zusammenhang korrigiert werden sollen. Zu den Standardkorrekturen gehören:

- **Stichprobengewichtung:** Studien mit großen Stichproben gehen bei der Berechnung des Gesamtkoeffizienten mit einem höheren Gewicht ein als Untersuchungen, die auf kleinen Stichprobengrößen basieren.
- **Korrektur unzureichender Kriteriumsreliabilität:** Berufliche Erfolgsmaße erfassen den beruflichen Erfolg nicht perfekt. Da dieser den empirischen Zusammenhang r_{xy} mindernde Fehlereinfluss nicht dem eignungsdiagnostischen Prädiktor anzulasten ist, wird diese Unreliabilität des Kriteriums statistisch korrigiert. In Abbildung 1 wird das durch den „eigentlich interessierenden Zusammenhang ρ_{xy}“ veranschaulicht.
- **Direkte Varianzeinschränkungskorrektur:** In Personalauswahlsituationen erhalten in der Regel nicht alle diagnostizierten Bewerbenden eine Zusage. Diese Personen fehlen dementsprechend später in einer Validierungsstudie, da zu ihnen keine Berufserfolgskriterien vorliegen. Wenn die Entscheidung zur Zu- bzw. Absage auf Grundlage des erforschten eignungsdiagnostischen Verfahrensergebnisses erfolgt, kann die resultierende Varianzeinschränkung durch eine einfache Verrechnung korrigiert werden.

Im Laufe der Jahrzehnte wurden zusätzliche, z. T. deutlich komplexere Korrekturformeln entwickelt, die die spezifischen Erhebungsbedingungen der Primärstudien (z. B. unterschiedliche Formen von indirekten Varianzeinschränkungen) und die generellen Rahmenbedingungen für Forschungsstudien (z. B. geringere Publikationschancen für Studien mit nicht signifikanten Befunden) berücksichtigen (vgl. Schmidt & Hunter, 2015). Erwähnenswert ist beispielsweise die Metaanalyse von Huffcutt, Culbertson und Weyhrauch (2014) zur Validität von eignungsdiagnostischen Interviews, bei der immerhin sechs Formen von Varianzeinschränkung unterschieden werden (z. B. Auswahl erfolgte auf Grundlage einer Kombination aus Interview und weiteren Verfahren, oder die Auswahl erfolgte sequentiell mit einer Vorauswahl auf Grundlage anderer Verfahren).

3. Wesentliche Aussagen von Schmidt und Hunter (1998)

Schmidt und Hunter publizieren ihre Übersichtsstudie 1998 im *Psychological Bulletin*, einer der einflussreichsten englischsprachigen psychologischen Fachzeitschriften mit einem Fokus auf teildisziplinübergreifenden Themenstellungen. Die Studie wird häufig fälschlicherweise als Metaanalyse bezeichnet. Tatsächlich wird unter dem Titel *„The Validity and Utility of Selection Methods in Personnel Psychology: Practical and Theoretical Implications of 85 Years of Research Findings“* eine Übersicht zu bis dato publizierten metaanalytischen Befunden ohne substantielle Neuberechnungen gegeben.

Schmidt und Hunter (1998) unterscheiden in ihrer Ergebnisdarstellung zwischen zwei Kriteriumsbereichen: Vorhersagequalität für berufliche Leistung (Job Perfomance) sowie für Ausbildungs- und Weiterbildungserfolg (Training Performance). Im Mittelpunkt des Interesses stehen regelmäßig die Befunde zur Validität eignungsdiagnostischer Verfahren für berufliche Leistung. Eine Zusammenschau der wichtigsten Befunde dazu (verkürzte Darstellung nach Schuler & Höft, 2001) zeigt Tabelle 1. Für eine grobe Gliederung wurden die Verfahren orientiert am trimodalen Ansatz der Personalpsychologie (vgl. z. B. Höft & Schuler, 2019) in drei Gruppen aufgeteilt: Eigenschaftsorientierte Verfahren erfassen als beruflich relevant eingestufte psychologische Konstrukte. Bei simulationsorientierten Verfahren werden berufserfolgskritische Situationen nachgestellt und die Bewährung der Personen in diesen Simulationen bewertet. Und biografieorientierte Ansätze schließlich erfassen anforderungsrelevantes Verhalten aus der Vergangenheit der Person. Einige zugeordneten Verfahren können in der konkreten Umsetzung teilweise auch eine Mischcharakteristik aufweisen. So werden im Einstellungsgespräch nicht nur biografische Fragen gestellt, sondern häufig beispielsweise auch situative Fragen.

