

Allgemeine Intelligenz und beruflicher Erfolg in Deutschland

Vertiefende und weiterführende Metaanalysen

Jochen Kramer

Zusammenfassung. Die Zusammenhänge von Intelligenz (*general mental ability*, GMA) mit den Karriereerfolgsindikatoren *Einkommen* und *berufliche Entwicklung* werden erstmals anhand von Studien aus Deutschland spezifiziert. Darüber hinaus stellen die vorliegenden Metaanalysen Updates früherer Analysen dar, die die Validität von GMA in Bezug auf subjektiv bewertete *Arbeitsleistung* und berufsbezogene *Lernleistung* in Deutschland untersucht haben. Die Anzahl der inkludierten Studien zu beiden Leistungsbereichen konnte mehr als verdoppelt werden. Die vorliegenden Analysen werden nach der Methode von Hunter und Schmidt (2004) vorgenommen. Methodische Neuerungen sind die Berücksichtigung insignifikanter nicht berichteter Korrelationen mit Hilfe eines Substitutionsverfahrens und die indirekte Korrektur der Varianzeinschränkung im Prädiktor. Dadurch werden akkuratere Schätzungen der Validität von GMA im Hinblick auf die zwei Leistungsbereiche ermöglicht. Die Ergebnisse zur *operationalen Validität* von GMA in Bezug auf subjektiv bewertete Arbeitsleistung (operationale Validität = .66, $k = 18$, $N = 2\,739$), berufsbezogene Lernleistung (.62, $k = 210$, $N = 30\,451$), das Einkommen (.35, $k = 7$, $N = 1\,980$) und die berufliche Entwicklung (.33, $k = 9$, $N = 1\,817$) in Deutschland werden auf mögliche Veröffentlichungs- und Quellenverzerrungen hin untersucht. Implikationen der Befunde für die weitere Forschung (z. B. zu welchen Testverfahren, Kriterien und Berufsgruppen Validierungsstudien fehlen) und die Praxis der Personalauswahl werden besprochen.
Schlüsselwörter: Intelligenz, Arbeitsleistung, berufsbezogene Lernleistung, Berufserfolg, Meta-Analyse

General mental ability and occupational success in Germany: Further metaanalytic elaborations and amplifications

Abstract. For the first time, the relationships between general mental ability (GMA) and indicators of career success (income and advancement) in a sample of German studies are examined. Furthermore, the current study expands the scope of prior analyses on the validity of GMA regarding job performance ratings and training performance ratings in Germany. Compared to prior meta-analyses more than twice the number of primary studies were included for each performance domain. The statistical procedure of Hunter and Schmidt (2004) was used for the analyses. Two methodical improvements were applied: with the help of a substitution procedure the not reported values of insignificant correlations were included and the range restriction of GMA was indirectly corrected. This allows a more accurate estimation of the validity of GMA regarding each performance domain. The operational validities of GMA in respect of job performance ratings (operational validity = .66, $k = 18$, $N = 2\,739$), training performance ratings (.62, $k = 210$, $N = 30\,451$), income (.35, $k = 7$, $N = 1\,980$), and advancement (.33, $k = 9$, $N = 1\,817$) in Germany were tested for possible publication and source biases. Implications are discussed for further research (e.g., lack of primary studies in respect of different GMA tests, occupational success criteria, and occupational groups) and for practical application in personnel selection.

Key words: intelligence, job performance, training performance, career success, meta-analysis

In mehreren Metaanalysen wurde gezeigt, dass allgemeine mentale Fähigkeiten (*general mental abilities*, GMA) deutlich mit verschiedenen Kriterien beruflichen Erfolgs korreliert sind (vgl. für eine Übersicht Bobko, Roth, Potosky, 1999; Campion in Morgeson, Campion, Dipboye, Hollenbeck, Murphy & Schmitt, 2007; Kramer, 2009). F. L. Schmidt und Hunter (1998) verglichen die prädiktive Validität verschiedener Personalauswahlverfahren und kamen

zu dem Schluss, dass GMA der beste Prädiktor für berufsbezogene Lernleistung (*training performance*) und Arbeitsleistung (*job performance*) darstellt. Darüber hinaus konnten Ng, Eby, Sorensen and Feldman (2005) zeigen, dass GMA prädiktive Validität in Bezug auf das *Einkommen* aufweist. Die Prädiktion von beruflichem Erfolg stellt laut Borkenau et al. (2005) inzwischen ein wesentliches Forschungsfeld der Persönlichkeitspsychologie dar. Die Ergebnisse dazu basierten bis zur Jahrtausendwende jedoch nahezu ausschließlich auf Studien aus den USA. Ob sie auf andere Kulturen generalisierbar sind, wurde in Frage gestellt (Bertua, Anderson & Salgado, 2005; Salgado, Anderson, Moscoso, Bertua & de Fruyt, 2003 a): In Deutschland z. B. werden andere Intelligenztests verwendet (Strukturtests statt Omnibusverfahren; vgl. F. L.

Diese Arbeit entstand an der Abteilung für Arbeits-, Organisations- und Wirtschaftspsychologie des Institutes für Psychologie der Universität Bonn. Ich danke Herrn Prof. Gerhard Blickle, Bonn, für seine Unterstützung bei der Entstehung dieser Arbeit und den zahlreichen Personen, die unpublizierte Studien oder weitergehende Informationen zu publizierten Studien zur Verfügung gestellt haben.

Schmidt & Hunter, 2004) und andere Leistungskriterien eingesetzt: Engagement und Kooperativität (umfeldbezogene Leistung nach Schmitt, Cortina, Ingerick & Wiechmann, 2003) spielen eine größere Rolle bei der Bewertung von Arbeitsleistung (Bertua et al., 2005), das duale Ausbildungssystem führt dazu, dass Berufsschulnoten und betriebliche Beurteilungen in die Bewertung von Ausbildungsleistung einfließen. Schließlich führt das dreigliedrige Schulsystem zu einer starken Vorselektion der Bewerber (Schuler & Höft, 2006) und damit einer eingeschränkteren Varianz von GMA. Dies wirkt sich verringernd auf die beobachtbare Validität von GMA aus (Hunter & Schmidt, 2004).

Diese Überlegungen gaben Anlass zu Replikationen der Metaanalysen mit Studien aus verschiedenen europäischen Ländern (z. B. Salgado et al., 2003a), Großbritannien (Bertua et al., 2005) sowie Deutschland (z. B. Hülshager, Maier & Stumpp, 2007). Sie zeigen, dass Intelligenz auch in diesen Kulturen einen validen Prädiktor beruflicher Leistung darstellt (Hülshager & Maier, 2008). Um den Nutzen von Personalauswahlinstrumenten bestimmen zu können, reicht es jedoch nicht aus zu wissen, dass GMA valide ist. Vielmehr ist es notwendig, die Höhe der Validität zu kennen (Hunter, 1986). Das Ziel der vorliegenden Studie ist es deshalb, die operationale Validität von GMA in Bezug auf subjektive *Berufsleistungs-* und objektive *Berufserfolgskriterien* in Deutschland einzuschätzen. Die Akkuratheit der Schätzung hängt dabei maßgeblich davon ab, dass möglichst viele der in Deutschland durchgeführten Studien berücksichtigt werden; ideal wäre eine Vollerhebung aller Studien. Die vorliegende Studie wurde parallel zu der Metaanalyse von Hülshager et al. (2007) durchgeführt und kann in verschiedener Hinsicht als deren Update angesehen werden: Es konnten mehr als doppelt so viele Studien zu berufsbezogener Lernleistung und Arbeitsleistung inkludiert werden. Zusätzlich werden erstmals für Deutschland objektive Berufserfolgskriterien (Einkommen und berufliche Entwicklung) berücksichtigt. Dies wurde ermöglicht durch die Spezifizierung von Inklusionsregeln auf der Basis von Rahmenmodellen zu GMA und beruflichem Erfolg sowie der Substitution von Effektstärken, die wegen Insignifikanz nicht berichtet wurden. Schließlich wird ein neues, *indirektes* Verfahren zur Korrektur von Varianzeinschränkungen verwendet, von dem F. L. Schmidt, Oh und Le (2006) zeigten, dass es im vorliegenden Fall angemessener ist als die bisher verwendete *direkte* Korrekturmethode.

Rahmenmodell von GMA

Um entscheiden zu können, welche Verfahren als GMA-Tests angesehen werden können, wurde ein Rahmenmodell von GMA gewählt: das Berliner Intelligenzstrukturmodell (BIS; Jäger, 1982, 1984). Im BIS-Modell wird GMA als Integral von drei inhaltsbezogenen Fähigkeiten – verbales, figurales und numerisches Denken – und von vier operationalen Fähigkeiten – Verarbeitungskapazität, Merkfähigkeit, Bearbeitungsgeschwindigkeit und Kreativität – konzipiert.

Idealerweise enthält ein GMA-Test Aufgaben aus allen zwölf möglichen Kombinationen inhaltsbezogener und operationaler Fähigkeiten. Dies ist in den meisten GMA-Tests nicht der Fall, abgesehen vom BIS-Test (Jäger, Süß & Beauducel, 1997) und dem BIS-r-DGP (Deutsche Gesellschaft für Personalwesen e. V. [DGP], 1996), die beide auf Basis des BIS-Modells entwickelt wurden. Es wurden daher folgende Regeln spezifiziert, um zu adäquaten Inklusions- und Exklusionsentscheidungen zu kommen: 1. Die Tests sollen die BIS-Operation Verarbeitungskapazität berücksichtigen, die nach Jäger et al. (1997) für GMA besonders wesentlich ist, und mindestens zwei verschiedene BIS-Inhalte umfassen. 2. Allgemein als GMA-Tests akzeptiert und deshalb berücksichtigt werden Matrizen-Tests, obwohl sie ausschließlich figurale Verarbeitungskapazität abdecken.

Diesen Regeln entsprechend konnten auch Skalen zur Erfassung von GMA aus Eignungstestbatterien berücksichtigt werden, die neben mentalen auch andere Fähigkeiten umfassen. Sie wurden in den bereits vorliegenden Metaanalysen aus Deutschland exkludiert. Dabei handelt es sich insbesondere um den *Berufseignungstest* (BET; Schmale & Schmidtke, 1995), eine deutsche Adaptation der *General Aptitude Test Battery* (GATB; U. S. Employment Service, 1970), zu dem 70 Studien vorliegen.

Rahmenmodell von beruflichem Erfolg

Zur Entscheidung, welche Maße als Indikatoren für beruflichen Erfolg angesehen und inkludiert werden können, wurde ein Rahmenmodell von Dette, Abele und Renner (2004) gewählt und modifiziert: Dette et al. unterscheiden die Erfolgsbereiche *spezifische Arbeit* und *globale Laufbahn*. Ergänzend dazu wird der Erfolgsbereich *berufsbezogener Lernleistung* betrachtet. Berufsbezogene Lernleistung kann als Mediator zwischen GMA und Arbeitsleistung angesehen werden, da GMA als die Fähigkeit zu Lernen definiert werden kann (Thorndike nach Hunter & Schmidt, 1996). Erfolgreiches Lernen resultiert in umfassenderem Fachwissen, mit dem dann Arbeitsleistung prädiagnostiziert wird (F. L. Schmidt, Hunter & Outerbridge, 1986).

Zu jedem Erfolgsbereich können objektive Maße (neutrale Kennzahlen) und subjektive Einschätzungen (Vergleiche mit Normen, Zielen oder anderen Personen) als Indikatoren dienen. Beide Datenarten können folgenden drei Datenquellen entstammen: Fremdbeurteilungen, Selbstberichten oder Akten bzw. Dokumenten. Die sich daraus ergebenden 18 möglichen Indikatorgruppen werden in der Praxis nicht alle verwendet, sind unterschiedlich aussagekräftig und zum Teil nur gering miteinander korreliert (vgl. Dette et al., 2004). Aus diesen Gründen wurden vier hinreichend homogene Indikatorgruppen gebildet und in getrennten Metaanalysen untersucht: 1. *spezifische Arbeitsleistung*, gemessen mit Fremdbeurteilungen, 2. *berufsbezogene Lernleistung*, die Ausbildungs- und Weiterbildungsleistungen in Form von Berufsschulnoten und betrieblichen Beurteilungen umfasst, 3. *Einkommensmaße*, z. B. das Einkommen zu einem be-

stimmten Zeitpunkt, die Einkommensentwicklung oder die Tarifgruppe und 4. *berufliche Entwicklung*, die Beförderungs- und Positionsindikatoren umfasst.

Substitution insignifikanter, nicht berichteter Effektstärken

Üblicherweise werden Studien exkludiert, wenn relevante Korrelationen erhoben, aber nicht berichtet werden. Alternativ dazu können die fehlenden Werte substituiert werden, z. B. mit Null (vgl. Funke, Krauß, Schuler & Stapf, 1987) oder mit regressionsanalytisch gewonnenen Substituten (vgl. Rustenbach, 2003). Die Exklusion der Studien führt i. d. R. zu einer Überschätzung von Effekten, die deutlich positiv von Null verschieden sind, da systematisch *insignifikante* Studien nicht berücksichtigt werden. Das Gegenteil ist bei der Substitution mit Null der Fall, die i. d. R. zu einer Unterschätzung führt, wenn der wahre Effekt positiv von Null verschieden ist. Regressionsanalytische Substitutionsverfahren können nur bei zufällig, nicht aber bei systematisch fehlenden Werten eingesetzt werden. Einen Überblick über die verschiedenen gängigen Substitutionsverfahren gibt z. B. Rustenbach (2003).

