

Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung – eine Metaanalyse

Benedikt Hell, Sabrina Trapmann, Sonja Weigand und Heinz Schuler

Zusammenfassung. Das Interesse an Kriterien und eignungsdiagnostischen Verfahren zur Auswahl von Studierenden im Rahmen der Hochschulzulassung hat in den vergangenen Jahren im deutschsprachigen Raum stark zugenommen. Während die Deutsche Gesellschaft für Psychologie (1997) Auswahlgesprächen kritisch gegenübersteht, wird der Einsatz dieser Verfahrensklasse von verschiedenen Hochschulvertretern und Bildungspolitikern befürwortet. Zur Bestimmung der prognostischen Validität von Interviews im Rahmen der Hochschulzulassung wird eine Metaanalyse von $k = 44$ Einzelstudien durchgeführt, die Auswahlgespräche als Prädiktor und Studiennoten als Erfolgskriterium berücksichtigen. Die mittlere korrigierte Validität von Interviews beträgt $\rho = .16$. Von insgesamt 11 geprüften Moderatoren der Validität erweist sich der Strukturierungsgrad der Interviews als bedeutsam. Unstrukturierte Interviews erreichen eine korrigierte Validität von $\rho = .11$ gegenüber einer korrigierten Validität von $\rho = .21$ für strukturierte Varianten dieser Verfahrensklasse. Die für strukturierte und unstrukturierte Auswahlgespräche berechneten Validitätskoeffizienten sind generalisierbar. Die inkrementelle Validität von Auswahlgesprächen gegenüber Schulnoten und Studierfähigkeitstests ist gering.

Schlüsselwörter: Metaanalyse, Interview, Auswahlgespräch, Studienerfolgsprognose, Auswahl von Studierenden

The validity of admission interviews – a meta-analysis

Abstract. Interest in the criteria and psychological assessment procedures that should be used to determine university admission has grown considerably in recent years in German-speaking countries. While the German Psychological Society (Deutsche Gesellschaft für Psychologie, 1997) is skeptical about admission interviews, several university representatives and education policy-makers favor this method. In order to assess the predictive power of interviews in the context of university admission, we have conducted a meta-analysis ($k = 44$ studies), in which the admission interview is the predictor and exam results are the criterion for academic performance. The mean corrected validity for interviews is $\rho = .16$. Among 11 validity moderators examined, the degree of structure in the interview proved significant. Unstructured interviews reach a corrected validity of $\rho = .11$ compared to a corrected validity of $\rho = .21$ for structured interviews. The computed validity coefficients for admission interviews can be generalized. The incremental validity of interviews to school grades and admission tests is low. Key words: meta-analysis, admission interview, academic performance prediction, admission criteria

Auswahlgespräche werden international in vielen Industriestaaten – z. B. in Australien, Großbritannien, Irland, Japan, Korea und den USA – zur Auswahl von Studierenden herangezogen (Deidesheimer Kreis, 1997). Insbesondere im anglo-amerikanischen Kulturraum haben Interviews als Bestandteil von Auswahlverfahren eine längere Tradition (Drolshagen, 1996; Heyneman & Fägerlind, 1988; Mitchell, 1990), im deutschen Sprachraum hingegen wurden Interviews bislang nur zurückhaltend eingesetzt (Heine, Briedis, Didi, Haase & Trost, 2005). In der Zeit von 1973 bis 1985 fanden an staatlichen Universitäten der Bundesrepublik Deutschland überhaupt keine Auswahlgespräche statt, in der Zeit davor und danach allenfalls in den medizinischen Disziplinen oder an privaten Hoch-

schulen. Erst in den letzten Jahren werden Interviews auch in anderen Disziplinen an öffentlichen Hochschulen durchgeführt.

Die psychometrische Qualität und insbesondere die Validität von Auswahlgesprächen im Kontext der Hochschulzulassung wurden bisher lediglich narrativ zusammengefasst. Trost (1996) berichtet, dass die Objektivität und Reliabilität von unstrukturierten Interviews gering ist und dass diejenigen Merkmale am zuverlässigsten erfasst werden, die in der Gesprächssituation unmittelbar beobachtet werden können. Die Prognosekraft der Interviewergebnisse für das Kriterium Studienerfolg sind laut Trost (1996, S. 63) „sehr gering“. Rindermann und Oubaid (1999) bewerten die Aussagekraft von Auswahlgesprächen im Rahmen der Hochschulzulassung anhand von zwei empirischen Studien und konstatieren, dass sich nach dem Kriterium Auswahlgespräch zugelassene Studierende *selbst* als „kooperativer, am Studienfach interessierter, kontaktfähiger und kontaktorientierter sowie

Dieser Beitrag entstand im Kontext des Projekts „Eignungsdiagnostische Auswahl von Studierenden“, das im Rahmen des Aktionsprogramms „StudierendenAuswahl“ des Stifterverbands für die Deutsche Wissenschaft und der Landessstiftung Baden-Württemberg durchgeführt wird.

sozial kompetenter“ (S. 182) einstufen. Morris (1999, p. 480) schließt seinen Review internationaler Studien mit der wenig hilfreichen Erkenntnis ab, dass die empirische Befundlage mehrdeutig ist: „Overall, results of studies into the fundamental issue of whether the admission interview is a reliable, valid and useful tool are equivocal.“ Dies führt zu der komfortablen Situation, dass diejenigen Hochschulen, die Interviews einsetzen, und auch diejenigen Hochschulen, die keine Interviews einsetzen, ihre Auswahlpraxis rechtfertigen können.

Während die Validität für Schulnoten (D: Schuler, Funke & Baron-Boldt, 1990; Trapmann, Hell, Weigand & Schuler, in Druck; GB: Peers & Johnston, 1994; USA: Burton & Ramist, 2001) und für Studierfähigkeitstests (D: Hell, Trapmann & Schuler, 2007; USA: Scholastic Assessment Test, ehemals Scholastic Aptitude Test, davor Scholastic Achievement Test, SAT: Burton & Ramist, 2001; Graduate Record Examination, GRE: Goldberg & Alliger, 1992; Kuncel, Hezlett & Ones, 2001; Morrison & Morrison, 1995; Law School Admission Test, LSAT: Linn & Hastings, 1984) mit metaanalytischen Techniken untersucht worden ist, fehlt eine quantitative Integration der Validitätsbefunde für Auswahlgespräche im Rahmen der Hochschulzulassung bislang. Die Prognosekraft von Auswahlgesprächen für den Berufserfolg und den Erfolg in beruflichen Trainings hingegen wurde in einigen Metaanalysen thematisiert. Nach der bislang umfassendsten Analyse von McDaniel, Whetzel, Schmidt und Maurer (1994) hängt die Validität der analysierten Interviews vom *Inhalt* (situationsbezogen, berufsbezogen oder psychologisch), der *Form* (strukturiert vs. unstrukturiert) und von der Art des *Kriteriums* (Berufserfolg, Trainingserfolg oder Beschäftigungsdauer) ab. Nach Anwendung von Prädiktor-Reliabilitäts- und Streuungskorrekturen berechnen sie für unstrukturiert geführte Interviews eine durchschnittliche Validität für Berufserfolg von .33 (unkorrigierter Mittelwert der Korrelationen: .18) und für strukturierte Varianten eine Validität von .44 (unkorrigierter Mittelwert der Korrelationen: .24). Die Prognosekraft hinsichtlich des Erfolgs in beruflichen Trainingmaßnahmen wird nach McDaniel et al. nicht durch den Strukturierungsgrad der Interviews moderiert. Für dieses Kriterium errechnen Sie eine Validität von .36 (unkorrigierter Mittelwert der Korrelationen: .23).