Zu beachten sind zunächst die Rahmenbedingungen der Studienauswahl, auf die bereits Schuler und Höft (2001, S. 105) hingewiesen haben:

- Es handelt sich um eine selektive Auswahl von Befunden. Bei konkurrierenden Metaanalysen zum gleichen Thema, z. B. zur Validität von Per-

	Eignungsdiagnostisches Verfahren	Val	mR	inkrV
Eigenschaft	Allgemeine kognitive Fähigkeitstests	.51		
	Integritätstests	.41	.65	.14
	Gewissenhaftigkeitstests	.31	.60	.09
	Fachkenntnistests	.48	.58	.07
	Interessen	.10	.52	.01
Simulation	Probezeit	.44	.58	.07
	Arbeitsproben	.54	.63	.12
	Assessment Center	.37	.53	.02
Biografie	Graphologie	.02	.51	.00
	Biografische Daten	.35	.52	.01
	Unstrukturiertes Einstellungsgespräch	.36	.55	.04
	Strukturiertes Einstellungsgespräch	.51	.63	.12

Tabelle 1
Überblick zu den wesentlichen Ergebnissen der Übersichtsarbeit von Schmidt und Hunter (1998) (adaptiert aus Schuler & Höft, 2001)

Anmerkungen: Als Artefakte wurden die Kriteriumsunreliabilität und Varianzeinschränkungen korrigiert. Val ρ = kriteriumsbezogene Validität (Kriterium: Leistungsbeurteilung zumeist über Vorgesetztenurteile); mR = geschätzte multiple Korrelation R bei gemeinsamer Berücksichtigung des jeweiligen Verfahrens zusammen mit einem allgemeinen kognitiven Fähigkeitstest; inkrV = inkrementelle Validität, d. h. der Validitätszuwachs durch Hinzunahme des zweiten Prädiktors zusätzlich zum allgemeinen kognitiven Fähigkeitstest

sönlichkeitsverfahren, wurde eine „wohlwollende“ Auswahl getroffen, bei der die jeweils höchste gefundene Validität in die Übersicht übernommen wurde.
- Die Koeffizienten stammen aus unterschiedlichen Metaanalysen und wurden nur hinsichtlich Kriteriumsunreliabilität und direkten Varianzeinschränkungen korrigiert. Als Kriterium wird zumeist ein globales Vorgesetztenurteil herangezogen. Andere Randaspekte (z. B. Kriterien zur Aufnahme oder von Primärstudien wurden nicht parallelisiert).
- Die meisten Metaanalysen basieren (auch heute noch) auf US-amerikanischen Primärstudien. Durch andere gesetzliche Rahmenbedingungen, andere Berufszuordnung usw. sind nicht alle Befunde direkt auf deutsche Verhältnisse übertragbar. So gibt es in Deutschland immer noch wenige Befunde zu Integritätstests, die das Ziel haben, kontraproduktives betriebliches Verhalten vorherzusagen.