In der vorliegenden Metaanalyse sollen möglichst alle erhältlichen Informationen zur Validität von GMA genutzt werden, auch wenn sie nur in dichotomer Form angegeben sind (signifikant vs. nicht signifikant), und dabei jede systematische Über- oder Unterschätzung vermieden werden. Um dies zu erreichen, wird ein Substitutionsverfahren eingesetzt, das auf den *berichteten insignifikanten* Korrelationen basiert (s. u.).

Indirekte Korrektur der Varianzeinschränkung im Prädiktor

In der zweiten Auflage ihres Handbuches zur Validitätsgeneralisierung stellen Hunter und Schmidt (2004) eine Möglichkeit vor, die Varianzeinschränkung im Prädiktor auch dann adäquat zu korrigieren, wenn die Selektionsvariable nicht mit der Prädiktorvariablen (GMA) identisch ist. Diese – als *indirekte* Korrektur bezeichnete – Prozedur berücksichtigt, dass Messfehler und Varianzeinschränkung im Prädiktor abhängig voneinander sein können. Anhand eines konkreten Beispiels zeigen Hunter, Schmidt und Le (2006), dass die Verwendung der direkten statt der indirekten Korrektur zu einer Unterschätzung des wahren Effektes um ca. 25 % führen kann.

Fragestellung

Welche operationale (d. h. um Varianzeinschränkung im Prädiktor und Messfehler im Kriterium korrigierte) Validität weist GMA auf, wenn a) auch Operationalisierungen von GMA verwendet werden, die in den bisherigen Metaanalysen nicht berücksichtigt wurden, b) neben subjektiv

bewerteter Arbeitsleistung und berufsbezogener Lernleistung auch die Karriereerfolgsindikatoren Einkommen und berufliche Entwicklung betrachtet werden, c) Studien, die insignifikante Korrelationen nicht berichten (nach Substitution der betroffenen Werte) ebenfalls inkludiert werden und d) die Varianzeinschränkung im Prädiktor indirekt statt direkt korrigiert wird?

Methode

Die maßgeblichen methodischen Entscheidungen zur Durchführung der Metaanalyse werden im Folgenden beschrieben. Eine umfassende Darstellung der methodischen Vorgehensweise findet sich bei Kramer (2009).

Literaturrecherche

Zur Suche nach geeigneten Studien wurden verschiedene Strategien angewendet: Es wurden Psyndex-Recherchen nach veröffentlichten Studien der Jahrgänge 1977 bis 2006 durchgeführt. Gesucht wurde 1. in den Zusammenfassungen der Artikel nach Kombinationen von Stichworten, die sich auf GMA (Fähigkeit, Intellekt, Intelligenz, IQ, kognitiv, mental), Kriterien (Arbeitsleistung, Beförderung, berufliche Entwicklung, beruflicher Erfolg, Beurteilung, Bewertung, Eignung, Einkommen, Gehalt, Karriere, Laufbahn, Lohn, Platzierung, Produktivität, Qualifikation, Tauglichkeit) und die gewünschte Stichprobe (Ausbildung, Beruf, Bewerber, Personal, Umschulung, prognostische Validität, Weiterbildung) beziehen, 2. nach weiteren Veröffentlichungen der Autoren, die mehrfach zum Thema veröffentlicht haben (Althoff, Funke, Kersting, Marschner, Schmidt-Atzert, Schuler, Thielepape), 3. mithilfe von Psyndex-Schlagworten, denen die gefundenen Artikel häufig zugeordnet waren (z. B. Intelligenz, prädiktive Validität) und 4. nach Studien, die die gängigen Intelligenztests verwendet haben. Darüber hinaus wurden 5. die Manuale der gängigen Intelligenztests nach Validierungsstudien durchgesehen, 6. die Quellen berücksichtigt, die in bereits vorliegenden Metaanalysen mit deutschen Studien verwendet worden sind (Hülshager, et al., 2007; Hülshager, Maier, Stumpp & Muck, 2006; Salgado & Anderson, 2003; Salgado et al., 2003 a; Salgado et al., 2003 b), 7. die Literaturverzeichnisse der geeigneten Quellen nach weiteren Artikeln durchsucht und 8. verschiedene Institutionen angeschrieben, von denen angenommen wurde, dass sie über unveröffentlichtes Material verfügen. Dies waren im Einzelnen die arbeits- und organisationspsychologischen Fachgruppen des Berufsverbandes Deutscher Psychologinnen und Psychologen (BDP) sowie der Deutschen Gesellschaft für Psychologie (DGPs), die Deutsche Gesellschaft für Personalführung (DGFP), die Deutsche Gesellschaft für Personalwesen (DGP), die Deutsche Gesellschaft für Luft- und Raumfahrt (DLRA), die Sparkassenakademie, die Gruppe Wehrpsychologie des Streitkräfteamtes sowie das Zentrum für Psychologische Information und Dokumentation (ZPID).

Inklusionskriterien

Zur Entscheidung, ob eine Studie berücksichtigt werden konnte, wurden Inklusions- bzw. Exklusionsregeln definiert, die 1. die Intelligenzmaße, 2. die Erfolgskriterien, 3. die untersuchte Stichprobe und 4. die statistischen Zusammenhangsmaße betreffen. Als Maße für *GMA* wurden anerkannte Intelligenztests angesehen, die in einschlägigen Überblickswerken zur Intelligenzdiagnostik besprochen werden (Brambring, 1983; Brickenkamp, 2002; Holling, Preckel & Vock, 2004; Kanning, 2002; Sarges & Wottawa, 2004). Testverfahren, die dort nicht beschrieben sind, wurden berücksichtigt, wenn die Testitems dem BIS-Modell zugeordnet werden konnten. Als hinreichend relevant wurde eine Operationalisierung angesehen, wenn sie die BIS-Operation Verarbeitungskapazität mit unterschiedlichen Aufgabeninhalten abdeckt. Kontaminationen mit Testitems, die sich nicht dem BIS-Modell zuordnen lassen, wurden in begrenztem Umfang akzeptiert, insbesondere dann, wenn der fragliche Test hoch mit anerkannten Intelligenzmaßen korreliert. Das Gewicht der Kontaminationen an den Metaanalysen ist gering (zwischen 0.0 und 6.4%; vgl. Kramer, 2009).

Als Indikatoren für *berufsbezogene Lernleistung* wurden Ausbildungs- und Weiterbildungsleistungen akzeptiert, die in Form von Berufsschulnoten oder betrieblichen Beurteilungen durch Vorgesetzte vorliegen. Nicht berücksichtigt wurden Studienleistungen sowie objektive Maße, z. B. die Ausbildungsdauer oder Fehlerquoten.

Zur *spezifischen Arbeitsleistung* liegen überwiegend subjektive Fremdeinschätzungen vor, objektive Maße werden selten berichtet. Sie wurden deshalb exkludiert. Selbstbeurteilungen werden ebenfalls exkludiert, da sie in der personalpsychologischen Praxis weniger anerkannt sind als Fremdbeurteilungen, der Monosourcebias vermieden werden kann und Fremdbeurteilungen mehr Leistungsaspekte umfassen als Selbstbeurteilungen (Hunter & Schmidt, 1996). Die Fremdbeurteilungen stammen überwiegend von Vorgesetzten und sind Einschätzungen der *typischen* Arbeitsleistung. Maximalleistungen, wie sie z. B. in Assessment Centern erhoben werden, wurden nicht inkludiert.

Als Karriereindikatoren wurden *Einkommensmaße* (Einkommen zu einem bestimmten Zeitpunkt, Einkommensentwicklung, Tarifgruppe) und Aspekte der *beruflichen Entwicklung* (Beförderungen, Hierarchiestufe, Prestige, Anzahl der Mitarbeiter, Erlaubnis zur Teilnahme an Führungsnachwuchstrainings, Führungsverantwortung) berücksichtigt.

Berücksichtigt wurden ausschließlich Sich-Bewerbende und Berufstätige aus dem früheren Deutschen Reich und der Bundesrepublik Deutschland. Studien aus Österreich, der deutschsprachigen Schweiz und der Deutschen Demokratischen Republik waren nicht zugänglich. Als Effektmaß wird die Produkt-Moment-Korrelation von Pearson verwendet. Andere Effektmaße werden nach Glass,

McGaw und Smith (1984) sowie Hunter und Schmidt (2004) in Pearson *rs* transformiert.

Kodierung

Kodiert wurden allgemeine Angaben zur Studie, Informationen zu Prädiktoren, Kriterien und den Effektstärken. Alle Kodierungen wurden unabhängig von zwei Ratern vorgenommen. Zur Prüfung der Übereinstimmung der Ratings wurden κ - und ICC_{unjust} -Koeffizienten berechnet, die im Mittel bei .93 liegen (Range = .72 bis 1.00; vgl. für das Kodierschema und die Übereinstimmungen pro Variable Kramer, 2009).

Substitution insignifikanter, nicht berichteter Korrelationen

Korrelationen, die in den Primärstudien deshalb nicht berichtet wurden, weil sie insignifikant sind, wurden in einem fünfschrittigen Verfahren substituiert: 1. Für alle *insignifikanten* Korrelationen, die *berichtet* sind (r_b), wurde der positive kritische Wert berechnet, ab dem sie statistisch signifikant wären ($\alpha = 5\%$, einseitig). 2. Die Korrelationen aus Schritt 1 wurden zu ihren kritischen Werten (r_k) ins Verhältnis gesetzt: r_b/r_k . 3. Die im zweiten Schritt errechneten Relationen wurden gemittelt. Die mittleren Relationen betragen für die Arbeitsleistungsstudien $M = .34$ ($SD = .50$, Anzahl Korrelationen = 31). Die entsprechenden Ergebnisse sind für Lernleistungsstudien $M = .39$ ($SD = .45$, $n = 707$), für Einkommensstudien $M = .46$ ($SD = .35$, $n = 8$) und für Studien zur beruflichen Entwicklung $M = .23$ ($SD = .34$, $n = 5$). 4. Für die einzelnen insignifikanten *nicht berichteten* Korrelationen wurden ebenfalls die zugehörigen kritischen Werte berechnet. Dies ist möglich, da für die Berechnung außer der Stichprobengröße, der Testseitigkeit (einseitig) und dem Signifikanzniveau (5%) keine weiteren Angaben erforderlich sind. Schließlich wurden 5. die *nicht berichteten* Korrelationen geschätzt, in dem ihr kritischer Wert jeweils mit der mittleren Relation aus Schritt 3 multipliziert wurde. Die daraus resultierenden Substitute liegen zwischen .02 und .14.

Es wird angenommen, dass diese Substitution weder zu einer systematischen *Überschätzung* (wie bei der Studienexklusion) noch zu einer systematischen *Unterschätzung* (wie bei der Substitution mit Null) führt. Kramer (2009) verglich die drei Arten mit insignifikanten nicht berichteten Korrelationen umzugehen miteinander. Er führte dazu auch Metaanalysen durch, bei denen die kritischen Werte selbst als Substitute verwendet wurden. Die mittleren Effekte dieses Vorgehens stellen eine Obergrenze dar, die nur dann erreicht werden könnte, wenn alle nicht berichteten Korrelationen gerade eben signifikant würden. Kramer (2009) zeigte, dass die *Substitution auf Basis der Relationen* zu mittleren Effekten führt, die zwischen denen bei Nullsubstitution und denen bei Substitution mit den kritischen Werten liegen. Hingegen übertreffen die mittleren Effekte bei *Exklusion* der betroffenen Studien diejenigen bei Substitution mit den kritischen Werten.

Kumulierung abhängiger Werte

Werden zu einer Stichprobe mehrere Effektstärken berichtet, sind diese zu einem Wert zu aggregieren, um eine einheitliche Gewichtung der Studien zu erreichen (Hunter & Schmidt, 2004). Sind sie hingegen statistisch unabhängig voneinander, werden die verschiedenen Effektstärken als getrennte Studien betrachtet. Die Aggregation der abhängigen Maße erfolgte durch Bildung linearer Verbände, bei denen die Korrelationen zwischen den verschiedenen GMA-Maßen bzw. Kriterien berücksichtigt werden. In den Fällen, in denen die benötigten Korrelationen nicht bekannt sind, wurden die Mittelwerte der verschiedenen Validitäten berechnet. Bei der Aggregation wurden ggf. auch Gewichtungen vorgenommen: verschiedene GMA-Tests innerhalb einer Studie wurden so aggregiert, dass sie jeweils dasselbe Gewicht am Gesamtmaß einnehmen; mehrere GMA-Subfaktoren wurden so gewichtet, dass ihr Aggregat dem Gesamtintelligenzwert möglichst genau entspricht. Im Falle mehrerer Kriterien innerhalb einer Studie wurden die Gewichte so gewählt, dass unterschiedliche Quellen dasselbe Gewicht erhalten (z. B. Berufsschulnoten und betriebliche Beurteilungen).

Artefaktkorrekturen

Die metaanalytische Vorgehensweise nach Hunter und Schmidt (2004) zeichnet sich dadurch aus, dass verschiedene Mängel der Primärstudien (sog. Artefakte) korrigiert werden können. Artefakte beeinflussen die zu beobachtende Korrelation zwischen GMA und den Kriterien. Die in den Primärstudien enthaltenen Informationen erlauben die Korrektur des Stichprobenfehlers, der Messfehler in den GMA- und Kriterienmaßen sowie der Varianzeinschränkung im Prädiktor. Einen Überblick über weitere Artefakte geben Hunter und Schmidt (2004).

Der Stichprobenfehler wird durch Gewichtung der Studien mit n korrigiert. Werden keine weiteren Korrekturen vorgenommen, liegt eine *bare-bone* Analyse vor.