Diese Erkenntnisse bieten zwar Anhaltspunkte für Fragen der Hochschulzulassung – zumal Berufserfolg als erweitertes distales Studienerfolgskriterium insbesondere bei stark praxis- und anwendungsorientierten Studiengängen angesehen werden kann – sind jedoch nicht direkt auf proximale Studienerfolgskriterien übertragbar. Mit der vorliegenden Metaanalyse wird diese Lücke geschlossen, indem die Aussagekraft von Auswahlgesprächen für Studienerfolg untersucht wird. Der Studienerfolg kann grundsätzlich sehr unterschiedlich operationalisiert werden. Im allgemeinen Sprachgebrauch versteht man unter „Erreichen des Studienziels“, dass ein einmal angefangenes Studium mit dem Hauptexamen abgeschlossen wird. Alternativ oder ergänzend bieten sich andere Indikatoren des Studienerfolgs an: Erreichte Noten in Zwischen- oder

Abschlussprüfungen, erreichte Noten in einzelnen Lehrveranstaltungen, Studiendauer, Prüfungswiederholungen, Beurteilung des Studienerfolgs – beispielsweise durch Dozenten oder Kommilitonen – oder Angaben über die Zufriedenheit mit dem Studium via Selbsteinschätzung. Mit dem Ziel einer möglichst umfassenden Betrachtungsweise wurde für die vorliegende Metaanalyse nach Studien zu allen aufgeführten Studienerfolgskriterien gesucht, aber lediglich zum Erfolgskriterium Studiennoten konnten wir eine für eine Metaanalyse ausreichende Anzahl von Einzelstudien ausmachen – zu den Kriterien Studienabbruch, Studiendauer und Zufriedenheit wurden jeweils nicht mehr als zwei Studien ermittelt. Die Tatsache, dass die meisten Einzelstudien auf Studiennoten als Validierungskriterium zurückgreifen, beruht wohl zum einen auf der Auffassung, dass Studiennoten eine angemessene Quantifizierung der Studienleistung auf einer definierten Skala darstellen, aber vermutlich auch auf dem Umstand, dass die für eine Validierung notwendigen Daten im Falle von Studiennoten in der Regel zentral verwaltet werden und somit vergleichsweise ökonomisch zu erheben sind.

Methode

Metaanalytischer Ansatz

Es wurde die auf Hunter, Schmidt und Jackson (1982) zurückgehende Methode der Validitätsgeneralisierung nach Raju, Burke, Normand und Langlois (1991) angewendet, die eine im Vergleich zu anderen Methoden genauere Schätzung der wahren Validität und ihrer Varianz liefert (Mendoza & Reinhardt, 1991). Dieses Verfahren bietet die Möglichkeit, bei vollständig oder partiell fehlenden studienspezifischen Artefaktinformationen (z. B. Reliabilitätsangaben oder Angaben zur Variabilitätseinschränkung) diese zu schätzen, ohne auf hypothetische Artefaktverteilungen zurückzugreifen. Weiterhin spricht für den Ansatz von Raju et al. (1991), dass er bei der Korrektur der Validitäten den Stichprobenfehler der Artefakte berücksichtigt und nicht auf der Vorannahme beruht, dass Validität, Varianzinformation, Prädiktor- und Kriteriumsreliabilität unkorreliert sind. Eine handhabbare technische Umsetzung dieses Ansatzes bietet das Computerprogramm von Raju und Fleer (2003).

Ein sehr wichtiger Gesichtspunkt bei der Problemformulierung ist die Beantwortung der Frage, wie die möglichen Unterschiede in den berichteten Ergebnissen der Studien erklärt werden können (Hedges & Vevea, 1998). Im Rahmen des so genannten Fixed-Effects-Modells (FE-Modell) werden die zu schätzenden Parameter als in der Grundgesamtheit von Studien invariant angenommen. Unterschiede in den berichteten Effekten werden nur auf unsystematische Unterschiede zurückgeführt. In einer Situation, in der von der Heterogenität von Effekten a priori ausgegangen wird, sollte stattdessen das so genannte Random-Effects-Modell (RE-Modell) verwendet werden. Hier denkt man sich die Grundgesamtheit aus Stu-

dien zusammengesetzt, die unterschiedliche Merkmale aufweisen und denen verschiedene wahre Effektstärken zu Grunde liegen können. Da wir davon ausgehen, dass den Einzelstudien nicht *eine* Effektstärke zu Grunde liegt, sondern der Populationsparameter von Studie zu Studie variiert, ist bei der vorliegenden Fragestellung grundsätzlich die Anwendung des RE-Modells angezeigt (vgl. Erez, Bloom & Wells, 1996; Hunter & Schmidt, 2000); neueste Simulationsstudien kommen jedoch zu dem Ergebnis, dass RE-Modelle erst bei einer Mindestanzahl von 32 Primärkoeffizienten zuverlässige Schätzungen liefern (Schulze, 2004), daher wird in Fällen, in denen die Anzahl der Primärkoeffizienten über 32 liegt, das RE-Modell herangezogen, und in Fällen, in denen die Anzahl der Primärkoeffizienten unter 32 liegt, das FE-Modell angewendet. Der Vollständigkeit halber werden aber immer die Ergebnisse beider Modelle berechnet und dokumentiert.

Als Effektmaß zur Quantifizierung der Primärstudien-ergebnisse wurden alle statistischen Maße in bivariate Produkt-Moment-Korrelationen überführt. Bei der Transformation der Effektstärken kamen spezifische statistische Rekonstruktionsmethoden zur Anwendung (Hunter & Schmidt, 2004; Rustenbach, 2003). Fisher-Z-Transformationen wurden nicht verwendet, da weder durch Monte-Carlo-Untersuchungen (Schulze, 2004) noch in realen Metaanalysen (Hunter & Schmidt, 2004) eine gesteigerte Genauigkeit der Schätzungen von ρ durch Fisher-Z-Transformationen erreicht werden konnte.

Literaturrecherche und Kodierung

Bei der Literaturrecherche kamen multiple Suchstrategien zum Einsatz. Ziel war es, alle seit 1980 veröffentlichten Studien zur Validität von Auswahlgesprächen im Hochschulkontext zu entdecken. Zunächst wurden die Datenbanken PsycINFO und PsycDEX zur Recherche herangezogen. In einem zweiten Schritt wurden einschlägig arbeitende Kollegen konsultiert. Darüber hinaus wurden alle Studien daraufhin geprüft, ob sie Verweise auf andere Studien enthielten, die bislang unentdeckt geblieben waren. Es wurden sämtliche Publikationsformen akzeptiert.

Tabelle 1. Geographische Herkunft der Stichproben: Land

	<i>k</i>	<i>N</i>
Australien	1	39
Deutschland	6	262
Großbritannien	1	123
Israel	5	367
Kanada	5	332
Mexiko	1	292
Österreich	4	205
Pakistan	1	374
Tschechien	1	92
USA	19	2219

Anmerkungen: *k* = Anzahl der unabhängigen Stichproben; *N* = Stichprobenumfang.

Insgesamt konnten so 28 Publikationen mit $k = 44$ unabhängigen Stichproben zusammengetragen werden. Der Median der Stichprobengröße liegt bei $n = 53$ mit einer Streubreite von $n = 23$ bis $n = 596$ Interviewten ($SD = 118.6$). Insgesamt beläuft sich die Stichprobe auf 4305 Probanden. Die größte Stichprobenanzahl für eine Publikation liegt bei $k = 5$.

Ein größerer Teil der Stichproben entstammt dem nord-amerikanischen Kulturraum, gefolgt von Deutschland, Israel und Österreich (Tabelle 1). Angesichts der Verbreitung von Interviews in den USA wundert es nicht, dass hier auch die meisten Evaluationsstudien durchgeführt wurden. Desgleichen war auch zu erwarten, dass das Studienfach Medizin überrepräsentiert ist, da Auswahlgespräche im Rahmen medizinischer Studiengänge besonders häufig eingesetzt werden (Tabelle 2).