Was ist nun gemäß Schmidt und Hunter hinsichtlich der Validität eignungsdiagnostischer Verfahren festzustellen? Zunächst ist als beruhigender Befund festzustellen, dass alle etablierten diagnostischen Verfahren aus den unterschiedlichen eignungsdiagnostischen Diagnoseansätzen (eigenschafts-, simulations- und biografieorientierte Verfahren) eine substantielle kriteriumsbezogene Validität für Berufserfolg (überwiegend operationalisiert über Vorgesetztenurteile, vgl. zu möglichen Problem diesbezüglich den Beitrag von Moser & Gunnesch-Luca, 2023, in diesem Heft) aufweisen. Die Spitzenreiter kommen aus allen drei Verfahrenskategorien: Arbeitsproben (ρ = .54), allgemeine kognitive Fähigkeitstests (ρ = .51) sowie strukturierte Einstellungsgespräche (ρ = .51). Schmidt und Hunter weisen kognitiven Fähigkeitstests allerdings eine besondere Relevanz als eignungsdiagnostische Prädiktoren zu, da sie eine sehr gute kriteriumsbezogene Validität aufweisen und gleichzeitig für sie eine hervorragende Kosten-Nutzen-Relation besteht.

Ergänzend führen Schmidt und Hunter (1998) Berechnungen zur inkrementellen Validität anderer Prädiktoren ergänzend zu kognitiven Fähigkeitsverfahren an (letzte zwei Spalten in Tabelle 1). Hier zeigt sich beispielsweise, dass die Hinzunahme eines unstrukturierten Einstellungsgesprächs nur wenig Zusatzinformationen liefert, während das (von vornherein validere) strukturierte Einstellungsgespräch deutlich bessere Ergebnisse erzielt. Zu beachten ist dabei aber, dass diese Angaben nicht aus empirischen Studien mit kombinierter Verfahrensauswahl stammen, sondern berechnet wurden: Wenn die Korrelation r_{X1X2} zwischen zwei Prädiktoren X1 und X2 sowie deren jeweilige Korrelation r_{X1Y} sowie r_{X2Y} mit Berufserfolg aus anderen Studien (hier: unterschiedlichen anderen Metaanalysen) bekannt sind, können ihre Gewichte in einer multiplen Regression sehr einfach theoretisch ohne reale Umsetzung berechnet werden.

In späteren Publikationen (z. B. Schmidt, 2002; Schmidt & Hunter, 2004) betonten die Autoren immer wieder die besondere Relevanz der Intelligenzmaße für die berufliche Eignungsdiagnostik und führten hierzu teilweise auch intensive Debatten (z. B. Schmidt, Le, Oh & Shaffer, 2007, sowie die damit verbundenen Diskussionsbeiträge). Gleichzeitig zeigt die eher zurückhaltende Einsatzhäufigkeit von kognitiven Fähigkeitstests in der Praxis (vgl. Hell et al., 2023, in diesem Heft), dass hier eine deutliche Wissenschafts-Praxis-Divergenz in der Bewertung der Verfahrensnützlich-

keit besteht. Leider können die Originalautoren die wissenschaftliche Debatte nicht weiter führen: John Hunter starb bereits 2002, Frank Schmidt dann 2021.

4. Aktualisierung von Sackett, Zhang, Berry und Lievens (2022)

Sackett, Zhang, Berry und Lievens legten 2022 (ein Preprint wurde bereits 2021 diskutiert) erstmalig eine systematische Neubewertung der metaanalytischen Gesamtlage zur Validität eignungsdiagnostischer Verfahren vor. Sie aktualisierten damit 24 Jahre nach der Publikation von Schmidt und Hunter den Überblick durch die Berücksichtigung neuerer Metaanalysen. Neben einigen Details bei der Angleichung der metaanalytischen Ergebnisdarstellungen (z. B. setzen sie jetzt kriteriumsangepasste Reliabilitätskorrekturen ein) rückt in der Arbeit speziell der Umgang mit unterschiedlichen Formen der Varianzeinschränkung bei den erfassten Prädiktor-Kriteriums-Zusammenhängen in den Vordergrund.