Zur vollständigen Korrektur der Messfehler von GMA-Tests sind nach Hunter und Schmidt (2004) Reliabilitätskennwerte aus kombinierten Paralleltest-Retest-Designs erforderlich (*coefficient of equivalence and stability*, CES; Cronbach, 1947), die i. d. R. nicht berichtet werden. F. L. Schmidt, Le und Ilies (2003) schlagen deshalb vor, CES auf Basis des *coefficient of equivalence* (CE nach Cronbach, 1947) zu schätzen: $CES = CE - .05$. Die CE-Reliabilitäten werden in Form von Testhalbierungs-, Paralleltest-Koeffizienten oder internen Konsistenzen den Testmanualen der GMA-Tests oder weiteren Studien entnommen, die nicht varianzeingeschränkte Stichproben untersuchen (vgl. Kramer, 2009). Die Reliabilität von Prädiktorverbänden wurde gemäß Mosier (1943, nach Hunter und Schmidt, 2004) berechnet. Dabei wurden die bei der Aggregation vorgenommenen Gewichtungen ebenfalls berücksichtigt. Die mit dieser Vorgehensweise ermittelten Prädiktorreliabilitäten betragen für die Gruppe der Berliner Intelligenzstrukturtests (BIS-4, Jäger et al., 1997; BIS-r-

DGP, DGP, 1996) im Mittel .87, für die Intelligenzstruktur-Tests (u. a. IST 70, Amthauer, 1973; IST 2000 R, Amthauer, Brocke, Liepmann & Beauducel, 2001) .87, für die Matrixtests (u. a. CFT 3, Weiß, 1980; FRT, Daniels, 1962; SPM, SPMP, Raven, Raven & Court, 2000, 2003) .89, für die Wilde Intelligenztests (WIT, Jäger & Althoff, 1983, 1994; WIT-2, Kersting, Althoff & Jäger, 2008) .91 sowie für das Leistungsprüfsystem (LPS, Horn, 1962, 1983) und das Prüfsystem für Schul- und Bildungsberatung (PSB, Horn, 1969) .94. Konnte die Reliabilität für eine Testversion nicht ermittelt werden, wurde sie mit der mittleren Reliabilität der anderen Verfahren aus der entsprechenden Testgruppe geschätzt. Die CES-Reliabilitäten der Testverfahren, für die keine Informationen vorliegen (*GATB-basierte* und *andere Verfahren* mit Ausnahme der AZUBI-Tests von Görlich & Schuler, 2007 sowie Schuler & Klingner, 2005 [$r_{xx} = .89$]), wurden mit dem Gesamtmittel ($r_{xx} = .90$) geschätzt. Eine detaillierte Aufstellung der inkludierten Testverfahren findet sich bei Kramer (2009).

Zur Korrektur der Kriterienmessfehler wurden Reliabilitäten verwendet, die in früheren Studien metaanalytisch bei Berufstätigen ermittelt wurden. Die Reliabilität von Leistungsbeurteilungen durch Vorgesetzte wurde von Rothstein (1990) in Abhängigkeit von der Dauer ermittelt, mit der sich Vorgesetzte und Mitarbeiter kennen, sowie von Viswesvaran, Schmidt und Ones (2002) in Abhängigkeit von den eingeschätzten Leistungsdimensionen. Beide Metaanalysen beziffern die mittlere Reliabilität der Leistungsbeurteilungen durch Vorgesetzte auf $r_{yy} = .52$. Diese Reliabilität wurde zur Korrektur der Arbeitsleistungskriterien und der betrieblichen Lernleistungsbeurteilungen verwendet. Für berufsbezogene Lernleistungen, die Berufsschulnoten enthalten, wurde auf Arbeiten von Baron-Boldt und Kollegen zurückgegriffen (Baron-Boldt, Funke & Schuler, 1989; Baron-Boldt, Schuler & Funke, 1988), die die Reliabilität von Ausbildungsleistungen in Deutschland mit $r_{yy} = .64$ angeben. Die Karriereerfolgskriterien Einkommen und berufliche Entwicklung wurden nicht in Bezug auf Messfehler korrigiert, da es sich überwiegend um neutrale Kennzahlen handelt (vgl. Ng et al., 2005).

Schließlich konnte die Varianzeinschränkung im Prädiktor korrigiert werden. Dazu wurden die in den Studien berichteten Streuungen (SD_{Sp}) zu der Streuung in der Population der Erwerbstätigen (SD_{Pop}) in Beziehung gesetzt ($u = SD_{Sp}/SD_{Pop}$). Die Populationsstreuung wurde den Testmanualen entnommen und – wenn es sich bei der Population um die Gesamtbevölkerung handelte – wie von Sackett und Ostgaard (1994) vorgeschlagen, zur Schätzung der Populationsstreuung der erwerbstätigen Bevölkerung um 10% reduziert.

Zur Schätzung der nicht berichteten Varianzeinschränkungen wurde berücksichtigt, dass sie abhängig vom Komplexitätsniveau der Berufe sein können: Salgado et al. (2003b) zeigten anhand europäischer Studien, dass die Varianzeinschränkung von GMA zunimmt, je größer die berufliche Komplexität ist. Die Unterschiede in den Varianzeinschränkungen zwischen verschiedenen komplexen Berufen in den vorliegenden Studien aus Deutschland zeigen dies lediglich tendenziell. Die Operationalisierung der

Berufskomplexität erfolgte mit den Skill Leveln der *International Standard Classification of Occupations* (ISCO-88; International Labour Office, 1990). Diese Klassifizierung liegt für Berufe in Deutschland vor (vgl. Geis, 2006). Der mittlere u -Wert liegt für komplexere Berufe (Skill Level 3) bei .75, für weniger komplexe Berufe (Skill Level 2) bei .79. Varianzeinschränkungen für die komplexesten (Skill Level 4) und einfachsten Berufe (Skill Level 1) konnten in Ermangelung an verwertbaren Angaben nicht berechnet werden. Die Varianzeinschränkungen der Studien, für die keine Angaben vorliegen, wurden geschätzt. Dabei wurde für die komplexeren Berufe das Mittel von Skill Level 3 ($u = .75$), für die einfacheren das Mittel von Skill Level 2 ($u = .79$) und für Stichproben, die nicht eindeutig einem Skill Level zugeordnet werden können, das Gesamtmittel ($u = .78$) verwendet. Die Varianzeinschränkungen wurden mit der indirekten Methode nach Hunter und Schmidt (2004) korrigiert, die dann angemessen ist, wenn die Selektionsvariable nicht mit der Prädiktorvariable identisch ist.

Aggregation unabhängiger Effekte

Zur Aggregation der Effekte über die verschiedenen Studien hinweg, wurde die als Validitätsgeneralisierung bezeichnete Vorgehensweise von Hunter und Schmidt (2004) gewählt, die im vorliegenden Gegenstandsbereich die Übliche ist (Schulze, 2004). Um die Einflüsse der Artefaktkorrekturen auf das Ergebnis einschätzen zu können, wurde neben den Ergebnissen der bare-bone Analyse, bei der nur der Stichprobenfehler korrigiert wird, die operationale Validität berechnet. Bei der operationalen Validität wird auf eine Korrektur der Unreliabilität von GMA beim mittleren Effekt verzichtet, da die Unreliabilitäten Eigenschaften der Intelligenztests selbst sind und keine externen Einflüsse darstellen (Hunter & Schmidt, 2004). Um die Variabilität der Studieneffekte nicht zu unterschätzen, wird die Unreliabilität im Prädiktor jedoch bei der Schätzung der Standardabweichung der operationalen Validität berücksichtigt. Die Validitätsgeneralisierung nach Hunter und Schmidt wird von Schulze (2004) als *conditionally random effects procedure* bezeichnet, da abhängig von der Höhe der geschätzten Standardabweichung der operationalen Validität (\widehat{SD}_p) entweder vom Vorliegen eines Modells fester Effekte (*fixed effects model*, $\widehat{SD}_p = 0$) oder eines Modells zufallsvariabler Effekte (*random effects model*, $\widehat{SD}_p > 0$) ausgegangen wird.

Zur Interpretation der mittleren Effekte werden zwei Kenngrößen berechnet: der Anteil der durch die Artefakte aufgeklärten Varianz (VE) und die untere Grenze des 90% Glaubwürdigkeitsintervalls (CV; nicht zu verwechseln mit dem Konfidenzintervall, CI, s. u.). Beide Werte geben Hinweise auf die Generalisierbarkeit des mittleren Effektes (Kemery, Mosholder & Roth, 1987): Liegen weniger als 90% der korrigierten Studieneffekte über Null, kann die Richtung des Effektes nicht generalisiert werden. Der mittlere Effekt kann lediglich als Mittel heterogener Studien angesehen werden (*situational specificity*). Liegt die untere Grenze des 90% CV hingegen über Null, kann die Richtung des Effektes generalisiert werden (*transportability*).

Werden zudem mehr als 75% der beobachteten Varianz zwischen den Studien durch die Artefakte aufgeklärt, kann auch die Höhe des Effektes generalisiert werden (*cross-situational consistency*).

Wie akkurat die Schätzung des mittleren Effektes erfolgt, verdeutlichen die Konfidenzintervalle (CI), die nach Whitener (1990) berechnet werden. Sie geben Auskunft darüber, ob Unterschiede zu Null bzw. zu geschätzten mittleren Effekten anderer Metaanalysen als bedeutsam angesehen werden können. Auf die Berechnung von Signifikanzen wird verzichtet, denn „the point estimate and confidence interval are more informative, and less likely to be misinterpreted, than the significance test“ (Borenstein, 2005, S. 196). Darüber hinaus wurde geprüft, ob Veröffentlichungs- oder Quellenverzerrungen vorliegen können (s. u.).

Ergebnisse

Beschreibung der inkludierten Studien

Zum Erfolgsbereich *Arbeitsleistung* konnten 18 Studien identifiziert werden ($N = 2\,739$), die zwischen 1963 und 2006 durchgeführt wurden; drei davon sind nicht veröffentlicht. Von 59.8% der Gesamtstichprobe ist das Geschlecht bekannt: ein Drittel davon ist weiblich. Das mittlere Alter der Teilnehmer ist in sieben Studien angegeben, es liegt im Bereich von 17.1 bis 36.0 Jahren ($M = 25.3$, $SD = 8.4$). 33.1% verfügen über die (Fach-)Hochschulreife, 24.8% über die mittlere Reife, 0.4% verfügen maximal über einen Hauptschulabschluss. Das Bildungsniveau der verbleibenden 41.8% ist unbekannt. Die Berufsgruppen und Komplexitätsniveaus der Berufe nach ISCO-88 können ebenso wie die verwendeten Prädiktoren Tabelle 1 entnommen werden. Die relative Bedeutung der einzelnen BIS-Zellen enthält Tabelle 2. In ihr wird deutlich, dass bei allen Erfolgsbereichen alle drei BIS-Inhaltsfacetten vertreten sind, während bei den Operationen vor allem Verarbeitungskapazität erfasst wird. Die Bewertungen der Arbeitsleistung sind heterogen: es wurden zwischen einem und 40 Items einem bis drei Ratern vorgelegt. Dabei wurden verschiedene Aspekte aufgabenbezogener, adaptiver sowie umfeldbezogener Leistung eingeschätzt (vgl. Kramer, 2009). Die Validitäten von GMA sind in 14 Studien prädiktiv (Δ_i : $M = 65.1$, $SD = 27.2$ Monate), in drei Studien konkurrenz und in einer Studie retrograd ($\Delta_i = -49$ Monate). Insignifikante Effekte werden in einer Studie nicht berichtet (Gewicht an der bare-bone Analyse: 2.6%).

Mit *berufsbezogener Lernleistung* befassen sich 210 Studien ($N = 30\,451$), 190 davon thematisieren Ausbildungsleistung und 20 Weiterbildungsleistung. Überwiegend werden Berufsschulnoten berichtet (Gewicht an der bare-bone Analyse: 52.5%), seltener betriebliche Beurteilungen (9.5%). Die Studien, bei denen es unklar ist, aus welcher Quelle die Beurteilung stammt, machen 33.3% an der bare-bone Analyse aus. Die Studien stammen aus den Jahren 1928 bis 2006. Studien vor 1960 haben ein geringes Gewicht an der Analyse (0.4%). 25 Studien sind nicht ver-