Regelrechte Interviewsysteme, wie das Behavior Description Interview (Janz, 1982), das Situational Interview (Latham, Saari, Pursell & Campion, 1980) oder das Multimodale Interview (Schuler, 1992) werden im Anwendungsfeld der Studienzulassung bislang offenbar kaum eingesetzt. Lediglich das in der Studie von Sue-Chan und Latham (2004) untersuchte Verfahren versteht sich als Situational Interview, das von Meister (2004) analysierte Instrument als Multimodales Interview.

Für den Beurteilungsprozess wurde ein Kodierleitfaden¹ entwickelt, der sämtliche zu kodierende Studienmerkmale ausführlich dokumentiert. Je Validitätskoeffizient wurden 79 Einzelinformationen zu den folgenden Aspekten kodiert: *Generelle Angaben zur Studie* (Herkunftsland der Studie, Jahr der Publikation), *Charakterisierung der Stichprobe* (Anzahl der in der Studie enthaltenen Stichproben, Größe der Stichproben, Geschlechterverteilung, Verteilung der Ethnien, Durchschnittsalter, Heterogenität der Stichprobe hinsichtlich verschiedener Kurse, Jahrgänge und Hochschulen, Selektionsquote und Selektionsmethode), *Charakterisierung des Prädiktors* (Inhalt, Bezeichnung des Instruments, deskriptive Statistiken, An-

Tabelle 2. Stichproben nach Studienfächern

	<i>k</i>	<i>N</i>
Sprach- und Kulturwissenschaften	1	75
Rechts-, Wirtschafts- und Sozialwissenschaften	13	908
Medizinische Disziplinen	22	1994
Kunst, Kunstwissenschaft, gestalterische Studiengänge	2	100
Lehramt	4	265
Gemeinsame Betrachtung mehrerer Studiengänge	2	963

Anmerkungen gem. Tabelle 1.

¹ Der Kodierleitfaden kann beim Erstautor angefordert werden.

gaben zur Reliabilität differenziert nach den verschiedenen Reliabilitätsmaßen), *Charakterisierung des Kriteriums* (Art des Kriteriums, Anzahl der Abstufungen, deskriptive Statistiken, Angaben zur Reliabilität differenziert nach den verschiedenen Reliabilitätsmaßen), *Informationen zur Validität* (verwendete Statistik, verwendete Korrekturen, Koeffizienten, methodische Besonderheiten) und *weitere Moderatorvariablen* (Studienfach, Studienabschnitt, Hochschulniveau, zeitlicher Abstand zwischen Prädiktor- und Kriteriumsbestimmung, Selektionskontext vs. freiwillige Teilnahme, Art der Publikation).

Vorgehensweise bei der Berechnung der Metaanalyse

Es wurden alle Primärstudien in die Analyse integriert, die Korrelationen zwischen Interviewergebnissen und Studiennoten oder andere in Korrelationen umrechenbare Effektmaße berichten (z. B. t-Statistiken, Ergebnisse von Varianzanalysen). Lagen mehrere Korrelationen vor (mehrere Interviewskalen, ohne dass ein Gesamtwert berichtet wird oder mehrere Einzelnoten als Erfolgskriterium, ohne dass eine Durchschnittsnote berichtet wird), so wurden diese unter Berücksichtigung der Interkorrelationen gepoolt oder – wo dies nicht möglich war – gemittelt. Die Reliabilität der gepoolten Prädiktoren wurde mit der Formel nach Mosier geschätzt (Hunter & Schmidt, 2004). Hierzu wurde das Programm von Schmidt und Le (2004) verwendet.

Zwei Studien wurden ausgeschlossen, da das Erfolgskriterium ungeeignet war bzw. das methodische Vorgehen nur unzureichend dokumentiert wurde. Für jede Stichprobe geht nur ein Wert in die Analyse ein, sämtliche Koeffizienten stammen somit von voneinander unabhängigen Stichproben.

Zur Bestimmung einer möglicherweise vorliegenden Publikationsverzerrung (publication bias) wurde dem Vorschlag von Wang und Bushman (1998) gefolgt und ein Normal-Quantil-Plot der korrigierten Effektstärken angefertigt. Dem Ergebnis zufolge liegt keine Publikationsverzerrung vor.

Im Ergebnisteil werden sowohl Konfidenz- als auch Kreditabilitätsintervalle berichtet. Das Kreditabilitätsintervall bezieht sich auf die Variabilität der Effekte über Studien hinweg. Hierzu wird die Varianz der korrigierten Validitäten σ_p^2 herangezogen. Ist die untere Grenze des 90%-Kreditabilitätsintervalls größer als Null, liegen 90% der ρ -Verteilung über Null. Werden 100% der Varianz durch die berücksichtigten Artefakte aufgeklärt, so entspricht der Kreditabilitätswert dem korrigierten Validitätskoeffizienten. Das Konfidenzintervall hingegen zeigt das Vertrauensintervall um ρ an. Hierzu wird der Stichprobenfehler herangezogen. Der Stichprobenfehler ist für die Berechnung des Kreditabilitätsintervalls irrelevant, da dieser bei der Schätzung von σ_p als Artefakt angesehen und eliminiert wird.

Artefaktquellen und Korrekturen

Es wurde eine Gewichtung der Einzelkorrelationen mit den Stichprobengrößen durchgeführt, um auf Grund der unterschiedlichen Stichprobengrößen zu erwartende Werteschwankungen zwischen den Primärstudien zu kontrollieren. Weiterhin wurden Korrekturen von Artefaktquellen durchgeführt, die zu einer artifiziellen Verwässerung der beobachteten Effektstärkevarianz beitragen. Bei allen Korrelationen wurde eine Kontrolle des Stichprobenfehlers vorgenommen. Auch die Reliabilität des Kriteriums sowie Varianzeinschränkungen (Selektion der Studierenden durch das durchlaufene Auswahlverfahren, durch Studienabbrüche bedingte Varianzeinschränkungen) wurden berücksichtigt (simultane Berücksichtigung der Artefakte durch einen compound artifact multipliziert; Hunter & Schmidt, 2004). Auf eine Korrektur der Reliabilität der Interviews wurde verzichtet.

Die Studien enthielten keine verlässlichen Informationen zur Reliabilität der Studiennoten. Daher wurde für die Korrektur der Reliabilität des Kriteriums *Durchschnittsnote* der von Kuncel, Hezlett und Ones (2001, 2004) ermittelte Wert $r_{tt} = .83$ verwendet. Die Reliabilität von *Einzelnoten* kann aus der Reliabilität der Durchschnittsnoten abgeleitet werden, indem die Spearman-Brown-Prophecy-Formel angewendet wird. Angenommen, eine Durchschnittsnote mit einer Reliabilität von $r_{tt} = .83$ besteht aus zwei Einzelnoten, so beträgt die Reliabilität der Einzelnoten gemäß Prophecy-Formel $r_{tt} = .71$. Setzt sich die Durchschnittsnote aus drei Einzelnoten zusammen, so ergibt sich eine Schätzung der Reliabilität der Einzelnoten von .62 usw.; neuere empirische Studien, die die Generalisierbarkeitstheorie zur Schätzung der Reliabilität von Hochschulnoten nutzen, kommen auf einen Bereich zwischen .55 und .70 für mündliche Prüfungen im Fach Psychologie (Novy, Kopel & Swank, 1996) und von .65 für mündliche Prüfungen im Fach Medizin (Wass, Wakeford, Neighbour & van der Vleuten, 2003). Für die Korrektur der Reliabilität von Einzelnoten wird bei der vorliegenden Metaanalyse der aus der Spearman-Brown-Prophecy-Formel extrapolierte Wert von $r_{tt} = .71$ angenommen (Gesamtnote, die sich aus zwei Einzelnoten zusammensetzt).