Wie schon erwähnt, wurden bereits in den älteren Metaanalysen standardmäßig direkte Varianzeinschränkungen korrigiert. Sie treten auf, wenn Personen auf der Grundlage des Verfahrensergebnisses eingestellt wurden und nur für diese im Bewerbungsprozess erfolgreichen Personen Kriteriumsdaten vorliegen. Üblicherweise werden diese Einschränkungen mittels Thorndikes Case-II-Formel (vgl. Schmidt & Hunter, 2015) korrigiert. Erst in neueren Metaanalysen werden auch indirekte Varianzeinschränkung berücksichtigt, die entstehen, wenn zusätzliche Informationen (= weitere eignungsdiagnostische Verfahren) die Selektionsentscheidung beeinflussen und/oder als geeignet diagnostizierte Personen das Stellenangebot ablehnen. Soweit hierzu weitere Daten vorliegen, kommt üblicherweise eine von Hunter, Schmidt und Le (2006) entwickelte Formel (Case IV) zum Einsatz.

Sackett et al. (2022) stellen in ihrer Arbeit zunächst eine Simulationsstudie vor (Table 1 in ihrem Artikel), in der sie die Auswirkungen variierender Randbedingungen auf die Korrekturen indirekter Varianzeinschränkungen gemäß der Formel von Schmidt et al. (2006) untersuchen. Hierbei zeigt sich folgendes:

- Der Einfluss der indirekten Varianzeinschränkung auf den Zusammenhang r_{X1Y}, verursacht durch eine Vorauswahl auf Basis eines anderen Prädiktors X2, ist häufig vernachlässigbar.
- Auch konkurrente Validierungsstudien mit bestehenden Stelleninhabern, die mit anderen eignungsdiagnostischen Verfahren ausgewählt wurden, sind häufig nicht substantiell betroffen.

Sackett et al. (2022) kommen zu dem Schluss, dass in der Vergangenheit für viele Studien zu pauschal eine Korrektur im Sinne der direkten Varianzeinschränkung angewandt wurde. Resultat war eine deutliche Überkorrektur. Als Beispiel soll die bereits erwähnte Metaanalyse von Huffcutt et al. (2014) zur Validität von (un-)strukturierten Einstellungsgesprächen dienen. In Tabelle 2 (eigene Darstellung nach Sackett et al., 2022, p. 2056) sind hier die Originalangaben von Huffcutt et al. den neuen Schätzungen von Sackett et al. gegenübergestellt. Während die Koeffizientenschätzungen nach den ersten beiden Korrekturschritten (Stichprobengewichtung der Primärstudienergebnisse, Korrektur der Unreliabilität des Kriteriums) noch eine vergleichbare Höhe aufweisen, divergieren sie deutlich im letzten Schritt nach der Korrektur der Varianzeinschränkung: Huffcutt et al. nehmen an, dass bei einer Vielzahl der berücksichtigten Primärstudien starke Selektionseffekte mit einem deutlichen Varianzeinschränkungseffekt auftraten. Als Beispiel kann die Teilgruppe der konkurrenten Validierungsstudien dienen. Bei diesem Studiendesign werden bestehende Stelleninhaberinnen und Stelleninhaber in simulierten Einstellungsgesprächen befragt. Für die kriteriumsbezogenen Validitätsanalysen werden aktuelle berufliche Leistungsdaten herangezogen. Huffcutt et al. gehen pauschal davon aus, dass von den ursprünglichen Bewerbenden für diese Zielpositionen 50% über eine Vorauswahl mittels kognitiver Fähigkeitstests ausgeschlossen worden waren, von den restlichen Personen sollten 10% in einem finalen Auswahlgespräch ein Stellenangebot erhalten haben. Dadurch ergibt sich eine sehr strenge Selektionsquote von 5%. Es resultiert bei beiden Gesprächsvarianten eine sehr deutliche Korrektur (ρ = .36 für unstrukturierte Gespräche und beeindruckende ρ = .70 für strukturierte Gespräche). Mit Bezug zu ihrer Simulationsstudie stufen Sackett et al. hingegen die Selektionseffekte bei konkurrenten Studien als vernachlässigbar ein und führen nur für prädiktive Studien eine Varianzkorrektur durch. Nach der stichprobengewichteten Aggregation verharren die Koeffizientenschätzungen praktisch auf dem Niveau der vorlaufenden Korrekturen (ρ = .18 und ρ = .45).