Tabelle 1. Häufigkeiten und Gewichte der kategorialen Variablen

Variable	Arbeitsleistung		Lernleistung		Einkommen		berufliche Entwicklung	
	k	w (%)	k	w (%)	k	w (%)	k	w (%)
<i>Publikationsart</i>								
Testmanuale	4	29.8	81	49.0	0	0.0	0	0.0
Artikel (Zeitschriften, Beiträge in Hrsg.-Werken)	5	11.2	69	27.9	1	14.9	2	22.6
Dissertationen, Monographien (ohne Testmanuale)	6	50.9	35	10.3	1	38.7	1	35.4
nicht veröffentlicht	3	8.1	25	12.9	5	46.4	6	42.0
<i>ISCO-88 Berufs-Hauptgruppe</i>								
0 Militär	0	0	7	4.3	0	0.0	0	0.0
1 Leitende Berufe	2	4.9	1	0.2	1	3.5	1	3.0
2 Wissenschaftler	2	9.0	2	0.7	0	0.0	0	0.0
3 Techniker und Gleichrangige	4	30.7	50	17.1	0	0.0	1	2.7
4 Bürokräfte	3	37.6	35	18.9	2	48.8	2	46.0
5 Dienstleistungsberufe	0	0	2	0.5	0	0.0	0	0.0
6 Landwirtschaft/Fischerei	0	0	0	0.0	0	0.0	0	0.0
7 Handwerk	1	2.7	86	40.4	0	0.0	0	0.0
8 Maschinenbediener	2	1.1	8	6.9	0	0.0	0	0.0
9 Hilfsarbeitskräfte	0	0	0	0.0	0	0.0	0	0.0
gemischt/unklar	4	13.9	19	11.0	4	47.6	5	47.5
<i>ISCO-88 Skill Level</i>								
ohne	2	4.0	8	4.5	1	3.5	1	3.9
1 (geringe Komplexität)	0	0.0	0	0.0	0	0.0	0	0.0
2	6	41.1	138	72.0	2	48.8	2	46.0
3	4	30.7	50	17.1	0	0.0	1	2.7
4 (hohe Komplexität)	2	9.0	2	0.7	0	0.0	0	0.0
gemischt/unklar	4	13.9	12	5.8	4	47.6	5	47.5
<i>GMA-Testgruppe</i>								
BIS ^a	1	3.6	9.0 ^f	2.9	0	0.0	0	0.0
GATB ^b -basiert	0	0.0	72.7 ^f	43.1	0	0.0	0	0.0
IST ^c	9	43.8	49.8 ^f	22.7	2	25.4	2	27.6
LPS/PSB ^d	0	0.0	12.5 ^f	7.3	2	25.8	1	10.6
Matrizentests	2	1.1	12.7 ^f	3.8	0	0.0	0	0.0
WIT ^e	1	2.6	29.8 ^f	9.9	1	3.5	1	3.9
andere	5	49.0	23.5 ^f	10.4	2	45.3	5	58.0
<i>Art der Validität</i>								
prädiktiv	14	84.8	197	97.7	4	62.8	6	73.6
konkurrent	3	11.6	8	1.8	1	6.6	2	23.7
retrograd	1	3.6	1	0.3	1	14.9	0	0.0
unklar	0	0.0	4	0.1	1	15.7	1	2.7
Gesamt	18	100.0	210	100.0	7	100.0	9	100.0

Anmerkungen: ^aBerliner Intelligenzstruktur-Tests; ^bGeneral Aptitude Test Battery; ^cIntelligenzstruktur-Tests; ^dLeistungsprüfsystem/Prüfssystem für Schul- und Bildungsberatung; ^eWilde Intelligenztests; ^fMehrfachnennungen möglich: Die verschiedenen GMA-Tests werden gleichwertig berücksichtigt, die betroffenen Studien anteilig aufgeteilt.

Tabelle 2. Gewichte der BIS-Zellen in den bare-bone Analysen

BIS-Operationen	BIS-Inhalte			
	verbal	numerisch	figural	Gesamt
Verarbeitungskapazität				
Arbeitsleistung	37.4	21.4	26.7	85.5
Lernleistung	29.8	26.5	26.9	83.2
Einkommen	23.3	22.1	30.4	75.8
Entwicklung	27.2	24.4	28.3	79.8
Bearbeitungsgeschwindigkeit				
Arbeitsleistung	4.2	2.2	0.2	6.6
Lernleistung	2.6	1.4	2.2	6.2
Einkommen	12.6	4.6	3.9	21.2
Entwicklung	8.8	3.4	0.0	12.2
Einfallsreichtum				
Arbeitsleistung	0.2	0.2	0.2	0.5
Lernleistung	0.9	0.2	0.3	1.4
Einkommen	1.3	0.0	0.0	1.3
Entwicklung	0.0	0.0	0.0	0.0
Merkfähigkeit				
Arbeitsleistung	5.0	2.0	0.5	7.4
Lernleistung	3.3	1.3	0.7	5.3
Einkommen	0.0	1.8	0.0	1.8
Entwicklung	0.0	1.6	0.0	1.6
Gesamt				
Arbeitsleistung	46.8	25.7	27.5	100.0
Lernleistung	36.6	29.4	30.1	96.1 ^a
Einkommen	37.3	28.4	34.3	100.0
Entwicklung	35.9	29.4	28.3	93.6 ^a

Anmerkung: ^aDie in Bezug auf 100% fehlenden Angaben lassen sich keiner Facette zuordnen (Kontaminationen).

öffentlich. 27.5% der Gesamtstichprobe sind männlich, 8.1% weiblich, von den restlichen 64.4% ist das Geschlecht nicht bekannt. Das mittlere Alter der Studienteilnehmer ist in 72 Studien berichtet: es liegt zwischen 15.8 und 41.0 Jahren ($M = 21.1$, $SD = 6.0$). Das Bildungsniveau der Stichproben ist zu 69.4% unbekannt, 11.5% haben die (Fach-)Hochschulreife, 10.6% die mittlere Reife und 8.7% maximal den Hauptschulabschluss. Es überwiegen handwerkliche Berufe und GATB-basierte Testverfahren (vgl. Tabelle 1). 197 Studien haben prädiktiven Charakter (Δ_i bei 86 Studien angegeben: $M = 28.5$, $SD = 16.0$ Monate), eine Studie berichtet eine retrograde Validität ($\Delta_i = -121$ Monate), in den restlichen 12 Studien werden GMA und Leistung zeitnah erhoben. 14 Studien berichten insignifikante Korrelationen nicht, ihr Anteil an der bare-bone Analyse beträgt 4.4%.

Einkommensmaße werden in sieben Studien als Kriterium verwendet ($N = 1\,980$). Sie stammen aus den Jahren 1975 bis 2006, lediglich zwei davon sind veröffentlicht. 60.6% der Gesamtstichprobe sind männlich, 39.4% weiblich, von 14.0% ist das Geschlecht unbekannt. Das mittlere Alter in den fünf Studien, die es berichten, liegt zwischen 16.9 und 39.0 Jahren (Δ_i : $M = 30.6$, $SD = 8.4$). Das Bildungsniveau von 14.9% der Gesamtstichprobe ist unbekannt, 38.2% können dem (Fach-)Hochschulreife-

veau, 33.1% dem mittlere Reife-niveau und 13.9% maximal dem Hauptschulniveau zugeordnet werden. In vier Studien werden verschiedene ISCO-88 Berufsgruppen unterschiedlichen Komplexitätsniveaus untersucht, zwei Studien mit einem Gewicht von insgesamt 48.8% an der bare-bone Analyse betrachten Bürokräfte (ISCO-88 Hauptgruppe 2). Je zwei Studien nutzen den IST 2000 R oder das LPS zur Erfassung von GMA, die restlichen drei Studien andere Verfahren (vgl. Tabelle 1). Die Art der Validität ist in vier Studien prädiktiv, in je einer konkurrent, retrograd oder nicht zuzuordnen. Eine Studie berichtet insignifikante Korrelationen nicht (bare-bone Gewicht: 3.5%).

Die berufliche Entwicklung ist Gegenstand von neun Studien ($N = 1\,817$) aus den Jahren 1975 bis 2006. Von 27.6% der untersuchten Personen wird das Geschlecht nicht berichtet, 38.7% sind männlich, 33.7% weiblich. Die Altersverteilung entspricht derjenigen der Einkommensstudien. 42.4% der Stichprobe verfügen über die (Fach-) Hochschulreife, 31.8% über die mittlere Reife und 1.3% maximal über den Hauptschulabschluss. Das Bildungsniveau von 24.4% ist unbekannt. Ein Großteil der Stichprobe sind Bürokräfte (ISCO-88 Hauptgruppe 2) oder können weder eindeutig einer Berufsgruppe noch einem Berufskomplexitätsniveau zugeordnet werden. Bei den verwen-

Tabelle 3. Ergebnisse der bare-bone Analysen

Kriterium	<i>k</i>	<i>N</i>	\bar{r}	SD_{beo}	SD_{art}	$\widehat{SD}_{\bar{r}}$	VE (%)	CI	fs_k
Arbeitsleistung	18	2 739	.329	.239	.073	.227	9.4	.22–.44	42
Lernleistung	210	30 451	.371	.151	.072	.133	22.9	.35–.39	570
Einkommen	7	1 980	.247	.129	.056	.117	18.7	.15–.34	11
Entwicklung	9	1 817	.232	.143	.067	.127	21.7	.14–.33	12

Anmerkungen: *k* = Anzahl der Studien; \bar{r} = mittlere *n*-gewichtete Korrelation; SD_{beo} = beobachtete Standardabweichung; SD_{art} = durch den Stichprobenfehler erklärte Standardabweichung; $\widehat{SD}_{\bar{r}}$ = geschätzte Standardabweichung von \bar{r} ; VE = Anteil der durch den Stichprobenfehler erklärten Varianz; CI = 95% Konfidenzintervall; fs_k = faile-safe *k*.

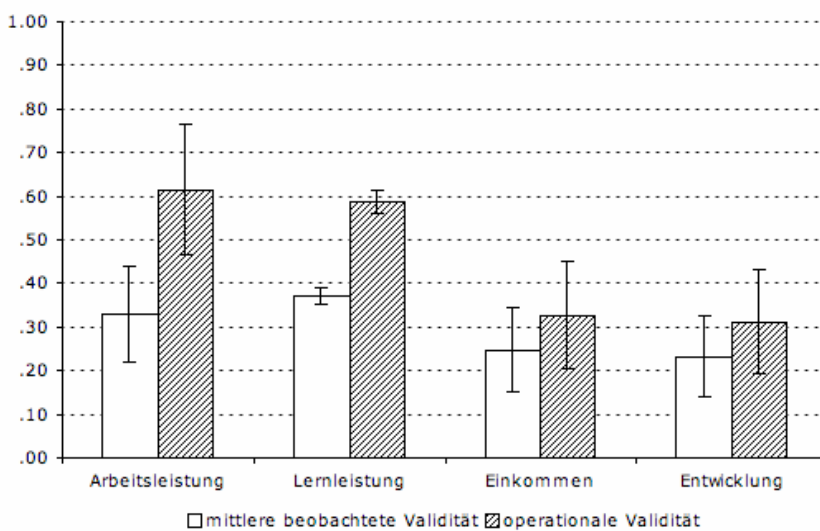
deten GMA-Maßen handelt es sich überwiegend um Eigenentwicklungen, in je zwei Studien wird der IST 2000 R und eine deutsche Adaptation des WPT (Wonderlic Inc, 1996) eingesetzt. Sechs Studien berichten prädiktive, zwei Studien konkurrenente und eine Studie eine nicht eindeutig zuordenbare Validität. Zwei Studien mit einem Gewicht von 20.2% an der bare-bone Analyse berichten insignifikante Korrelationen nicht.

Metaanalytische Ergebnisse

Die *n*-gewichtete mittlere Korrelation zwischen GMA und *Arbeitsleistung* beträgt $\bar{r} = .33$, mit *Lernleistung* korreliert GMA zu $\bar{r} = .37$, mit *Einkommen* zu $\bar{r} = .25$ und mit *beruflicher Entwicklung* zu $\bar{r} = .23$ (Tabelle 3). Das Konfidenzintervall der Lernleistungsstudien ist aufgrund der großen Stichprobe schmal. Es wären 570 Studien mit Nulleffekt erforderlich, um die mittlere Korrelation unter .10 zu drücken. Weniger akkurat sind die Schätzungen der mittleren Effekte der drei anderen Kriterien. Die Konfidenzintervalle der Lernleistungsanalyse einerseits und der Karriereerfolgsanalysen andererseits überlappen sich nicht: die Validität von GMA in Bezug auf Lernleistung ist daher deut-

lich höher als in Bezug auf das Einkommen und die berufliche Entwicklung. Die Validitäten sind unbeeinflusst von den Artefaktkorrekturen, die zur Berechnung der operationalen Validität vorgenommen werden. Durch die *n*-Gewichtung wurde ausschließlich der Stichprobenfehler korrigiert, der zwischen 9.4% und 22.9% der beobachteten Varianz aufklärt.

Die operationalen Validitäten nach Korrektur der Varianzeinschränkung im Prädiktor und der Unreliabilität im Kriterium (vgl. Tabelle 4) betragen im Falle von *Arbeitsleistungsstudien* $\hat{\rho} = .62$, *Lernleistungsstudien* $\hat{\rho} = .59$, *Einkommen* als Kriterium $\hat{\rho} = .33$ und *beruflicher Entwicklung* als Kriterium $\hat{\rho} = .31$. Die Unterschiede zwischen den $\hat{\rho}$ s der Arbeitsleistungs- und Lernleistungsstudien einerseits und der Karriereerfolgsstudien andererseits sind bedeutsam (vgl. Abbildung 1). Die unteren Grenzen der 90% Glaubwürdigkeitsintervalle sind größer als Null, die durch die Artefaktkorrekturen aufgeklärte Varianz liegt weit unter 75%. Damit lässt sich die Richtung aller Effekte verallgemeinern: GMA korreliert positiv mit den Kriterien. Die Höhe der Korrelation kann jedoch nicht verallgemeinert werden, sie hängt von weiteren Einflussfaktoren ab (*transportability*).



Anmerkung: Fehlerbalken = 95% Konfidenzintervalle.

Abbildung 1. Ergebnisse der Metaanalysen.