Über alle Einzelstichproben dieser Metaanalyse hinweg ergibt sich ein durchschnittlicher Korrekturwert für die Reliabilität des Kriteriums von $r_{tt} = .82$ ($SD = .02$).

Weiterhin werden die Validitätskoeffizienten hinsichtlich vorhandener Einschränkungen der Variabilität korrigiert. Der Wertebereich der u-Koeffizienten liegt bei den einbezogenen Stichproben zwischen .50 und 1.0, der Mittelwert beträgt .82 ($SD = .09$). Somit liegen die Variabilitätseinschränkungskorrekturen in der Größenordnung des von Kuncel et al. (2001) aus insgesamt 3466 einzelnen Varianz-Quotienten berechneten mittleren u-Werts von .76. Der Wert fällt in der vorliegenden Metaanalyse etwas höher aus, da die Analyse von Kuncel et al. (2001) Leistungstests zum Gegenstand hatte. Bei dieser Verfahrensklasse sind strengere Selektionsquoten als bei Interviews üblich.

Moderatoranalysen

Da sich der Strukturiertheitsgrad von Interviews bereits in mehreren Metaanalysen zur Prognose des Berufserfolgs als bedeutsam erwiesen hat (Cortina, Goldstein, Payne, Davison & Gilliland, 2000; Huffcutt et al., 1996; Huffcutt & Arthur, 1994; Wright, Lichtenfels & Pursell, 1989; Wiesner & Cronshaw, 1988), wird erwartet, dass sich dieser Effekt auch im Kontext der Studienzulassung replizieren lässt. Angesichts der oftmals nur kurzen und unzureichenden Beschreibung der Auswahlgespräche in den Primärstudien wurde analog zu Huffcutt, Conway, Roth und Stone (2001, p. 902) das Merkmal Strukturiertheitsgrad dichotom konzipiert. Als *strukturiert* wurden solche Interviews klassifiziert, in denen allen Bewerbern eine Reihe von vorgegebenen anforderungsbezogenen Fragen in gleichartiger Weise gestellt wurden. Des Weiteren sollte die Bewertung auf unterschiedlichen Dimensionen erfolgen, die im Anschluss an das Gespräch nach einem vorab festgelegten Schema kombiniert wurden. Interviews, die nicht mindestens zwei der Kriterien Anforderungsbezug, vorgegebene Fragen und multiple Bewertungsdimensionen erfüllten, galten als *unstrukturiert*.

Als weitere potenzielle Moderatoren der Validität wurden das Jahr der Publikation, der zeitliche Abstand zwischen Interviewdurchführung und Kriteriumsleistung, das mittlere Alter der Probanden, die Heterogenität der Stichprobe in Bezug auf unterschiedliche Hochschulen, verschiedene Jahrgänge und verschiedene Fächer, die Anzahl der Abstufungen im Kriterium, der Studienabschnitt, in dem die Kriterienleistung erbracht worden ist, die Art der Publikation (Peer reviewed- vs. nicht Peer reviewed-Publikation) sowie das Studienfach kodiert.

Zur Entscheidung, ob ein Moderator vorliegt, wird sowohl der Q-Test (Hedges & Olkin, 1985) als auch die 75-Prozent-Regel nach Hunter und Schmidt (2004) herangezogen. Diese beiden Kennwerte geben Hinweise auf das Vorhandensein von Moderatoren. Da der Q-Test in vielen Situationen eine zu geringe Power aufweist (Harwell, 1997), wird zusätzlich die 75-Prozent-Regel verwendet. Die 75-Prozent-Regel zieht zur Entscheidung, ob ein Moderator vorliegt, die Varianzaufklärung an den Studieneffekten durch die Artefakte (in der vorliegenden Studie: Stichprobenumfang, Unreliabilität Kriterium, Variabilitätseinschränkung) heran. Liegt diese über 75 Prozent, so wird davon ausgegangen, dass kein Moderator vorliegt. Diese „Daumenregel“ hat sich als relativ robust erwiesen (Sackett, Harris & Orr, 1986).

Ergebnisse

Die ermittelten Validitätskoeffizienten weisen einen unkorrigierten ungewichteten Mittelwert von $r = .15$ ($SD = .16$) und einen unkorrigierten gewichteten Mittelwert von $r = .11$ ($SD = .11$) auf. Die geringste Validität beträgt $-.13$ (Posthuma & Sommerfreund, 1985), die höchste $.59$ (Tompkins & Harkins, 1990). Der unkorrigierte ungewich-

Tabelle 3. Ergebnisse der Metaanalyse differenziert nach Form des Interviews

	N	k	ρ	σ_p^2	90% KW	%	SE_{FE}	95% KI _{FE}	SE_{RE}	95% KI _{RE}
Interview, alle Formen	4305	44 ¹	.126	.004	.041	72.4	.016	[.094, .158]	.019	[.089, .163]
Korrektur Reliabilität Kriterium										
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			.157	.002	.101	90.1	.020	[.118, .196]	.021	[.116, .198]
Interview, unstrukturiert	2136	14 ²	.076	.002	.026	83.5	.023	[.030, .122]	.026	[.026, .126]
Korrektur Reliabilität Kriterium										
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			.109	.000	.109	100.0	.032	[.046, .171]	.028	[.054, .164]
Interview, strukturiert	2169	30 ²	.175	.002	.114	86.9	.023	[.131, .219]	.024	[.128, .223]
Korrektur Reliabilität Kriterium										
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			.205	.002	.142	89.4	.026	[.154, .256]	.028	[.151, .259]

Anmerkungen: ¹ RE-Modell sollte zur Interpretation herangezogen werden. ² FE-Modell sollte zur Interpretation herangezogen werden. N = Größe der Gesamtstichprobe; k = Anzahl der unabhängigen Stichproben; ρ = Mittlere korrigierte Validität; σ_p^2 = Geschätzte Varianz von ρ ; 90% KW = 90%-Kreditabilitätswert für die Verteilung der wahren Werte; % = Prozentualer Anteil der Gesamtvarianz, der durch die Artefakte erklärt wird; SE_{FE} = Standardfehler ρ ; Modell: Feste Effekte; 95% KI_{FE} = Konfidenzintervall mit $p = 95\%$ um ρ ; Modell: Zufällige Effekte; SE_{RE} = Standardfehler ρ ; Modell: Zufällige Effekte; 95% KI_{RE} = Konfidenzintervall mit $p = 95\%$ um ρ ; Modell: Zufällige Effekte.

tete Wert fällt höher als der unkorrigierte gewichtete Wert aus, da zwei Studien mit großem Stichprobenumfang relativ geringe Validitätskennwerten aufweisen (Gehrlein, Dipboye & Shahani, 1993; Rahbar, Vellani, Sajan, Zaidi & Akbarali, 2001). Durch Anwendung der Gewichtung- und Korrekturalgorithmen ergeben sich eine mittlere korrigierte Validität von $\rho = .13$ nach Korrektur der Reliabilität des Kriteriums und von $\rho = .16$ nach zusätzlicher Korrektur der Variabilitätseinschränkung (Tabelle 3). Die unteren Grenzen der 90 %-Kredibilitätsintervalle für die Verteilung der wahren Werte liegen für beide Fälle über Null, so dass eine positive Validität für Auswahlgespräche generalisiert werden kann. Auch die Konfidenzintervalle umfassen den Wert Null nicht.

Durch die Artefakte werden 72.4 bzw. 90.1 Prozent der Varianz der wahren Werte aufgeklärt. Gemäß der 75-Prozent-Regel würde man bei einer Varianzaufklärung unter 75-Prozent die Existenz von Moderatoren der Validität prüfen, bei einer Varianzaufklärung über diesem Schwellenwert von einer weiteren Prüfung absehen (Schmidt, Hunter, Pearlman & Hirsh, 1985). Weiteren Aufschluss über das Vorhandensein von Moderatoren der Validität liefert der Q-Test, der signifikant ausfällt, also für das Vorliegen von Moderatoren spricht (RE-Modell, Korrektur Reliabilität Kriterium und Variabilitätseinschränkung; $Q_{Total} = 59.77, p < .05$).