5. Vergleich der Übersichten 1998 und 2022

In Tabelle 3 werden die Schätzungen von Hunter und Schmidt (1998) den neueren Berechnungen von Sackett et al. (2022) gegenübergestellt. Hierbei werden die eignungsdiagnostischen Verfahren wieder getrennt nach den konzeptionellen Zugängen (Eigenschaft, Simulation und Biografie) aufgelistet.

Tabelle 2

Gegenüberstellung der Koeffizientenschätzungen aus Huffcutt et al. (2014) und Sackett et al. (2022) zur Validität von (un-)strukturierten Einstellungsgesprächen

	Stichproben-gewichtete mittleres *r*	**Zstl. Korrektur der Unreliabilität des Kriteriums**	**Zstl. Korrektur der Varianzeinschränkung im Prädiktor**
	Korrekturen bei Huffcutt et al. (2014)		
Unstrukturierte Einstellungsgespräche	.13	.18	.36
Strukturierte Einstellungsgespräche	.36	.49	.70
	Korrekturen bei Sackett et al. (2022)		
Unstrukturierte Einstellungsgespräche	.13	.17	.18
Strukturierte Einstellungsgespräche	.36	.46	.45

Anmerkungen: Sackett et al. haben die ursprünglich von Huffcutt et al. unterschiedenen Strukturierungslevel 1 und 2 in der Kategorie „unstrukturierte Einstellungsgespräche" sowie die Level 3 und 4 zur Kategorie „strukturierte Einstellungsgespräche" zusammengefasst und die entsprechenden Ergebnisse stichprobengewichtet gemittelt. In der Gesamtübersicht berücksichtigen Sackett et al. zusätzliche Daten aus einer weiteren Metaanalyse. Dadurch werden dort leicht abweichende Koeffizienten angegeben (.19 und .42).

Zunächst einmal ist festzuhalten, dass Sackett et al. (2022) in einem umfangreichen Tabellenwerk deutlich mehr Detailinformationen zu den einzelnen Validitätsschätzungen liefern. In Tabelle 3 werden diese Informationen auszugsweise gezeigt. Neben der stichprobengewichteten Punktschätzung *r* und der (nach Kriteriumsunreliabilität und identifizierter Varianzeinschränkung korrigierten) Validität ρ wird auch die verbliebene Variabilität der Schätzungen angegeben über die Standardabweichung *SD* und der unteren Grenze des 80%-Credibility-Intervalls. Wenn *SD* groß ist, ist der mittlere Wert nur begrenzt aussagekräftig. Wenn die untere Grenze des Schätzintervalls noch unter 0 liegt, muss die ermittelte Validität mit deutlicher Vorsicht betrachtet werden. Vor dem Hintergrund der häufig unklaren Varianzeinschränkungssituation wurde auf die Post-hoc-Berechnung von inkrementellen Validitäten von Verfahrenskombinationen verzichtet.

Die über die unterschiedlichen Metaanalysen hinweg angepassten Koeffizienten zeigen in Teilbereichen ein deutlich anderes Bild im Vergleich zu Schmidt und Hunter (1998):