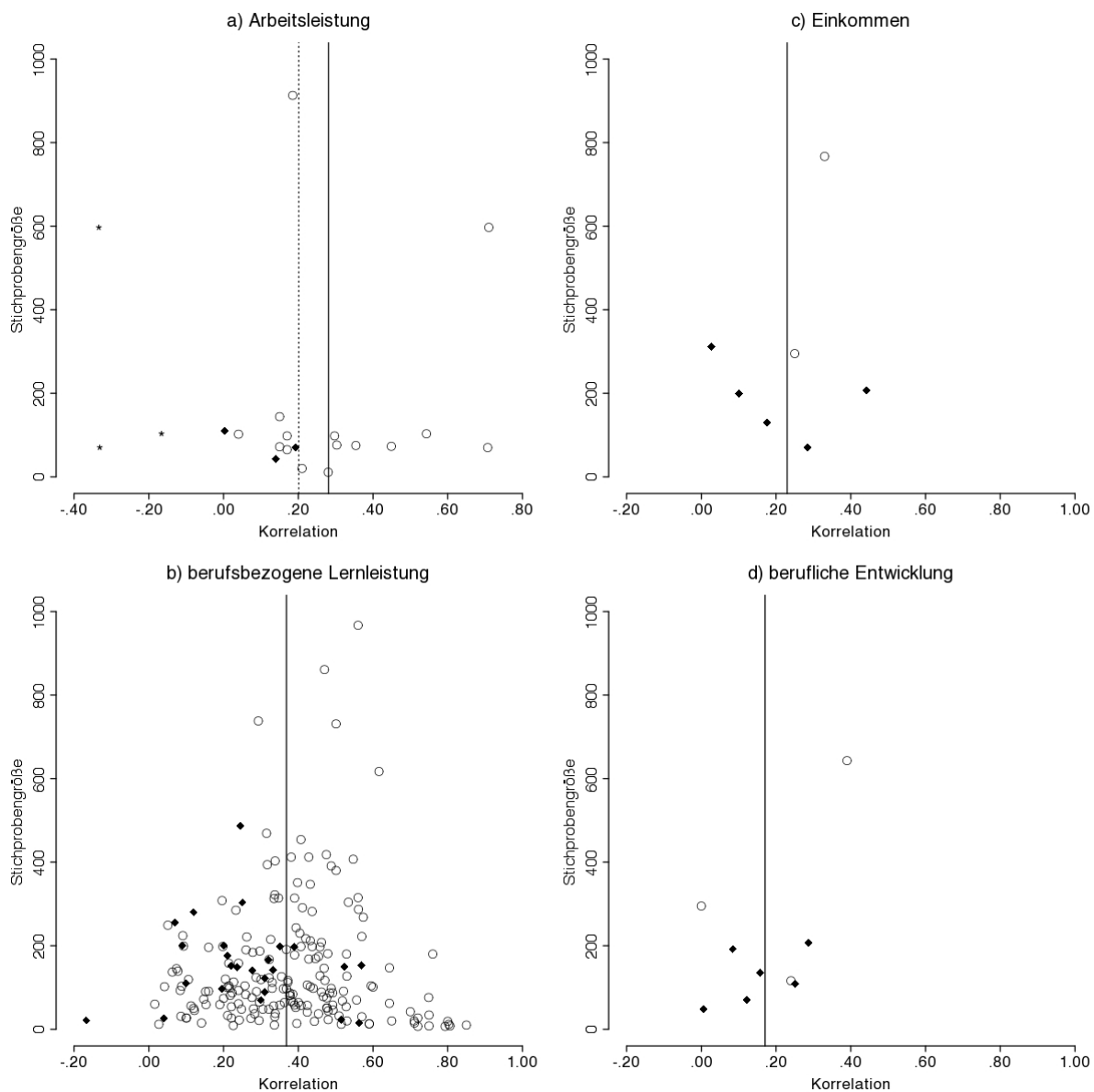
Veröffentlichungs- und Quellenverzerrungen

Zur Prüfung, ob Veröffentlichungsverzerrungen vorliegen, werden üblicherweise Funnel Plots erstellt (Sterne, Becker & Egger, 2005; vgl. Abbildung 2) und auf Symmetrie geprüft. Asymmetrische Funnel Plots können ein Hinweis darauf sein, dass eine Veröffentlichungsverzerrung vorliegt (Sutton & Pigott, 2005), wenn wenig Studien in den Extrembereichen der Verteilung vorhanden sind. Ob ein Funnel Plot asymmetrisch ist, kann mit zwei Kennwerten (L_0^+ und R_0^+) von Duval (2005) geprüft werden, die den Überhang von Studien angeben, die extrem hohe Werte berichten. Duval (2005) stellt auch ein *trim-and-fill* Verfahren vor, mit dem die Höhe der potenziellen Veröffentlichungsverzerrung eingeschätzt werden kann, wenn ein asymmetrischer Funnel Plot vorliegt.

Tabelle 4. Operationale Validitäten

Kriterium	r_{xx}	r_{yy}	u_x	$\hat{\rho}$	$\widehat{SD}_{\text{beo}}$	SD_{art}	$\widehat{SD}_{\hat{\rho}}$	VE(%)	CI	CV	Gen.
Arbeitsleistung	.880	.520	.743	.615	.323	.101	.307	9.7	.47–.76	.11	TRA
Lernleistung	.894	.639	.785	.587	.199	.101	.171	25.9	.56–.61	.31	TRA
Einkommen	.899	–	.780	.327	.165	.066	.151	16.1	.20–.45	.08	TRA
Entwicklung	.894	–	.779	.312	.182	.085	.161	21.7	.19–.43	.05	TRA

Anmerkungen: r_{xx} = mittlere Reliabilität des Prädiktors; r_{yy} = mittlere Reliabilität des Kriteriums; u_x = mittlere beobachtete Varianzeinschränkung; $\hat{\rho}$ = geschätzte operationale Validität; SD_{beo} = beobachtete Standardabweichung; SD_{art} = durch die Artefakte erklärte Standardabweichung; $\widehat{SD}_{\hat{\rho}}$ = geschätzte Standardabweichung von $\hat{\rho}$; VE = Anteil der durch die Artefakte erklärten Varianz; CI = 95% Konfidenzintervall; CV = untere Grenze des 90% Glaubwürdigkeitsintervalls, Gen. = Generalisierung: TRA = transportability.



Anmerkungen: o publiziert, \blacklozenge unpubliziert, * importiert (trim-and-fill Analyse), — \bar{r} , \bar{r} (trim-and-fill Analyse).

Abbildung 2. Funnel Plots.

Eine Asymmetrie liegt möglicherweise bei den *Arbeitsleistungsstudien* vor, da $R_0^+ > 0$. Mit der iterativen trim-and-fill Methode von Duval (2005) werden in drei Schritten drei Studien identifiziert, die zunächst exkludiert werden (*trim*-Schritte) und anschließend mit symmetrischen Pendanten wieder inkludiert werden (*fill*-Schritt). Die mittlere beobachtete Korrelation der Arbeitsleistungsstudien sinkt dabei von .33 auf .19 (vgl. Abbildung 2 a) bei steigender Varianz (von .06 auf .11). Für die Studien zum Karriereerfolg ist die Prüfung auf Veröffentlichungsverzerrung nicht erforderlich, da in diesem Erfolgsbereich überwiegend unveröffentlichte Studien vorliegen.

Der Funnel Plot der *Lernleistungsstudien* kann als symmetrisch angesehen werden ($L_0^+ = R_0^+ = 0$), eine Veröffentlichungsverzerrung scheint demnach nicht vorzuliegen. Vergleicht man die unterschiedlichen Veröffentlichungsarten der Lernleistungsstudien, fällt jedoch auf, dass die 25 nicht veröffentlichten Studien die geringste beobachtete Validität aufweisen ($\bar{r} = .25$, 95% CI: $.20 < \bar{r} < .30$), gefolgt von den 35 Monographien (ohne Testmanuale, $.27 < \bar{r} = .32 < .37$), 69 Artikeln in Zeitschriften oder Herausgeberwerken ($.29 < \bar{r} = .34 < .38$) und schließlich den 81 Studien, die in Testmanualen veröffentlicht wurden ($.41 < \bar{r} = .43 < .46$). Auch die operationalen Validitäten der unpublizierten Studien sind deutlich kleiner ($.33 < \hat{\rho} = .40 < .47$) und die der Testmanuale deutlich größer ($.64 < \hat{\rho} = .66 < .69$) als diejenigen der Monographien ($.46 < \hat{\rho} = .53 < .60$) und Artikel ($.50 < \hat{\rho} = .55 < .61$). Ob vergleichbare Unterschiede zwischen verschiedenen Veröffentlichungsarten auch bei Arbeitsleistungs- und Karriereerfolgsstudien vorliegen, kann wegen der vergleichsweise geringen Studienanzahl in diesen Analysen nicht untersucht werden. Bei der Interpretation dieser Ergebnisse ist zu berücksichtigen, dass sich die verschiedenen veröffentlichten Studien auch in anderer Hinsicht unterscheiden können. Die Unterschiede zwischen ihnen können deshalb nur als Hinweis auf eine möglicherweise vorliegende Verzerrung gewertet werden.

Diskussion

Die vorliegenden Metaanalysen haben gezeigt, dass GMA auch dann valide ist, wenn a) auch GMA-Operationalisierungen verwendet werden, die in den bisherigen Metaanalysen nicht berücksichtigt wurden, b) neben subjektiv bewerteter Arbeits- und berufsbezogener Lernleistung auch die Karriereerfolgsindikatoren *Einkommen* und *berufliche Entwicklung* betrachtet werden sowie c) Studien, die insignifikante Korrelationen nicht berichten, inkludiert werden. Durch die Inklusion zusätzlicher Studien und die Verwendung der indirekten statt direkten Korrektur der Varianzeinschränkung im Prädiktor, kann die Höhe der Validität von GMA akkurater als bisher eingeschätzt werden.

Gottfredson (2003) betrachtet berufliche und Lebensaufgaben im Hinblick darauf, wie stark sie mit GMA (genauer: dem *g*-Faktor) zusammenhängen. Sie stellt fest, dass die höchsten Zusammenhänge mit Bildungsaufga-

ben zu finden sind, die geringsten mit dem Einkommen und mittelhohe mit der beruflichen Entwicklung und der Arbeitsleistung. In anderen Worten: je distaler die Kriterien sind, desto geringer sind die Zusammenhänge mit GMA. Die Befunde aus Deutschland entsprechen dieser Feststellung in der Tendenz: GMA weist die höchste beobachtete Validität mit berufsbezogener Lernleistung auf, gefolgt von subjektiv bewerteter Arbeitsleistung und schließlich den Karriereerfolgskriterien *Einkommen* und *berufliche Entwicklung*, die sich nicht wesentlich voneinander unterscheiden (vgl. Abbildung 1). Von den Artefaktkorrekturen profitiert vor allem die Analyse der Arbeitsleistungsstudien: ihre operationale Validität ist geringfügig höher als die der Lernleistungsstudien.

Theoretische und praktische Implikationen

Mit den vorliegenden Metaanalysen werden die Beziehungen zwischen GMA und beruflichen Erfolgskriterien in Deutschland aufgezeigt. Diese zu kennen ist eine Voraussetzung für die Entwicklung von Theorien zum Einfluss mentaler Fähigkeiten auf beruflichen Erfolg: „Before theories can be developed, we need to be able to pin down the relations between variables“ (F. L. Schmidt & Hunter, 2001, S. 51).

Obwohl die Validität von GMA anhand amerikanischer Studien schon in den 1990er Jahren nachgewiesen und diese Validität von F. L. Schmidt und Hunter (1998) im Vergleich mit anderen Auswahlverfahren als sehr gut beschrieben wurde, werden GMA-Tests in Deutschland vergleichsweise selten eingesetzt (Ryan, MacFarland, Baron & Page, 1999). Schuler, Hell, Trapmann, Schaar und Boramir (2007) stellen sogar eine Abnahme des GMA-Einsatzes in der Personalauswahl zwischen 1993 und 2007 fest. Als Ursache geben die 125 von Schuler et al. (2007) befragten Unternehmen an, GMA nicht für valide zu halten. Dieser Einwand kann mit der vorliegenden Metaanalyse erneut entkräftet werden. Als weiteren Einwand gegen den Einsatz von GMA-Tests wird angeführt, dass sie auf wenig Akzeptanz bei den Bewerbern stoßen. Um die Akzeptanz zu verbessern, sind deshalb Tests zu entwickeln, die einen hohen Anforderungsbezug aufweisen, wie es auch in der DIN 33430 gefordert wird (Deutsches Institut für Normung e. V., 2002). Dieses Ziel wurde bei der Entwicklung einiger jüngerer GMA-Tests bereits berücksichtigt, z. B. den AZUBI-Tests (Görllich & Schuler, 2007; Schuler & Klingner, 2005), dem BIS-r-DGP (DGP, 1996) und dem WIT-2 (Kersting et al., 2008). Validierungsstudien zu den genannten Verfahren wurden in der vorliegenden Metaanalyse berücksichtigt.

Einschränkungen

Die Interpretation der korrigierten Korrelationen als Nachweise der *Validität* sehen Borsboom, Mellenbergh und van Herden (2004) als zu weitgehend an, da mit den Artefaktkorrekturen keine Validitätsmängel korrigiert werden. Sie schlagen deshalb vor, die korrigierten Korrelationen als Nachweis der *Qualität* von GMA anzusehen.

Die Varianzaufklärung in den durchgeführten Metaanalysen bleibt weit unter den von Hunter und Schmidt (2004) als Mindestmaß für Validitätsgeneralisierung angesehenen 75%. Dies spricht dafür, dass Moderatorvariablen die Höhe der GMA-Erfolgs-Korrelationen beeinflussen. Aufgrund der geringen Studienzahl – für eine aussagekräftige Metaanalyse werden mindestens sechs Studien benötigt (Levine, Spector, Menon, Narayanan & Cannon-Bowers, 1996; F. L. Schmidt, Hunter & Pearlman, 1981) – können Moderatoranalysen für Arbeitsleistungsstudien und Studien mit Einkommen bzw. beruflichen Entwicklungskriterien nicht durchgeführt werden. Die Studien zu berufsbezogener Lernleistung erlauben die getrennte Betrachtung unterschiedlicher GMA-Testgruppen, Lernleistungsindikatoren und Berufsgruppen. Auch bei getrennter Betrachtung dieser Moderatoren, verbleiben in den einzelnen Analysen deutlich mehr als 75% nicht aufgeklärte Varianz (Kramer, 2009).