Zur Prüfung der Relevanz der insgesamt 11 potenziellen Moderatoren der Validität wurden für jeden Moderator separate Q-Tests berechnet (Annahmen: RE-Modell, Korrektur Reliabilität Kriterium und Variabilitätseinschränkung). Da insgesamt 11 Moderatoren mittels Q-Tests geprüft werden, werden die Signifikanzgrenzen mittels Bonferroni-Adjustierung angepasst. Um eine Irrtumswahrscheinlichkeit von $p < .05$ bzw. $p < .01$ über alle Tests hinweg zu erreichen, müssen die Grenzen für die einzelnen Q-Tests bei $p < .005$ bzw. $p < .001$ liegen. Unter diesen Bedingungen wird keiner der Moderatoren signifikant.

Legt man bei den durchgeführten Q-Tests eine Fehlerwahrscheinlichkeit von $p = .05$ statt der adjustierten Fehlerwahrscheinlichkeit für jede einzelne der 11 Moderatoranalysen zu Grunde, dann wird der Moderator *Strukturiertheitsgrad der Interviews* signifikant ($Q_{Between} = 4.50, p < .05$).

Die Validität unstrukturierter Interviews beträgt $\rho = .08$ (Korrektur Reliabilität Kriterium) bzw. $\rho = .11$ (Korrektur Reliabilität Kriterium und Variabilitätseinschränkung), die Validität strukturierter Interviews $\rho = .18$ bzw. $\rho = .21$ (Tabelle 3). Die Varianzaufklärung durch die Artefakte liegt innerhalb der Moderatorabstufungen des Moderators Strukturiertheitsgrad über 75 Prozent. Gemäß der 75-Prozent-Regel ist also zu erwarten, dass kein weiterer Moderator wirksam ist. Die untere Grenze des 90 %-Kredibilitätsintervalls zeigt zudem an, dass für beide Interviewformen eine *positive* Validität generalisiert werden kann. Auch die auf Basis des Standardfehlers von ρ berechneten Konfidenzintervalle enthalten den Wert 0 nicht. Die mittlere korrigierte Validität von strukturierten und unstrukturierter Interviews ist somit signifikant von 0 unterschiedlich.

Um die mögliche Konfundierungen des Moderators Strukturiertheitsgrad mit den anderen betrachteten Moderatoren zu prüfen, wurden die Zusammenhänge mittels bivariater Korrelationen bzw. Chi-Quadrat-Tests geprüft. Lediglich die Variable Jahrgangsheterogenität vs. Jahrgangshomogenität (wurden zur Berechnung des Validitätskoeffizienten mehrere Kohorten zusammengefasst oder bezieht sich die Validitätsangabe auf *einen* Jahrgang) kovariiert signifikant ($r = .45, p < .01$) mit dem Moderator Strukturiertheitsgrad. Werden diese beiden Moderatoren simultan analysiert, so wird das regressionsanalytische Gesamtmodell nicht signifikant.

Von großer Bedeutung ist die *inkrementelle Validität* von Auswahlgesprächen gegenüber anderen Zulassungsverfahren. Die inkrementelle Validität zu Schulnoten und zu Studierfähigkeitstests interessiert besonders, da diese beiden Verfahren die höchste eigenständige prognostische Validität erbringen und zudem am häufigsten eingesetzt werden (Heine et al., 2006; Hell et al., 2007; Trapmann et al., in Druck). Für die Kalkulation der inkrementellen Validität müssen die eigenständige Validität der anderen Verfahren sowie die Interkorrelation zwischen Auswahlgesprächen und den anderen Verfahren bekannt sein. Die eigenständige Validität von Schulnoten kann mit .517 (Trapmann et al., in Druck) veranschlagt werden, die Validität von Studierfähigkeitstests mit .478 (Hell et al., 2007). Zur Bestimmung der Interkorrelation zwischen Auswahlgesprächen und den beiden anderen Auswahlverfahren wurden die dieser Metaanalyse zu Grunde liegenden Primärstudien nach entsprechenden Angaben durchgesehen. Interkorrelationen zwischen Auswahlgesprächen und Schulnoten berichten Alvidres und Whitworth (1981), Höschl und Kozeny (1997), Meredith, Dunlap und Baker (1982) sowie Shahani, Dipboye und Gehrlein (1991). Zur Bestimmung der Interkorrelation zwischen Auswahlgesprächen und Schulnoten wurden die gefundenen Interkorrelationen metaanalytisch zusammengefasst. Es ergibt sich eine hinsichtlich der Variabilitätseinschränkung korrigierte Interkorrelation von $\rho = .16$ (Varianzaufklärung durch die Artefakte: 100 %). Interkorrelationen zwischen kognitiven Fähigkeitstests und Auswahlgesprächen berichten Alvidres und Whitworth (1981), Hell (1998), Shamai, Gafni und Dekel (1995) sowie Sue-Chan und Latham (2004). Eine metaanalytische Verdichtung der berichteten Interkorrelationen ergibt einen korrigierten Wert von $\rho = .12$ (Varianzaufklärung durch die Artefakte: 53.4 %; 90 %-Kredibilitätsintervall reicht von .01 bis .23). Zur Bestimmung der inkrementellen Validität wurden auf Basis der ermittelten vollständigen Interkorrelationsmatrizen multiple Regressionen berechnet, in die in einem ersten Schritt (Modell 1) das validere und in einem zweiten Schritt (Modell 2) auch das weniger valide Verfahren als Prädiktor aufgenommen wurden. Bei den Schulnoten wurde angenommen, dass die Interkorrelation mit Interviews .16 beträgt, da die Artefakte bei der Schätzung der Interkorrelation die gesamte Varianz der Primärstudieneffekte aufklären, das Kredibilitätsintervall also ρ entspricht. Bei den Studierfähigkeitstests wurden zusätzlich zu der geschätzten Interkorrelation zwischen Testergebnissen und Interviewergebnissen von .12 auch noch die unteren und

Tabelle 4. Inkrementelle Validität von Auswahlgesprächen

Prädiktor/Modell	Prädiktorkombination	β	R	R^2
Inkrementelle Validität gegenüber Schulnoten				
Annahme: $r_{\text{Schulnoten, Interview}} = .16$				
Modell 1	Schulnoten	.517	.517	.267
Modell 2	Schulnoten	.505	.522	.273
	Interview	.076		
Inkrementelle Validität gegenüber Studierfähigkeitstests (SFT)				
Annahme: $r_{\text{SFT, Interview}} = .01$				
Modell 1	Studierfähigkeitstest	.478	.478	.228
Modell 2	Studierfähigkeitstest	.476	.502	.252
	Interview	.152		
Annahme: $r_{\text{SFT, Interview}} = .12$				
Modell 1	Studierfähigkeitstest	.478	.478	.228
Modell 2	Studierfähigkeitstest	.466	.489	.239
	Interview	.103		
Annahme: $r_{\text{SFT, Interview}} = .23$				
Modell 1	Studierfähigkeitstest	.478	.478	.228
Modell 2	Studierfähigkeitstest	.467	.480	.231
	Interview	.050		

oberen Kreditabilitätswerte (.01 bzw. .23) als geschätzte untere und obere Grenze der Interkorrelation zwischen Testergebnissen und Interviewergebnissen angenommen. Wie in Tabelle 4 zu sehen ist, kann durch den zusätzlichen Einsatz von Auswahlgesprächen die eigenständige Validität der Schulnoten von .517 auf .522 und die eigenständige Validität von Studierfähigkeitstests von .478 auf .502, .489 bzw. .480 (je nach Annahme hinsichtlich der Interkorrelation zwischen den Prädiktoren) gesteigert werden.