- Die Validitäten fallen über die unterschiedlichen einungsdiagnostischen Verfahren hinweg deutlich homogener und niedriger aus. Typisch sind korrigierte Koeffizienten zwischen .20 und .40 für die Mehrheit der berücksichtigten Prädiktoren. Alle Zugänge (Eigenschaft, Simulation, Biografie) weisen Verfahren mit zufriedenstellenden Einzelkoeffizienten auf.
- Kognitive Fähigkeitstests sind nicht mehr die prognosestärksten Prädiktoren (ρ = .31). Spitzenreiter sind nunmehr strukturierte Einstellungsgespräche (ρ = .42).
- Es liegen speziell im Bereich der eigenschaftsorientierten nicht-kognitiven Verfahren Jahrzehnte nach Schmidt und Hunter deutlich mehr Befunde vor. Während damals beispielsweise nur die Validität von Gewissenhaftigkeitstests berichtet wurde, sind jetzt alle Big5-Merkmale in unterschiedlichen Erhebungsformen (allgemein oder kontexualisiert, d. h. berufsbezogen ausgestaltet) erfasst. Hier weist die kontextualisierte Version von Gewissenhaftigkeit im Vergleich mit den anderen Merkmalen die höchste Validität auf (ρ = .25) und die nicht kontextualisierten Big5-Merkmale liegen in ihrer Validität alle unter ρ = .20. Zudem erzielen auch andere Konzepte, namentlich fähigkeitsbezogen gestaltete Emotionale Intelligenzverfahren (ρ = .30) oder auch Interessentests (ρ = .24) bedeutsame Werte.
- Bei den Simulationen muss die Schätzung zu Arbeitsproben aufgrund neuerer metaanalytischer Befunde deutlich nach unten korrigiert werden (ρ = .33). Assessment Center weisen eine substantielle, aber vergleichsweise geringe Validität auf (ρ = .26).
- Bei den Verfahren mit überwiegend biografieorientierten Inhalten ergibt sich ein deutlicher Abstand zwischen strukturierten und unstrukturierten Einstellungsgesprächen (ρ = .42 vs. ρ = .19). Bei Biografischen Fragebogen wird genauer differenziert nach dem Konstruktionsprinzip: Auf der Grundlage empirischer Kriteriumsdaten konzipierte Verfahren erzielen deutlich höhere Werte als Verfahren, die theoriebezogen nach rationalen Konstruktionsprinzipien gestaltet wurden (ρ = .22 vs. ρ = .38).

Diese Berechnungen wurden zwischenzeitlich von Sackett, Zhang, Berry und Lievens (2023) noch einmal an zwei Stellen korrigiert: Nach einer neueren Metaanalyse erzielen kognitive Fähigkeitstests nur noch eine Validität von ρ = .23, während der Befund für Assessment Center aufgrund einer angepassten Varianzeinschränkungskorrektur etwas höher ausfällt (ρ = .33; siehe hierzu auch den Beitrag von Kleinmann, 2023, in diesem Heft). Es bleibt abzuwarten, wohin sich der weitere wissenschaftliche Diskurs unter anderem aufgrund dieser Publikation entwickeln wird. Eine rege und aktive Diskussion zu Befunden, die bisher jahrzehntelang als gesichert galten, ist aber absehbar.

Tabelle 3

Gegenüberstellung der Validitätsschätzungen von Schmidt und Hunter (1998) sowie Sackett et al. (2022)

		Schmidt & Hunter (1998)	Sackett et al. (2022)			
	Eignungsdiagnostisches Verfahren	**Val ρ**	**Mittleres *r***	**Val ρ**	***SD* (ρ)**	**Unterer 80%-CV**
Eigenschaft	Allgemeine kognitiver Fähigkeitstests	**.51**	.24	**.31**	.14	.13
	Berufskenntnistests	**.48**	.31	**.40**	.13	.23
	Integritätstests	**.41**	.18	**.31**	.20	.05
	EI-Verfahren (persönlichkeitsbasiert)		.23	**.30**	.17	.08
	EI-Verfahren (fähigkeitsbasiert)		.17	**.22**	.05	.16
	Gewissenhaftigkeit (allgemein)	**.31**	.16	**.19**	.15	.02
	Gewissenhaftigkeitstest (kontextualisiert)		.19	**.25**	.00	.25
	Extraversion (kontextualisiert)		.16	**.21**	.08	.11
	Extraversion (allgemein)		.08	**.10**	.13	-.06
	Verträglichkeit (kontextualisiert)		.15	**.19**	.13	.02
	Emotionale Stabilität (allgemein)		.07	**.09**	.08	-.01
	Emotionale Stabilität (kontextualisiert)		.18	**.23**	.10	.10
	Offenheit für Erfahrung (allgemein)		.04	**.05**	.07	-.03
	Offenheit für Erfahrung (kontextualisiert)		.09	**.12**	.00	.12
	Verträglichkeit (allgemein)		.08	**.10**	.14	-.08
	Verträglichkeit (kontextualisiert)		.15	**.19**	.13	.02
	Interessen	**.10**	.17	**.24**	.25	-.08
Simulation	Arbeitsproben	**.54**	.26	**.33**	.09	.21
	Assessment Center	**.37**	.20	**.29**	.09	.17
	SJTs (wissensorientiert)		.20	**.26**	.10	.13
	SJTs (verhaltensorientiert)		.20	**.26**	.12	.11
Biografie	Strukturierte Einstellungsgespräche	**.51**	.32	**.42**	.19	.18
	Unstrukturierte Einstellungsgespräche	**.38**	.13	**.19**	.16	-.01
	Biodata-Fragebogen (rational kodiert)		.17	**.22**	.06	.14
	Biodata-Fragebogen (empirisch kodiert)	**.35**	.30	**.38**	.09	.26
	Berufserfahrung (in Jahren)	**.18**	.06	**.07**	.11	-.07