Eine höhere Varianzaufklärung könnte vermutlich erreicht werden, wenn spezifischere Artefaktkorrekturen vorgenommen werden könnten. Hierzu wären mehr Informationen in den Primärstudien nötig, z. B. zur Varianzeinschränkung in Prädiktor und Kriterium. Die zur Messfehlerkorrektur verwendeten Maße stammen entweder nicht aus Deutschland (wie für die Kriterien Arbeitsleistung und betriebliche Lernleistungsbeurteilungen) oder sind älteren Datums (wie für die Ausbildungsleistung).

Mehr Informationen wären auch erforderlich, um bei allen Studien, die mehrere Effektstärken berichten, angemessene Aggregationen mithilfe linearer Verbünde berechnen zu können. Waren die Korrelationen zwischen den unterschiedlichen Prädiktoren bzw. Kriterien nicht erhältlich, wurden Mittelwerte berechnet, die die wahren Effekte unterschätzen. Alternativ dazu können die betroffenen Studien exkludiert werden, was im vorliegenden Datensatz zu nahezu denselben Ergebnissen führt (Kramer, 2009).

Ebenfalls nicht zufriedenstellend beantwortet werden kann die Frage danach, welche Rolle die Berufskomplexität für die Validität von GMA in Deutschland spielt. Die Befunde der Arbeitsgruppen von F. L. Schmidt und Hunter (Hunter, 1986; F. L. Schmidt & Hunter, 1998) sowie Salgado et al. (2003b) haben nachgewiesen, dass mit steigender Berufskomplexität, gemessen mit dem *Dictionary of Occupational Titles* (DOT; U. S. Department of Labor, 1991), die Validität von GMA zunimmt. Dieser Befund konnte von Hülshager et al. (2007) und Kramer (2009) nicht repliziert werden, vermutlich weil für die komplexesten (ISCO-88 Skill Level 4) und einfachsten Berufe (Skill Level 1) in Deutschland nur wenige Studien vorliegen und die ISCO-88-Operationalisierung von Berufskomplexität weniger messgenau sein dürfte als die DOT-Operationalisierung.

Da es sich bei den untersuchten Studien um Korrelationsstudien handelt, können die Unterschiede zwischen den verschiedenen Metaanalysen *nicht* kausal interpretiert werden. Obwohl sich die Validität von GMA zwischen den verschiedenen Kriterien unterscheidet und bei Lern-

leistung auch Unterschiede je nach Veröffentlichungsart festgestellt werden konnten, ist doch nicht nachgewiesen, dass diese Unterschiede ihre Ursache (allein) in den verschiedenen Kriterien bzw. Veröffentlichungsarten haben.

Schließlich sind insbesondere die verwendeten Indikatoren für die Karriereerfolgs-kriterien vergleichsweise heterogen (statische und dynamische Einkommensmaße, Beförderungs- und Positionsmaße). Es wäre wünschenswert, über ausreichend viele Studien zu verfügen, um die unterschiedlichen Indikatoren in Moderatoranalysen getrennt berücksichtigen zu können.

Weiterer Forschungsbedarf

Mit Metaanalysen kann auch gezeigt werden, zu welchen Bereichen Studien fehlen. Die Validität von GMA in Deutschland ist insbesondere im Hinblick auf berufsbezogene Lernleistung untersucht. Wenig Studien sind zu den Erfolgsbereichen *spezifische Arbeit* und *globale Laufbahn* verfügbar. Es fehlen (noch) Studien, die neuere GMA-Tests (z. B. den IST 2000 R, Amthauer et al., 2001; oder den WIT-2, Kersting et al., 2008) einsetzen. Hülshagers 2006 getroffene Aussage bleibt deshalb gültig: „eine Reihe von Intelligenztests (z. B. MIT, Büscher et al., 1986; BOMAT, Hossiep, Turck & Hassella, 2001; HAWIE, Tewes, 1991) ... [wurden] noch nicht hinsichtlich ihrer Kriteriumsvalidität überprüft“ (S. 31).

Die Berufe, zu denen Validierungsstudien vorliegen, sind i. d. R. dinge- bzw. datenorientiert (Bürokräfte, Handwerker; vgl. Prediger, 1982). Validierungsstudien bei personenorientierten Berufen fehlen (z. B. Erziehungs-, Heil- und Pflegeberufe). Weitere Metaanalysen sollten durchgeführt werden, die die Validität alternativer Personalauswahlverfahren und -instrumente in Deutschland untersuchen. Hülshager und Maier (2008) weisen darauf hin, dass die Bedeutung spezifischer kognitiver Fähigkeiten (z. B. sprachliches Denken, rechnerisches Denken), neuerer Intelligenzkonstrukte (z. B. praktische Intelligenz, emotionale Intelligenz) sowie typischerweise gezeigter kognitiver Fähigkeiten (z. B. „typical intellectual engagement“, „need for cognition“ (Ackerman, 1994; Cacioppo, Petty & Kao, 1984 zitiert nach Hülshager & Maier, 2008) untersucht werden sollten.

Schlussfolgerung

Die vorliegenden Metaanalysen haben gezeigt, dass GMA valide in Bezug auf die subjektiv bewertete Arbeitsleistung, die berufsbezogene Lernleistung, das Einkommen und Indikatoren der beruflichen Entwicklung ist, wenn GMA und die Kriterien mit den in Deutschland üblichen Instrumenten operationalisiert und bei Stichproben aus Deutschland erhoben werden.

Literatur

In die Metaanalysen inkludierte Quellen sind mit einem Sternchen (*) markiert.

- *Althoff, K. (1968). *Untersuchungen zum Einfluss von Berufsausbildungen und Berufstätigkeiten auf Testergebnisse und Eignungsdispositionen*. Unveröffentlichte Dissertation, Universität Gießen.
- *Althoff, K. (1974). Kurzbericht über eine Bewährungskontrolle der Eignungsprognosen für Bewerber der allgemeinen Verwaltung in Nordrhein-Westfalen. *DGP-Informationen*, 13, 12–21.
- *Althoff, K. (1975). Die Vorhersage des Ausbildungserfolgs von Aufstiegsbeamten der Kriminalpolizei. *DGP-Informationen*, 14, 13–23.
- *Althoff, K. (1977). Zusammenhänge zwischen den Ergebnissen von Eignungstests und beruflicher Bewährung – dargestellt am Beispiel der Auswahl von Bewerbern für den gehobenen und höheren Polizeivollzugsdienst. *Schriftenreihe der Polizeiführungsakademie*, 1, 6–27.
- *Althoff, K. (1978). [Unbekannt. Zitiert nach Jäger & Althoff, 1994, S. 37]
- *Amelang, M. & Steinmayr, R. (2006 a). „Is there a validity for tests of emotional intelligence in explaining the variance of performance criteria?“. Unveröffentlichte Daten, Universität Heidelberg.
- *Amelang, M. & Steinmayr, R. (2006 b). Is there a validity increment for tests of emotional intelligence in explaining the variance of performance criteria? *Intelligence*, 34, 459–468.
- *Amthauer, R. (1973). *I-S-T 70: Intelligenz-Struktur-Test* (3. erw. Aufl.). Göttingen: Hogrefe.
- Amthauer, R., Brocke, B., Liepmann, D. & Beauducel, A. (2001). *I-S-T 2000 R. Intelligenz-Struktur-Test 2000 R*. Göttingen: Hogrefe.
- *Awiszus, M. (1967). Über den Vorhersagewert des Intelligenzstrukturtests (I-S-T) für die Leistung in einer Berufsaufbauschule. *Der Schulpsychologe*, 14, 50–61.
- *Balck, F., Leins, B. & Schröder, U. (1979). Gruppenunterschiede, Reliabilitäten und Validitäten des Berufseignungstests (BET) von Schmale und Schmidtke. *Psychologie und Praxis*, 23, 153–164.
- Baron-Boldt, J., Funke, U. & Schuler, H. (1989). Prognostische Validität von Schulnoten. Eine Metaanalyse der Prognose des Studien- und Ausbildungserfolgs. In R. Jäger, R. Horn & K. Ingenkamp (Hrsg.), *Tests und Trends 7* (S. 11–39). Weinheim: Beltz.
- Baron-Boldt, J., Schuler, H. & Funke, U. (1988). Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 2, 79–90.
- *Barthel, E., Fünfgelt, V. & Schuler, H. (1982). *Modellversuch zur Erschließung gewerblich-technischer Ausbildungsberufe für Mädchen. 2. Zwischenbericht der wissenschaftlichen Begleitung in Augsburg*. Unveröffentlichtes Manuskript, Universität Erlangen-Nürnberg.
- Bertua, C., Anderson, N. & Salgado, J. F. (2005). The predictive validity of cognitive ability tests: A UK meta-analysis. *Journal of Occupational and Organizational Psychology*, 78, 387–409.
- *Biernat, J. & Scholz, B. (1995). Eine Bewährungskontrolle des DGP-Auswahlverfahrens für den mittleren feuerwehrtechnischen Dienst. *DGP-Informationen*, 54, 21–30.
- *Blickle, G., Kramer, J. & Mierke, J. (2005). *Politische Fertigkeiten, Intelligenz und Arbeitsleistung*. Unveröffentlichte Daten.
- *Blickle, G., Witzki, A. & Schneider, P. B. (2006). *Längsschnittstudie zum selbstinitiierten Mentoring in Organisationen*. Unveröffentlichte Daten aus dem DFG-Projekt AZBL62.
- Bobko, P., Roth, P. L. & Potosky, D. (1999). Derivation and implications of a meta-analytic matrix incorporating cognitive ability, alternative predictors, and job performance. *Personnel Psychology*, 52, 561–589.
- Borenstein, M. (2005). Software for publication bias. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Hrsg.), *Publication bias in meta analysis: prevention, assessment and adjustments* (S. 193–220). Chichester: Wiley.
- Borkenau, P., Egloff, B., Eid, M., Henning, J., Kersting, M., Neubauer, A. C. & Spinath, F. M. (2005). Persönlichkeitspsychologie: Stand und Perspektiven. *Psychologische Rundschau*, 56, 271–290.
- Borsboom, D., Mellenbergh, G. J. & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111, 1061–1071.
- Brambring, M. (1983). Spezielle Eignungsdiagnostik. In K. J. Groffmann & L. Michel (Hrsg.), *Grundlagen psychologische Diagnostik* (Bd. 2, S. 414–481). Göttingen: Hogrefe.
- *Brandstätter, H. (1970). *Leistungsprognose und Erfolgskontrolle*. Bern: Huber.
- *Bretz, E. & Oldendörp, H. (1992). Bewährungskontrolle: Vorhersage des Ausbildungserfolgs im Angestelltenlehrgang I. *DGP-Informationen*, 51, 75–84.
- Brickenkamp, R. (Hrsg.). (2002). *Handbuch psychologischer und pädagogischer Tests* (3. Aufl., Bd. 1). Göttingen: Hogrefe.
- Cronbach, L. J. (1947). Test „reliability“: Its meaning and determination. *Psychometrika*, 12, 1–16.
- *Dahle, K.-P. & Erdmann, K. (2001). Aus der Berliner CRIME-Studie. Unveröffentlichte Daten.
- *Damitz, M., Eißfeldt, H., Grasshoff, D., Lorenz, B., Pecena, Y. & Schwert, T. (2000). *Validierung des DLR-Auswahlverfahrens für Nachwuchsfluglotsen der DFS Deutsche Flugsicherung GmbH: Ergebnisse des Projektes Qualitätssicherung*. Hamburg: Deutsches Zentrum für Luft- und Raumfahrt e. V.
- Daniels, J. C. (1962). *FRT – Figure reasoning test*. London: Lockwood.
- *Derner, N. (1983). Der Vorhersagewert verschiedener Eingangsvariablen für den Ausbildungserfolg von Metallfacharbeitern. *Lernzielorientierter Unterricht*, 4, 16–26.
- *Deter, B. & Herbig, V. (1992). *Zusammenhänge zwischen standardisierten Ausbildungstests und den betrieblichen Beurteilungen angehender Chemikanten im zweiten Lehrjahr*. Unveröffentlichtes Manuskript.
- Detle, D. E., Abele, A. E. & Renner, O. (2004). Zur Definition und Messung von Berufserfolg. Theoretische Überlegungen und metaanalytische Befunde zum Zusammenhang von externen und internen Laufbahnerfolgsmaßen. *Zeitschrift für Personalpsychologie*, 3, 170–183.
- Deutsche Gesellschaft für Personalwesen e.V. (DGP). (1996). *Berliner Intelligenzstruktur-Test der Deutschen Gesellschaft für Personalwesen e. V. (BIS-r-DGP)*. Hannover: Deutsche Gesellschaft für Personalwesen e.V.
- Deutsches Institut für Normung e.V. (2002). *DIN 33430: Anforderungen an Verfahren und deren Einsatz bei berufsbezogenen Eignungsbeurteilungen*. Berlin: Beuth.
- *Diepenbrock, W. & Thielepape, M. (1973). Untersuchung über die Vorhersage des Ausbildungserfolgs von Meisteranwärtern in der Automobilproduktion. *DGP-Informationen*, 12, 1–14.
- *Dreyer, H. & Grabitz, H. (1971). Über den Wert psychologischer Eignungsuntersuchungen bei der Auslese von potentiellen Führungskräften. *Psychologische Rundschau*, 22, 187–196.
- Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: prevention, assessment and adjustments* (pp. 127–144). Chichester: Wiley.
- *Ehlers, T. (1961). Über den Vorhersagewert einiger psychologischer Untersuchungsverfahren. *Psychologie und Praxis*, 5, 157–163.