Diskussion

Mit der vorliegenden Metaanalyse können die Befunde der einbezogenen Primärstudien integriert werden. Der größte Teil der Varianz der Validitätskoeffizienten wird durch die analysierten Artefakte Stichprobenumfang, Unreliabilität der Kriterien und Variabilitätseinschränkung erklärt. Im Ergebnis kann eine mit $\rho = .16$ zwar niedrige, aber durch das Kreditabilitätsintervall und auch das Konfidenzintervall abgesicherte positive Validität generalisiert werden. Die Validität von Auswahlgesprächen wird durch ihren Strukturierungsgrad moderiert. Diese Moderatorwirkung wurde bereits im Kontext der beruflichen Eignungsdiagnostik für die Validität von Interviews für beruflichen Erfolg nachgewiesen. Trivial ist dieses Erkenntnis allerdings nicht, da sich der Strukturierungsgrad von Interviews bei der Prognose des Erfolgs bei *beruflichen Trainingsmaßnahmen* nach McDaniel et al. (1994) nicht als Moderator auswirkt.

Worauf kann die Überlegenheit strukturierter Interviews zurückgeführt werden? Campion, Palmer und Cam-

pion (1997) konnten eine Reihe von Effekten der Interviewstrukturierung identifizieren: Sie wirkt sich u. a. positiv auf die Testwiederholungsreliabilität, die Interraterreliabilität (korrelative Maße), die Beurteilerübereinstimmung (Mittelwertsdifferenzen) und die interne Konsistenz von Interviewbeurteilungen aus. Weiterhin wirkt sich die Strukturierung nach Campion et al. (1997) positiv auf die Validität aus, indem die Messdefizienz (Ausmaß der nützlichen Information) und die Kontamination durch Fehler-effekte (z. B. positive Selbstdarstellung) reduziert werden.

Bei der Interpretation der vorliegenden Ergebnisse ist zu berücksichtigen, dass mit $k = 22$ die Hälfte der eingegangenen Stichproben aus den medizinischen Disziplinen stammen. Es ist denkbar, dass die unausgewogene Verteilung der Stichproben über verschiedene Studiengänge dafür verantwortlich ist, dass das Studienfach die Validität gemäß den angestellten Analysen nicht moderiert.

Im Vergleich zu Schulnoten und Leistungstests fällt die Validität von Auswahlgesprächen für das Studienerfolgskriterium Studiennoten gering aus. Die korrigierte mittlere Validität unstrukturierter Interviews von $\rho = .11$ bedeutet einen gemeinsamen Varianzanteil von Interviewergebnissen und Studiennoten von 1.2 Prozent. Als Prädiktor für Studiennoten sind unstrukturierte Interviews demnach unbrauchbar. Durch strukturierte Interviews können 4.2 Prozent der Kriterienvarianz erklärt werden. Die in Tabelle 4 aufgeführten inkrementellen Zuwächse in der Prognosekraft durch die gemeinsame Verwendung von Auswahlgesprächen mit Schulnoten bzw. Studierfähigkeitstests sind vom Betrag her so gering, dass der zusätzliche Einsatz von Interviews zur Prognose von *Studiennoten* nicht als sinnvoll erachtet werden kann, wenn bereits

valide Informationen wie Schulnoten oder Ergebnisse von Studierfähigkeitstests vorliegen. Die empirische Evidenz steht somit im Widerspruch zu der subjektiven Einschätzung der Interviewer (Wagoner & Gray, 1979) und auch der Bewerber (Hell & Schuler, 2005), dass Interviews besonders aussagekräftig sind.

Hinsichtlich anderer proximaler Erfolgskriterien wie Studiendauer, Zufriedenheit und Studienabbruch oder Berufserfolg als distales Studienerfolgskriterium liegen zu wenige Primärstudien vor, um eine zuverlässige Aussage über die Validität treffen zu können. Es ist denkbar, dass Auswahlgespräche für alternative Erfolgskriterien über eine größere Aussagekraft verfügen. Vorausgesetzt, die in der beruflichen Eignungsdiagnostik gewonnenen Erkenntnisse hinsichtlich der positiven Wirkung realistischer Tätigkeitsinformationen auf Fluktuationsraten (Phillips, 1998) können auf den Studienkontext übertragen werden, sollten zumindest Interviews, in denen Bewerber über Studieninhalte und -bedingungen informiert werden, für das Studienerfolgskriterium (fehlender) Studienabbruch prognostisch valide sein.

Die günstige Prognosekraft strukturierter Interviews für Berufserfolg hingegen ist durch mehrere Metaanalysen belegt (zusammenfassend Schuler, 2002). Eine *auf den Berufserfolg ausgerichtete Auswahl* von Studierenden ist daher in Fächern denkbar, mit denen ein klar definiertes Berufsbild (z. B. Lehramtsstudiengänge) verbunden ist. Auch für die Auswahl von Teilnehmern an Graduiertenschulen bzw. Doktoranden könnten Interviews auf Grund des klareren Zuschnitts des Tätigkeitsbereichs sinnvoll sein.

Für die unterschiedlichen Validitäten von Interviews für Studiennoten einerseits und Berufserfolg andererseits können mehrere Aspekte ursächlich sein. Als mögliche Erklärung bietet sich die Messqualität des Kriteriums an: Metaanalysen zur Prognose des Berufserfolgs nehmen meist niedrige Reliabilitäten für die Kriterienmessung an, was nach Korrektur zu entsprechend erhöhten Validitätskoeffizienten führt. Legte man – in der Annahme, dass die geschätzten Reliabilitäten der Studiennoten in der vorliegenden Metaanalyse zu optimistisch gewählt worden sind – die von McDaniel et al. (1994) verwendete geschätzte Kriterienreliabilität von .60 zu Grunde, so stiege die korrigierte Validität von $\rho = .11$ auf $\rho = .13$ (unstrukturierte Interviews) und von $\rho = .21$ auf $\rho = .24$ (strukturierte Interviews); beide Werte sind somit immer noch weit entfernt von den Koeffizienten McDaniels und seiner Kollegen, die für unstrukturierte Interviews eine korrigierte Validität für Berufserfolg von .33 und für strukturierte Interviews eine korrigierte Validität von .44 berechneten. Über weitere Erklärungen für die abweichende Validität kann nur spekuliert werden. Die Reliabilität von Auswahlgesprächen im Rahmen der Hochschulzulassung könnte geringer sein, da diese vermutlich zeitlich kürzer ausfallen als Interviews im Rahmen der Personalauswahl. Weiterhin ist denkbar, dass durch Personalfachleute geführte Interviews valider sind als von ungeübten Hochschullehrern geführte. Eine weitere Erklärung könnte darin liegen, dass sich in Interview-Ergebnissen „soziale“ oder „kommunikative Kompe-

tenzen“ widerspiegeln, die für den beruflichen Erfolg wichtiger sind als für Studiennoten. Zudem ist die vergleichsweise kurze Biographie von Studienplatzbewerbern vermutlich weniger aussagekräftig (vgl. günstige Validität der Berufserfahrung und des beruflichen Wissens für Auswahlgespräche im Kontext der Personalauswahl bei Huffcutt et al., 2001).

Zahlreiche der in diese Metaanalyse einbezogenen Einzelstudien berichten ihre Analysen und Ergebnisse nur partiell, so dass Parameter wie die Reliabilität der Noten und die Selektionsquoten teilweise geschätzt werden mussten. Es ist wünschenswert, dass bei zukünftigen Publikationen von empirischen Einzelstudien Inhalt, Durchführung und Auswertungsmodi des Interviews beschrieben werden und die relevanten Kennwerte (Selektionsquote, Dropoutquote, Reliabilitäten) berichtet werden. Weiterhin wäre es für die individuelle Bewertung der Einzelstudien und ganz besonders für metaanalytische Betrachtungen der Konstruktvalidität von Auswahlgesprächen notwendig, dass die Interkorrelationen zwischen Interviewergebnissen und anderen eingesetzten Verfahren dokumentiert werden.