Anmerkungen: EI = Emotionale Intelligenz; SJTs = Situational Judgement Tests; Mittleres r = stichprobengewichteter mittlerer Zusammenhang; Val ρ = Validitätsschätzung nach Korrektur der Kriteriumsunreliabilität und ggf. Varianzeinschränkungskorrektur; SD (ρ) = Verbleibende Standardabweichung der korrigierten Validitätsschätzungen, Unterer 80%-CV = Untere Grenze des 80%-Credibility-Intervalls

6. Fazit

24 Jahre nach der Publikation von Schmidt und Hunter führten Sackett et al. (2022) eine Aktualisierung und Neubewertung zur Validitätsbefundlage von eignungsdiagnostischen Verfahren durch. Neben der Integration neuerer Befunde steht dabei die Korrektur von Varianzeinschränkungen im Zentrum ihrer Arbeit. Gemäß eigener Simulationsergebnisse und einer kritischen Neubetrachtung der Erhebungssituation in den erfassten Primärstudien der früheren Jahre kommen Sackett et al. zu dem Befund, dass Varianzeinschränkungskorrekturen vielfach nicht adäquat eingesetzt wurden und oftmals zu großzügig korrigiert wurde. Bei ihrer Aktualisierung gingen sie nach eigener Aussage gemäß einem „Prinzip der konservativen Schätzung“ vor, d. h. sie unterließen im Zweifelsfall eine solche Korrektur.

Prinzipiell leistet die wissenschaftsbasierte berufliche Eignungsdiagnostik danach einen guten (aber bei der Betrachtung der erzielten Effektstärken aber durchaus ausbaufähigen) Beitrag zur Prognose des beruflichen Erfolgs. Im Verfahrenskanon können tätigkeitsspezifisch konzipierte Verfahren (wie strukturierte Einstellungsgespräche, Arbeitsproben oder biografische Fragebogen) gute Validitäten vorweisen. Laut Sackett et al. (2023) rücken damit simulations- und biografieorientierte Verfahren wieder mehr in den Vordergrund. Sie erfordern aber gleichzeitig mehr spezifischen Konstruktions- und Durchführungsaufwand als sofort einsetzbare eigenschaftsorientierte Verfahren wie kognitive Fähigkeitstests und persönlichkeitsorientierte Verfahren, die mit einer hervorragenden Kosten-Nutzen-Relation überzeugen (auf das Problem der ungleichen Chancen unterschiedlicher Bewerbersubgruppen bei den verschiedenen Verfahrensgruppen kann an dieser Stelle nur hingewiesen werden; vgl. dazu die Ausführungen bei Sackett et al., 2022, und zu einem möglichen Umgang damit in der Praxis beispielsweise Leenen, Stumpf & Höft, 2021).