- *Fünfgelt, V., Barthel, E. & Schuler, H. (1984). *Modellversuch zur Erschließung gewerblich-technischer Ausbildungsberufe für Mädchen. Abschlußbericht der wissenschaftlichen Begleitung in Augsburg*. Unveröffentlichtes Manuskript, Universität Hohenheim.
- *Funke, U. (1986). Die Validität verschiedener eignungsdiagnostischer Verfahren bei Lehrstellenbewerbern. *Zeitschrift für Arbeits- und Organisationspsychologie*, 30, 92–97.
- Funke, U., Krauß, J., Schuler, H. & Stapf, K. H. (1987). Zur Prognostizierbarkeit wissenschaftlich-technischer Leistungen mittels Personvariablen: Eine Metaanalyse der Validität diagnostischer Verfahren im Bereich Forschung und Entwicklung. *Gruppendynamik*, 18, 407–428.
- Geis, A. (2006). *Handbuch für die Berufsvercodung*. Mannheim: ZUMA.
- Glass, G. V., McGaw, B. & Smith, M. L. (1984). *Meta-analysis in social research* (3. Aufl.). Beverly Hills, CA: Sage.
- *Görlich, Y. & Schuler, H. (2004). *Arbeitsprobe zur berufsbezogenen Intelligenz – Technische und handwerkliche Tätigkeiten (AZUBI-TH) und berufliche Ausbildungsleistung*. Unveröffentlichte Daten.
- Görlich, Y. & Schuler, H. (2007). *Arbeitsprobe zur berufsbezogenen Intelligenz – Technische und handwerkliche Tätigkeiten (AZUBI-TH)*. Göttingen: Hogrefe.
- Gottfredson, L. S. (2003). g, jobs and life. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 293–342). Oxford: Pergamon.
- *Graudenz, H. (1982). Vorhersage des Ausbildungserfolges von Beamtenanwärtern des gehobenen Dienstes beim RP Darmstadt. *DGP-Informationen*, 42, 32–50.
- *Graudenz, H. (1984). Personalauswahl für den mittleren Justizvollzugsdienst in Hessen. Eine Bewährungskontrolle. *DGP-Informationen*, 44, 1–17.
- *Greif, S. (1972). *Gruppenintelligenztests – Untersuchungen am WIT, IST, LPS und AIT*. Frankfurt: Lang.
- *Herzog, (2004). *Eignungsfeststellung beim Zentrum für Nachwuchsgewinning der Marine. Eine Prüfung der soldatischen Eignungsprognosen auf ihre Vorhersagevalidität*. Unveröffentlichte Diplomarbeit, [Ort und Institution unbekannt].
- Holling, H., Preckel, F. & Vock, M. (2004). *Intelligenzdiagnostik*. Göttingen: Hogrefe.
- Horn, W. (1962). *Leistungsprüfsystem LPS*. Göttingen: Verlag für Psychologie.
- Horn, W. (1983). *Leistungsprüfsystem LPS. Handanweisung* (2. erw. und verb. Aufl.). Göttingen: Hogrefe.
- Horn, W. (1969). *Prüfsystem für Schul- und Bildungsberatung PSB. Handanweisung*. Göttingen: Hogrefe.
- *Hossiep, R. (1995). *Berufseignungsdiagnostische Entscheidungen. Zur Bewährung eignungsdiagnostischer Ansätze*. Göttingen: Hogrefe.
- Hülshager, U. R. (2006). *Persönlichkeit im Arbeitsleben: Zur Bedeutung von Persönlichkeitseigenschaften und Intelligenz für berufliche Leistung und beruflichen Erfolg*. Unveröffentlichte Dissertation, Universität Bielefeld.
- Hülshager, U. R. & Maier, G. W. (2008). Persönlichkeitseigenschaften, Intelligenz und Erfolg im Beruf. Eine Bestandsaufnahme internationaler und nationaler Forschung. *Psychologische Rundschau*, 59, 108–122.
- Hülshager, U. R., Maier, G. W. & Stumpp, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment*, 15, 3–18.
- Hülshager, U. R., Maier, G. W., Stumpp, T. & Muck, P. M. (2006). Vergleich kriteriumsbezogener Validitäten verschiedener Intelligenztests zur Vorhersage von Ausbildungserfolg in Deutschland. *Zeitschrift für Personalpsychologie*, 5, 145–162.
- Hunter, J. E. (1986). Cognitive ability, cognitive aptitudes, job knowledge, and job performance. *Journal of Vocational Behavior*, 29, 340–362.
- Hunter, J. E. & Schmidt, F. L. (1996). Intelligence and job performance: Economic and social implications. *Psychology, Public Policy, and Law*, 2, 447–472.
- Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2. Aufl.). Beverly Hills, CA: Sage.
- 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, 594–612.
- International Labour Office (Hrsg.). (1990). *International Standard Classification of Occupations: ISCO-88*. Geneva: International Labour Office.
- *Jäger, A. O. (1960). Zum prognostischen Wert psychologischer Eignungsuntersuchungen, eine Bewährungsstudie. *Psychologische Rundschau*, 6, 160–178.
- Jäger, A. O. (1982). Mehrmodale Klassifikation von Intelligenzleistungen. Experimentell kontrollierte Weiterentwicklung eines deskriptiven Intelligenzstrukturmodells. *Diagnostica*, 28, 195–226.
- Jäger, A. O. (1984). Intelligenzstrukturforschung: Konkurrierende Modelle, neue Entwicklungen, Perspektiven. *Psychologische Rundschau*, 35, 21–35.
- Jäger, A. O. & Althoff, K. (1983). *Der Wilde-Intelligenz-Test (WIT) – Ein Strukturdiagnostikum*. Göttingen: Hogrefe.
- Jäger, A. O. & Althoff, K. (1994). *Der Wilde-Intelligenz-Test (WIT) – Ein Strukturdiagnostikum* (2. rev. Aufl.). Göttingen: Hogrefe.
- Jäger, A. O., Süß, H. M. & Beauducel, A. (1997). *BIS-TEST Berliner Intelligenzstruktur-Test – Form 4*. Göttingen: Hogrefe.
- *Jungkunz, D. (1995). *Berufsausbildungserfolg in ausgewählten Ausbildungsberufen des Handwerks. Theoretische Klärung und empirische Analyse*. Weinheim: Deutscher Studien Verlag.
- *Jungkunz, D. & Bodinet, K. (1989). Korrelative Bedeutung von Testergebnissen, schulischer Vorbildung, Berufsschulnoten und Fähigkeitseinschätzungen für Berufsabschlussnoten. *Zeitschrift für Pädagogik*, 35, 711–730.
- *Jungkunz, D. & John, W. (1991). Bedingungsfaktoren bei der Vergabe von Zeugnisnoten an Auszubildende handwerklicher Ausbildungsberufe. *Zeitschrift für Berufs- und Wirtschaftspädagogik*, 87, 481–494.
- Kanning, U. P. (Hrsg.). (2002). *Handbuch personaldiagnostischer Instrumente*. Göttingen: Hogrefe.
- Kemery, E. R., Mosholder, K. W. & Roth, L. (1987). The power of the Schmidt and Hunter additive model of validity generalization. *Journal of Applied Psychology*, 72, 30–37.
- *Kersting, M. (1999). *Diagnostik und Personalauswahl mit computergestützten Problemlösenszenarien? Zur Kriteriumsvalidität von Problemlösenszenarien und Intelligenztests*. Göttingen: Hogrefe.
- *Kersting, M. (2003). Assessment Center: Erfolgsmessung und Qualitätskontrolle. In S. Höft & B. Wolf (Hrsg.), *Qualitätsstandards für Personalentwicklung in Wirtschaft und Verwaltung* (S. 72–93). Hamburg: Windmühle.
- *Kersting, M., Althoff, K. & Jäger, A. O. (2005). *Kriteriumsvalidität neuer Wilde Test*. Unveröffentlichte Daten.
- Kersting, M., Althoff, K. & Jäger, A. O. (2008). *Der WILDE Intelligenztest 2 (WIT-2)*. Göttingen: Hogrefe.
- *Kettel, J. & Simmat, W. (1969). Beitrag zu einem psychologischen Berufsprofil für Fernmeldehandwerker. *Psychologie und Praxis*, 13, 169–174.
- *Kleinevoss, R. (1983). Untersuchung zur Vorhersage des Ausbildungserfolges von Anwärtern des gehobenen Dienstes einer Bundesbehörde. *DGP-Informationen*, 43, 41–72.
- *Knebelau, M. (2002). *Evaluation des Wilde-Intelligenz-Tests (WIT) und seiner Funktion als eignungsdiagnostisches Instrument*. Unveröffentlichte Diplomarbeit, RWTH Aachen.
- *Kober, U. (1978). Eignungsuntersuchungs-Ergebnisse bei mehreren Ausbildungsberufen in einer Bewährungskontrolle

- durch verschiedene Methoden. *Psychologie und Praxis*, 22 (4), 168–177.
- Kramer, J. (2009). *Metaanalytische Studien zu Intelligenz und Berufsleistung in Deutschland*. Unveröffentlichte Dissertation, Universität Bonn.
- *Krause, D. E., Kersting, M. & Heggestad, E. D. (2006). Incremental validity of assessment center ratings over cognitive ability tests: A study at the executive management level. *International Journal of Selection and Assessment*, 14, 360–371.
- *Krause, D. E., Kersting, M., Heggestad, E. D. & Thornton, G. C. (2005, April). *Criterion validity of cognitive ability tests and assessment centers*. Paper presented at the 20th Annual Society of Industrial and Organizational Psychology (SIOP) Conference, USA.
- *Kreuscher, R. (1987). *Der Zusammenhang von Berufserfolg mit Handlungsstil und fünf weiteren, psychologischen Variablen bei selbständig im Außendienst tätigen*. Unveröffentlichte Diplomarbeit, Maximilians-Universität München.
- *Kuhl, J. & Greve, H. G. (1971). Validierung zweier Testsysteme zur Selektion und Beurteilung von Auszubildenden in Metall- und Elektroberufen. *Psychologie und Praxis*, 15, 81–88.
- *Lang, K. (1990). *Entwicklung und längsschnittliche Validierung einer eignungsdiagnostischen Testbatterie für Auszubildende gewerblich-technischer Berufe*. Unveröffentlichte Diplomarbeit, Justus-Liebig-Universität Gießen.
- Levine, E. L., Spector, P. E., Menon, P. E., Narayanan, L. & Cannon-Bowers, J. (1996). Validity generalisation for cognitive, psychomotor, and perceptual tests for craft jobs in the utility industry. *Human Performance*, 9, 1–22.
- *Linz, P. (1985). Die kognitive Leistungsfähigkeit und die berufliche Integration hörgeschädigter Jugendlicher. In W. F. Kugemann, S. Preiser & K. A. Schneewind (Hrsg.), *Psychologie und Lebenswirklichkeit. Festschrift zum 65. Geburtstag von Walter Toman* (S. 309–320). Göttingen: Hogrefe.
- *Lometsch, D. (1975). *Untersuchungen zur Vorhersage des Berufserfolgs von Führungsnachwuchskräften*. Unveröffentlichtes Manuskript, Hamburg: Deutsche Unilever GmbH.
- *Marcus, B., Schuler, H., Quell, P. & Hümpfer, G. (2002). Measuring counter productivity: development and initial validation of a German self-report questionnaire. *International Journal of Selection and Assessment*, 10, 18–35.
- *Marschner, G. (1964). Auswahl des Steigernachwuchses. *Psychologie und Praxis*, 8, 49–65.
- *Marschner, G. (1966). Betriebspsychologische Erfahrungen mit dem I-S-T-Amthauer bei Auswahluntersuchungen. *Psychologie und Praxis*, 10, 145–153.
- *Marschner, G. (1972). Auswahluntersuchungen bei Führungsnachwuchskräften. *Psychologie und Praxis*, 16, 116–127.
- *Marschner, G. (1978). Differential-psychologische Untersuchungen an industriellen Facharbeitern. *Zeitschrift für Arbeitswissenschaft*, 32, 222–224.
- *Marschner, G. (1981). Untersuchungen mit dem Lern- und Gedächtnis-Test (LGT-3) bei Facharbeitern. *Diagnostica*, 27, 261–265.
- *Marschner, G. (1983). Die geistige Leistungsfähigkeit im Verlauf der beruflichen Ausbildung. *Zeitschrift für Arbeitswissenschaft*, 37, 50–53.
- 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.
- *Nettelstroth, W. (2004). Intelligenz und Berufserfolg. Eine Validierung der fluiden und kristallinen Intelligenz an subjektiven und objektiven Berufserfolgskriterien. In W. Bungard (Hrsg.), *Psychologie und Wirtschaft leben* (S. 415–420). München: Hampp.
- *Neumann, E. (1938). Psychotechnische Eignungsprüfung und Anlernung im Flugmotorenbau. *Industrielle Psychotechnik*, 15, 111–162.
- Ng, T. W. H., Eby, L. T., Sorensen, K. L. & Feldman, D. C. (2005). Predictors of objective and subjective career success: A meta-analysis. *Personnel Psychology*, 58, 367–408.
- *Paff, G. (1963). Ergebnisse einer Bewährungskontrolle bei Apparatewärmern und Maschinisten. *FORFA-Briefe*, 2, 89–92.
- *Paff, G. (1966). *Schulleistung, Berufseignung und Bewährung*. Bern: Hans Huber.
- Prediger, D. J. (1982). Dimensions underlying Holland's hexagon: Missing link between interests and occupations? *Journal of Vocational Behavior*, 21, 259–287.
- Raven, J., Raven, J. C. & Court, J. H. (2000). *Raven manual: Section 3. SPM manual (including the parallel and plus versions)*. 2000 Edition. Oxford: Oxford Psychologists Press.
- Raven, J., Raven, J. C. & Court, J. H. (2003). *Raven's standard progressive matrices plus – SPMPPLS*. Manual. Wien: Schuffried.
- *Rolloff, H. P. (1928). Über Eignung und Bewährung. Forschungen zur industriellen Psychotechnik. *Beihefte zur Zeitschrift für Angewandte Psychologie*, 148.
- Rothstein, H. R. (1990). Interrater reliability of job performance ratings: Growth to asymptote level with increasing opportunity to observe. *Journal of Applied Psychology*, 75, 322–327.
- Rustenbach, S. J. (2003). *Metaanalyse. Eine anwendungsorientierte Einführung*. Bern: Huber.
- Ryan, A. M., MacFarland, 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.
- Sackett, P. R. & Ostgaard, D. J. (1994). Job-specific applicant pools and national norms for cognitive ability tests: Implications for range restriction corrections in validation research. *Journal of Applied Psychology*, 79, 680–684.
- Salgado, J. F. & Anderson, N. (2003). Validity generalization of GMA tests across countries in the European Community. *European Journal of Work and Organizational Psychology*, 12, 1–17.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C. & de Fruyt, F. (2003a). International validity generalization of GMA and cognitive abilities: A European community meta-analysis. *Personnel Psychology*, 56, 573–605.
- Salgado, J. F., Anderson, N., Moscoso, S., Bertua, C., de Fruyt, F. & Rolland, J.-P. (2003b). A meta-analytic study of general mental ability validity for different occupations in the European community. *Journal of Applied Psychology*, 88, 1068–1081.
- *Salzgitter Service und Technik GmbH (SZST). (2004). *Zusammenhang Intelligenzmaße und Leistungsmaße bei Auszubildenden der SZST* (früher PPS), Salzgitter: Unveröffentlichte Daten.
- Sarges, W. & Wottawa, H. (Hrsg.). (2004). *Handbuch wirtschaftspsychologischer Testverfahren. Band I: Personalpsychologische Instrumente* (2. überarb. und erw. Aufl.). Lengerich: Pabst.
- *Schäfer, N. (1986). *Eignung und Arbeitszufriedenheit. Beiträge zur Eignungsdiagnostik*. Frankfurt: Lang.
- Schmale, H. & Schmidtke, H. (1995). *Handanweisung für den Berufseignungstest* (3. korr. Aufl.). Bern: Huber.
- *Schmale, H. & Schmidtke, H. (2001). *Berufseignungstest BET Tabellenband* (4. überarb. und erg. Aufl.). Bern: Huber.
- Schmidt, F. L. & Hunter, J. E. (1998). Meßbare Personenmerkmale: Stabilität, Variabilität und Validität zur Vorhersage zukünftiger Berufsleistung und berufsbezogenen Lernens. In M. Kleinmann & B. Strauss (Hrsg.), *Potentialfeststellung und Personalentwicklung* (S. 15–43). Göttingen: Verlag für angewandte Psychologie.
- Schmidt, F. L. & Hunter, J. E. (2001). Meta-analysis. In N. Anderson, D. S. Ones, H. K. Sinangil & C. Viswesvaran (Hrsg.), *Handbook of industrial, work and organizational psychology* (Vol. 1, S. 51–70). London: Sage.