Auswahlgespräche im Rahmen der Hochschulzulassung müssen nicht zwangsläufig als reines Selektionsinstrument angesehen werden. Vielmehr stellt sich die Zulassungsproblematik aus gesellschaftlicher Perspektive eher als Zuordnungsentscheidung denn als Selektionsentscheidung dar. In Interviews können für abgelehnte Bewerber Platzierungsempfehlungen entwickelt werden und bei Bedarf kann auf Beratungsangebote hingewiesen werden. Solche Zusatzfunktionen sollten bei Überlegungen, Interviews einzuführen, berücksichtigt werden.

Literatur

- * Alvidres, M. D. & Whitworth, R. H. (1981). The development and validation of an entrance examination in a Mexican university. *Educational and Psychological Measurement*, 41, 503–509.
- * Balogun, J. A., Karacoloff, L. A. & Farina, N. T. (1986). Predictors of academic achievement in physical therapy. *Physical Therapy*, 66, 976–980.
- * Balogun, J. A. (1988). Predictors of academic and clinical performance in a baccalaureate physical therapy program. *Physical Therapy*, 68, 238–242.
- Burton, N. W. & Ramist, L. (2001). Predicting success in college: SAT® studies of classes graduating since 1980. *The College Board Research Report*, 2001-2.
- * Calkins, E. V., Willoughby, T. L. & Arnold, L. M. (1982). Predictors of performance of minority students in the first two years of a BA/MD program. *Journal of the National Medical Association*, 74, 625–632.
- Campion, M. A., Palmer, D. K. & Campion, J. E. (1997). A review of structure in the selection interview. *Personnel Psychology*, 50, 655–702.
- Cortina, J. M., Goldstein, N. B., Payne, S. C., Davison, H. K. & Gilliland, S. W. (2000). The incremental validity of interview scores over and above cognitive ability and conscientiousness scores. *Personnel Psychology*, 53, 325–351.
- Deidesheimer Kreis (1997). *Hochschulzulassung und Studieneignungstests*. Göttingen: Vandenhoeck und Ruprecht.

- Deutsche Gesellschaft für Psychologie (1997). Stellungnahme der Deutschen Gesellschaft für Psychologie (DGPs) zu Auswahlgesprächen der Hochschulen bei der Vergabe von Studienplätzen. *Psychologische Rundschau*, 48, 107–108.
- Drolshagen, S. (1996). Der Stellenwert von Interviews beim Zulassungsverfahren zum Medizinstudium im internationalen Vergleich. In R. Lohölter, K. Hinrichsen, G. Trost & S. Drolshagen (Hrsg.), *Das Interview bei der Zulassung zum Medizinstudium* (S. 81–130). Stuttgart: Schattauer.
- Erez, A., Bloom, M. C. & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, 49, 275–306.
- * Gehrlain, T. M., Dipboye, R. L. & Shahani, C. (1993). Nontraditional validity calculations and differential interviewer experience: Implications for selection interviews. *Educational and Psychological Measurement*, 53, 457–469.
- Goldberg, E. L. & Alliger, G. M. (1992). Assessing the validity of the GRE for students in psychology: A validity generalization approach. *Educational and Psychological Measurement*, 52, 1019–1027.
- Harwell, M. (1997). An empirical study of Hedges's homogeneity test. *Psychological Methods*, 2, 219–231.
- Hedges, L. V. & Olkin, I. (1985). *Statistical methods for meta-analysis*. London: Academic Press.
- Hedges, L. V. & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3, 486–504.
- Heine, C., Briedis, K., Didi, H.-J., Haase, K. & Trost, G. (2006). *Bestandsaufnahme von Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern*. Hannover: HIS-Kurzinformation A 3/2006.
- * Hell, B. (1998). *Evaluation des Aufnahmeverfahrens an der Fachhochschul-Studiengänge-Vorarlberg GmbH unter besonderer Berücksichtigung der prognostischen Validität*. Unveröffentlichte Diplomarbeit, Universität Bonn.
- Hell, B. & Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. *Empirische Pädagogik*, 19, 361–376.
- Hell, B., Trapmann, S. & Schuler, H. (2007). *Eine Metaanalyse der Validität von Studierfähigkeitstests im deutschsprachigen Raum*. Manuskript eingereicht zur Publikation.
- Heyneman, S. P. & Fägerlind, I. (Eds.). (1988). *University examinations and standardized testing*. World Bank Technical Paper Number 78. Washington, DC: The World Bank.
- * Hobfoll, S. E. & Benor, D. E. (1981). Prediction of student clinical performance. *Medical Education*, 15, 231–236.
- * Holt, T. B. O. & Dunlevy, C. L. (1992). The use of preadmission criteria to predict academic success in a 4-year respiratory care curriculum. *Respiratory Care*, 37, 439–443.
- * Hosford, R. E., Johnson, M. E. & Atkinson, D. R. (1984). Academic criteria, experiential background, and personal interviews as predictors of success in a counselor education program. *Counselor Education and Supervision*, 23, 268–275.
- * Höschl, C. & Kozeny, J. (1997). Predicting academic performance of medical students: The first three years. *American Journal of Psychiatry*, 154, 87–92.
- Huffcutt, A. I. & Arthur, W. (1994). Hunter and Hunter (1984) revisited: Interview validity for entry-level jobs. *Journal of Applied Psychology*, 79, 184–190.
- Huffcutt, A. I., Conway, J. M., Roth, P. L. & Stone, N. J. (2001). Identification and meta-analytic assessment of psychological constructs measured in employment interviews. *Journal of Applied Psychology*, 86, 897–913.
- Huffcutt, A. I., Roth, P. L. & McDaniel, M. A. (1996). A meta-analytic investigation of cognitive ability in employment interview evaluations: Moderating characteristics and implications for incremental validity. *Journal of Applied Psychology*, 81, 459–474.
- Hunter, J. E. & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, 8, 275–292.
- Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis – correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.
- Hunter, J., Schmidt, F. & Jackson, G. (1982). *Meta-analysis: cumulating research findings across studies*. London: Sage.
- Janz, T. (1982). Initial comparisons of patterned behaviour description interviews versus unstructured interviews. *Journal of Applied Psychology*, 67, 577–580.
- * Kulatunga-Moruzi, C. & Norman, J. R. (2002). Validity of admissions measures in predicting performance outcomes: The contribution of cognitive and non-cognitive dimensions. *Teaching and Learning in Medicine*, 14, 34–42.
- Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2001). A comprehensive meta-analysis of the predictive validity of the graduate record examinations: Implications for graduate student selection and performance. *Psychological Bulletin*, 127, 162–181.
- Kuncel, N. R., Hezlett, S. A. & Ones, D. S. (2004). Academic performance, career potential, creativity, and job performance: Can one construct predict them all? *Journal of Personality and Social Psychology*, 86, 148–161.
- Latham, G. P., Saari, L. M., Pursell, E. D. & Campion, M. A. (1980). The situational interview. *Journal of Applied Psychology*, 65, 422–427.
- * Levine, S. B., Knecht, H. G. & Eisen, R. G. (1986). Selection of physical therapy students: Interview methods and academic predictors. *Journal of Allied Health*, 15, 151.
- * Lucas, C. J. & Schmitz, C. D. (1991). A comparison of academic and affective variables as predictors of student-teaching performance. *College Student Journal*, 25, 518–531.
- Linn, R. L. & Hastings, C. N. (1984). A meta analysis of the validity of predictors of performance in law school. *Journal of Educational Measurement*, 21, 245–259.
- McDaniel, M. A., Whetzel, D. L., Schmidt, F. L. & Maurer, S. D. (1994). The validity of employment interviews: A comprehensive review and meta-analysis. *Journal of Applied Psychology*, 79, 599–616.
- * McManus, I. C. & Richards, P. (1986). Prospective survey of performance of medical students during preclinical years. *British Medical Journal*, 293, 124–127.
- * Meister, K. (2004). *Prognose des Erfolgs: Eine Untersuchung der Validität des Auswahlverfahrens der HfB für Studienanfänger*. Unveröffentlichte Diplomarbeit, Hochschule für Bankwirtschaft, Frankfurt am Main.
- Mendoza, J. L. & Reinhardt, R. N. (1991). Validity generalization procedures using sample-based estimates: A comparison of six procedures. *Psychological Bulletin*, 110, 596–610.
- * Meredith, K. E., Dunlap, M. R. & Baker, H. H. (1982). Subjective and objective admissions factors as predictors of clinical clerkship performance. *Journal of Medical Education*, 57, 743–751.
- Mitchell, K. J. (1990). Traditional predictors of performance in medical school. *Academic Medicine*, 65, 149–158.
- Morris, J. G. (1999). The value and role of the interview in the student admissions process: a review. *Medical Teacher*, 21, 473–481.
- Morrison, T. & Morrison, M. (1995). A meta-analytic assessment of the predictive validity of the quantitative and verbal components of the Graduate Record Examination with graduate grade point average representing the criterion of graduate success. *Educational and Psychological Measurement*, 55, 309–316.
- * Munro, B. H. (1985). Predicting success in graduate clinical specialty programs. *Nursing Research*, 34, 54–57.
- Novy, D. M., Kopel, K. F. & Swank, P. R. (1996). Psychometrics of oral examinations for psychology licensure: The Te-