Die Befundübersicht stützt den gelebten Praxisalltag, bei dem Einstellungsgespräche eine exponierte Position haben. Laut der Befundlage sollte der diagnostische Teil aber konform zu früheren Empfehlungen möglichst strukturiert erfolgen (siehe auch Schulte & Hiltmann, 2023, in diesem Heft). Beim Einsatz anderer Verfahren muss kritisch überlegt werden, wie und ob der allgemeine Validitätsbefund auf die eigene Erhebungssituation übertragen werden kann. Hier spielen bestehende Vorselektionen der Bewerbendengruppe genauso eine Rolle wie Anforderungen der spezifischen Tätigkeit oder relevante berufliche Erfolgskriterien (Schmidt & Hunter sowie Sackett et al. konzentrieren sich vorrangig auf Vorgesetztenurteile zur beruflichen Leistung).

Generell ist die eignungsdiagnostische Welt durch Sackett et al. (2022) nicht einfacher geworden, die Befunde deuten vielmehr darauf hin, dass sorgfältig konzipierte Verfahren und ein überlegtes Vorgehen bei der Gestaltung eignungsdiagnostischer Auswahlsysteme Schlüsselmerkmale für eine erfolgreiche Arbeit sind. Die unterschiedlichen Beiträge in diesem Themenheft geben hierzu wertvolle Hinweise.

Literatur

Döring, N. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften*. Berlin: Springer.

Höft, S. (2014). Erfolgsüberprüfung personalpsychologischer Arbeit. In H. Schuler & U. P. Kanning (Hrsg.), *Lehrbuch der Personalpsychologie* (3. Aufl., S. 1081-1136). Hogrefe.

Höft, S. & Schuler, H. (2019). Personalmarketing und Personalauswahl. In H. Schuler & K. Moser (Hrsg.), *Lehrbuch Organisationspsychologie* (6., Aufl., S. 47-108). Hogrefe.

Huffcutt, A. I., Culbertson, S. S., & Weyhrauch, W. S. (2014). Moving Forward Indirectly: Reanalyzing the validity of employment interviews with indirect range restriction methodology: Employment interview validity. *International Journal of Selection and Assessment*, *22*(3), 297-309.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594-612.

Kleinmann, M. (2023). Was schmeckt besser, Äpfel oder Birnen? Welche Vergleiche in der Eignungsdiagnostik Sinn machen. *Wirtschaftspsychologie, 25*(2), 96-106.

Leenen, R., Stumpf, S. & Höft, S. (2021). Adverse Impact und Kulturfairness in der Personalauswahl: Empirische Befundlage und Leitlinien für die Praxis. *Wirtschaftspsychologie, 23*(4), 94-105.

Moser, K., & Gunnesch-Luca, G. (2023). Kriteriumsvalidierung von Personalauswahlverfahren als Arena sozialer Normen. *Wirtschaftspsychologie, 25*(2), 107-119.

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.

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2023). Revisiting the design of selection systems in light of new findings regarding the validity of widely used predictors. *Industrial and Organizational Psychology, 1-18*. [Preprint]

Schmidt, F.L. (2002): The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1), 187-210.

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.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162–173.

Schmidt, F. L. & Hunter, J. E. (2015). *Methods of meta-analysis.* Sage.

Schmidt. F.L., Le, H. Oh, I.-S. & Shaffer, J. (2007). General mental ability, job performance, and red herrings: Responses to Osterman, Hauser, and Schmitt. *Academy of Management Perspectives, 21*(4), 64-76.

Schuler, H. & Höft, S. (2001). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (1. Aufl., S. 93-133). Hogrefe.

Schulte, S. & Hiltmann, M. (2023). Mit welchen Strukturierungselementen sich die Validität von Interviews steigern lässt. *Wirtschaftspsychologie, 25*(2), 185-196.

Corresponding Author:
Prof. Dr. Stefan Höft
Hochschule der Bundesagentur für Arbeit (HdBA)
Seckenheimer Landstr. 16
68163 Mannheim
stefan.hoeft@hdba.de