- Schmidt, F. L. & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Applied Psychology*, 86, 162–173.
- Schmidt, F. L., Hunter, J. E. & Outerbridge, A. N. (1986). Impact of job experience and ability on job knowledge, work sample performance, and supervisory ratings of job performance. *Journal of Applied Psychology*, 71, 432–439.
- Schmidt, F. L., Hunter, J. E. & Pearlman, K. (1981). Task differences as moderators of aptitude test validity in selection: A red herring. *Journal of Applied Psychology*, 66, 166–185.
- Schmidt, F. L., Le, H. & Iles, R. (2003). Beyond alpha: an empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods*, 8, 206–224.
- Schmidt, F. L., Oh, I.-S. & Le, H. (2006). Increasing the accuracy of corrections for range restriction: Implications for selection procedure validities and other research results. *Personnel Psychology*, 59, 281–305.
- *Schmidt, J. U. (1987). Zur Vorhersagbarkeit des Ausbildungserfolgs im gehobenen Verwaltungsdienst durch einen auf der Grundlage des Berliner Intelligenzstrukturmodells entwickelten Test. *DGP-Informationen*, 38, 29–45.
- *Schmidt-Atzert, L. & Deter, B. (1993). Intelligenz und Ausbildungserfolg: Eine Untersuchung zur prognostischen Validität des I-S-T 70. *Zeitschrift für Arbeits- und Organisationspsychologie*, 37, 52–63.
- *Schmidt-Atzert, L., Deter, B. & Jaekel, S. (2004). Prädiktion von Ausbildungserfolg: Allgemeine Intelligenz (g) oder spezifische kognitive Fähigkeiten? *Zeitschrift für Personalpsychologie*, 3, 147–158.
- *Schmidt-Atzert, L. & Eser, K.-H. (2006). *Intelligenz, Leistungsmotivation und Prognose des Ausbildungserfolgs bei lern- und mehrfachbehinderten Teilnehmern in Berufsbildungswerken*. Vortrag präsentiert auf der Konferenz „teilhabe2006.de – Ein Beruf gehört zum Leben“, Berlin.
- *Schmitt, A. A. (2006). *Faktoren für Ausbildungserfolg in den neu geordneten technischen Ausbildungsberufen unter dem besonderen Fokus sozialer Kompetenzen*. Unveröffentlichte Daten.
- Schmitt, N., Cortina, J. M., Ingerick, M. J. & Wiechmann, D. (2003). Personnel selection and employee performance. In W. C. Borman, D. R. Ilgen & R. J. Klimoski (Hrsg.), *Handbook of Psychology: Industrial and organizational psychology* (Vol. 12, S. 77–105). Hoboken, NJ: Wiley.
- *Schmitz, I. (2006). *Prognostische Validität des Auswahlverfahrens für Auszubildende bei zwei Sparkassen in Nordrhein-Westfalen*. Unveröffentlichte Diplomarbeit, Universität Bonn.
- *Schuler, H., Barthel, E. & Fünfgelt, V. (1980). *Modellversuch zur Erschließung gewerblich-technischer Ausbildungsberufe für Mädchen. 1. Zwischenbericht der wissenschaftlichen Begleitung in Augsburg*. Unveröffentlichtes Manuskript, Universität Erlangen-Nürnberg.
- *Schuler, H., Barthel, E. & Fünfgelt, V. (1984). Erfolg von Mädchen in gewerblich-technischen Ausbildungsberufen: Ein Modellversuch. *Zeitschrift für Arbeits- und Organisationspsychologie*, 2, 67–78.
- *Schuler, H., Funke, U., Moser, K. & Donat, M. (1995). *Personalauswahl in Forschung und Entwicklung. Eignung und Leistung von Wissenschaftlern und Ingenieuren*. Göttingen: Hogrefe.
- *Schuler, H., Funke, U., Moser, K., Donat, M., Barthelme, D., Bretz, E. & Diemand, A. (1991). *Abschlussbericht zum Forschungsprojekt: „Entwicklung eines eignungsdiagnostischen Personalauswahlsystems für Wissenschaftler und Ingenieure im Bereich industrieller Forschung und Entwicklung“*. Berichtszeitraum 1.4.1986 – 30.9.1990. Unveröffentlichtes Manuskript, Universität Hohenheim.
- Schuler, H., Hell, B., Trapmann, S., Schaar, H. & Boramir, I. (2007). Die Nutzung psychologischer Verfahren der externen Personalauswahl in deutschen Unternehmen. Ein Vergleich über 20 Jahre. *Zeitschrift für Personalpsychologie*, 6, 60–70.
- Schuler, H. & Höft, S. (2006). Konstruktorientierte Verfahren der Personalauswahl. In H. Schuler (Hrsg.), *Lehrbuch der Personalpsychologie* (2., überarb. und erw. Aufl., S. 101–144). Göttingen: Hogrefe.
- *Schuler, H. & Klingner, Y. (2002). *Arbeitsprobe zur berufsbezogenen Intelligenz – Büro- und kaufmännische Tätigkeiten (AZUBI-BK) und Intelligenz-Struktur-Test 70 (IST-70) und berufliche Ausbildungsleistung*. Unveröffentlichte Daten.
- *Schuler, H. & Klingner, Y. (2005). *Arbeitsprobe zur berufsbezogenen Intelligenz – Büro- und kaufmännische Tätigkeiten (AZUBI-BK)*. Göttingen: Hogrefe.
- *Schuler, H., Moser, K., Diemand, A. & Funke, U. (1995). Validität eines Einstellungsinterviews zur Prognose des Ausbildungserfolgs. *Zeitschrift für Pädagogische Psychologie*, 9, 45–54.
- Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge: Hogrefe & Huber.
- *Schwadorf, H. (2003). *Berufliche Handlungskompetenz*. Hohenheim: ibw.
- *Seggebruch, G. (1982). Zur Vorhersage des Berufserfolgs von Bewerbern für den mittleren Verwaltungsdienst in einer Kommunalverwaltung. *DGP-Informationen*, 42, 6–31.
- *Seggebruch, G. (1984). Vergleich der Ergebnisse aus Eignungsuntersuchungen und Zwischenprüfungen im gehobenen allgemeinen Verwaltungsdienst – eine Bewährungskontrolle. *DGP-Informationen*, 44, 56–74.
- *Steinmann, V. (1997). *Die Prognosekraft des Eignungstests Bankkaufmann für den späteren Berufserfolg: eine Längsschnittstudie*. Unveröffentlichte Diplomarbeit, Universität Köln.
- Sterne, J. A. C., Becker, B. J. & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: prevention, assessment and adjustments* (pp. 75–98). Chichester: Wiley.
- *Streitkräfteamt. (1996 a). *Arbeitsbericht*. Unveröffentlichte Daten.
- *Streitkräfteamt. (1996 b). *Arbeitsbericht*. Unveröffentlichte Daten.
- *Streitkräfteamt. (2001). *Arbeitsbericht*. Unveröffentlichte Daten.
- *Streitkräfteamt. (2004). *Arbeitsbericht des Psychologischen Dienstes*. Unveröffentlichte Daten.
- Sutton, A. J. & Pigott, T. D. (2005). Bias in meta-analysis induced by incompletely reported studies. In H. R. Rothstein, A. J. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: prevention, assessment and adjustments* (pp. 223–239). Chichester: Wiley.
- *Thielepape, M. (1980). Untersuchung über die Vorhersage des Ausbildungserfolgs von Krankenschwestern und -pflegern – eine Bewährungskontrolle. *DGP-Informationen*, 17, 21–43.
- *Thielepape, M. & Kersting, M. (2005 a). Erfolgskontrolle und Qualitätsoptimierung in der Personalauswahl: 2 Bewährungskontrollen. *DGP-Informationen*, 49, 2–20.
- *Thielepape, M. & Kersting, M. (2005 b). Evaluation und Qualitätsoptimierung in der Personalauswahl: Zwei Bewährungskontrollen. Unveröffentlichte Daten.
- U. S. Department of Labor. (1991). *Dictionary of occupational titles* (4. Aufl.). Washington, DC: U. S. Government Printing Office.
- U. S. Employment Service. (1970). *General aptitude test battery (GATB)*. Washington, DC: U. S. Employment Service.
- Viswesvaran, C., Schmidt, F. L. & Ones, D. S. (2002). The moderating influence of job performance dimensions on convergence of supervisory and peer ratings of job performance: Unconfounding construct-level convergence and rating difficulty. *Journal of Applied Psychology*, 87, 345–354.

- *Weiß, R. (1969). *Die Brauchbarkeit des Culture Free Intelligence Tests Skala 3 (CFT 3) bei begabungspsychologischen Untersuchungen*. Unveröffentlichte Dissertation, Universität Würzburg.
- *Weiß, R. H. (1980). *Grundintelligenztest Skala 3 – CFT 3*. Braunschweig: Georg Westermann.
- Whitener, E. M. (1990). Confusion of confidence intervals and credibility intervals in meta-analysis. *Journal of Applied Psychology*, 75, 315–321.
- *Wolf, B. (1990). Untersuchung zur Vorhersage des Ausbildungserfolges von Aufstiegsbewerbern für den gehobenen Polizeivollzugsdienst – Kurzbericht über eine Bewährungskontrolle. *DGP-Informationen*, 50, 51–66.
- *Wolff, P. & Voullaire, C. (1968). Eignungsbegutachtung von Körperbehinderten für einen Verwaltungsberuf – eine Bewährungskontrolle. *Diagnostica*, 14, 70–87.
- Wonderlic Inc. (Ed.). (1996). *Wonderlic personnel test (WPT-German version, forms A and B)*. Libertyville, IL: Wonderlic.
- *Zeißig, H. J. (1989). *Berufsfindung und Arbeitserprobung der Berufsförderungswerke, Möglichkeiten ihrer Ergänzung hin zum Assessment-Center*. Unveröffentlichte Dissertation, Universität Bielefeld.

Dipl.-Psych. Jochen Kramer

Institut für Psychologie
Rheinische Friedrich-Wilhelms Universität Bonn
Kaiser-Karl-Ring 9
53111 Bonn
E-Mail: jochen.kramer@uni-bonn.de