- xas examination as an example. *Professional Psychology: Research & Practice*, 27, 415–417.
- Peers, I. S. & Johnston, M. (1994). Influence of learning context on the relationship between A-level attainment and final degree performance: A meta-analytic review. *British Journal of Educational Psychology*, 64, 1–18.
- Phillips, J. M. (1998). Effects of realistic job previews on multiple organizational outcomes: A meta-analysis. *Academy of Management Journal*, 41, 673–690.
- * Posthuma, B. W. & Sommerfreund, J. (1985). Examination of selection criteria for a program in occupational therapy. *American Journal of Occupational Therapy*, 39, 441–445.
- * Rahbar, M. H., Vellani, C., Sajan, F., Zaidi, A. A. & Akbarali, L. (2001). Predictability of medical students' performance at the Aga Khan University from admission test scores, interview ratings and systems of education. *Medical Education*, 35, 374–380.
- * Ross, M. W. (1983). Predictive value of selection interviews in success in first-year medicine. *Medical Journal of Australia*, 10, 423–424.
- Raju, N. S., Burke, M. J., Normand, J. & Langlois, G. M. (1991). A new meta-analytic approach. *Journal of Applied Psychology*, 76, 432–446.
- Raju, N. S. & Fleer, P. F. (2003). *VG2M: A computer program for conducting validity generalization analysis*. Chicago, IL: Illinois Institute of Technology.
- Rindermann, H. & Oubaid, V. (1999). Auswahl von Studienanfängern durch Universitäten – Kriterien, Verfahren und Prognostizierbarkeit des Studienerfolgs. *Zeitschrift für Differentielle und Diagnostische Psychologie*, 20, 172–191.
- Rustenbach, S. J. (2003). Metaanalyse – Eine anwendungsorientierte Einführung. In K. Pawlik (Hrsg.), *Methoden der Psychologie*. Bern: Verlag Hans Huber.
- Sackett, P. R., Harris, M. M. & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlational data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology*, 71, 302–310.
- Schmidt, F. L., Hunter, J. E., Pearlman, K. & Hirsh, H. R. (1985). Forty questions about validity generalization and meta-analysis. *Personnel Psychology*, 38, 697–798.
- Schmidt, F. L. & Le, H. (2004). *Software for the Hunter-Schmidt meta-analysis methods*. University of Iowa, Department of Management & Organization, Iowa City, IA.
- Schuler, H. (1992). Das Multimodale Einstellungsinterview. *Diagnostica*, 38, 281–300.
- Schuler, H. (2002). *Das Einstellungsinterview*. Göttingen: Hogrefe.
- Schuler, H., Funke, U. & Baron-Boldt, J. (1990). Predictive validity of school grades: A meta-analysis. *Applied Psychology: An International Review*, 39, 89–103.
- Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.
- * Shahani, C., Dipboye, R. L. & Gehrlein, T. M. (1991). The incremental contribution of an interview to college admissions. *Educational and Psychological Measurement*, 51, 1049–1061.
- * Shamai, S., Gafni, O. & Dekel, Y. (1995). The admission tests to different programs of teachers' training college as predictors of the students' success. *College Student Journal*, 29, 171–176.
- * Sue-Chan, C. & Latham, G. P. (2004). The situational interview as a predictor of academic and team performance: A study of the mediating effects of cognitive ability and emotional intelligence. *International Journal of Selection and Assessment*, 12, 312–320.
- * Tompkins, L. S. & Harkins, C. J. (1990). Predicting academic success in a nontraditional program. *Journal of Allied Health*, 19, 15–24.
- Trapmann, S., Hell, B., Weigand, S. & Schuler, H. (in Druck). Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*.
- Trost, G. (1996). Die Bedeutung des Interviews für die Diagnose der Studieneignung. Darstellung der internationalen Forschungsergebnisse. In R. Lohölter, K. Hinrichsen, G. Trost & St. Drolshagen (Hrsg.), *Das Interview bei der Zulassung zum Medizinstudium* (S. 49–80). Stuttgart: Schattauer.
- * Trost, G. & Piel, P. (1991). Der Aufnahmewettbewerb der WHU – eine Zwischenbilanz. *Hochschulnachrichten aus der Wissenschaftlichen Hochschule für Unternehmensführung*, 6, 60–65.
- * Vargo, J. W., Madill, H. M. & Davidson, P. R. (1986). The pre-admission interview as a predictor of academic grades and fieldwork performance. *Canadian Journal of Occupational Therapy*, 53, 211–215.
- Wagoner, N. E. & Gray, G. T. (1979). Report on a survey of program directors regarding selection factors in graduate medical education. *Journal of Medical Education*, 54, 445–452.
- Wang, M. C. & Bushman, B. J. (1998). Using the normal quantile plot to explore meta-analytic data sets. *Psychological Methods*, 3, 46–54.
- * Walker, J. D., Killip, D. E. & Fuller, J. L. (1985). The significance of the admission interview in predicting students' performance in dental school. *Journal of Medical Education*, 60, 569–571.
- Wass, V., Wakeford, R., Neighbour, R. & Van der Vleuten, C. (2003). Achieving acceptable reliability in oral examinations: An analysis of the Royal College of General Practitioners membership examination's oral component. *Medical Education*, 37, 126–131.
- * Weiss, M., Lotan, I., Kedar, H. & Ben-Shakhar, G. (1988). Selecting candidates for a medical school: An evaluation of a selection model based on cognitive and personality predictors. *Medical Education*, 22, 492–497.
- Wiesner, W. H. & Cronshaw, S. F. (1988). A meta-analytic investigation of the impact of interview format and degree of structure on the validity of the employment interview. *Journal of Occupational Psychology*, 61, 275–290.
- Wright, P. M., Lichtenfels, P. A. & Pursell, E. D. (1989). The structured interview: Additional studies and a meta-analysis. *Journal of Occupational Psychology*, 62, 191–199.
- * Primärstudien, die in die Metaanalyse eingingen.

Dr. Benedikt Hell

Lehrstuhl für Psychologie (540F)
Universität Hohenheim
70593 Stuttgart
E-Mail: hell@uni-hohenheim.de