

Originalartikel

Die Validität von Schulnoten zur Vorhersage des Studienerfolgs – eine Metaanalyse¹

Sabrina Trapmann, Benedikt Hell, Sonja Weigand und Heinz Schuler

Lehrstuhl für Psychologie, Universität Hohenheim, Stuttgart

Zusammenfassung. Schulabschlussnoten sind trotz erweiterter Möglichkeiten der Universitäten und Kritik an der Hinlänglichkeit zur Vorhersage des Studienerfolgs nach wie vor das häufigste Einzelkriterium, das zur Auswahl von Studierenden in Deutschland herangezogen wird. Die vorliegende Metaanalyse integriert die seit 1980 publizierten europäischen Studien zum Zusammenhang von einzelnen und durchschnittlichen Schulnoten einerseits und dem Studienerfolg andererseits (insgesamt 83 Koeffizienten). Es ergeben sich mittlere korrigierte Validitätskoeffizienten in Höhe von .26 bis .53 für das Studienerfolgskriterium Studiennoten. Die höchste Validität weisen deutsche Schuldurchschnittsnoten auf ($\rho = .53$, korrigiert für die Reliabilität der Studiennoten und die Varianzeinschränkung durch Selektion). Moderatoranalysen zeigen signifikante Einflüsse des Landes, des Studienfachs, der Zeitspanne zwischen Schul- und Studienleistung und des Studienabschnitts der Studienleistung. Die für Schulnoten berechneten Validitätskoeffizienten sind generalisierbar positiv und von Null verschieden.

Schlüsselwörter: Metaanalyse, Schulnoten, Studienerfolgsprognose, Auswahl von Studierenden

The Validity of School Grades for Academic Achievement – A Meta-Analysis

Abstract. Although school grades are often criticized and universities in Germany are now allowed to use other criteria of selection, high school grades are still used most often to admit students to university. The present meta-analysis includes all European studies that have been published since 1980 and investigated the relationship between school grades and university grades (altogether 83 coefficients). Grade point average and individual subject grades were accepted as predictors. The mean corrected validity for school grades ranges from .26 to .53, using undergraduate or graduate grades to measure academic success. Highest validity was found for German grade point average ($\rho = .53$, corrected). Analyses of moderator-effects show significant influences of country of origin, study major, time between school and university grades, and study period. The computed validity coefficients for school grades are positive and different from zero.

Keywords: meta-analysis, school grades, academic performance prediction, academic admission criteria

1 Einleitung

Schulabschlussnoten sind trotz erweiterter Möglichkeiten der Universitäten durch die kürzlich erfolgte Revision des Hochschulrahmengesetzes von 1998 nach wie vor das häufigste Einzelkriterium, das zur Auswahl oder Vorauswahl von Studierenden in Deutschland herangezogen wird (Heine, Briedis, Didi, Haase & Trost, 2006). Eine Ursache hierfür ist sicherlich die leichte Verfügbarkeit der Schulabschlussnote bei Studienplatzbewerbern. Auch international kommt der schulischen Leistung bei der Zulassung zum

Hochschulstudium eine herausragende Rolle zu, da sie in fast allen Ländern als wesentliches und in zahlreichen Ländern als das einzige Entscheidungskriterium herangezogen wird (Deidesheimer Kreis, 1997).

Viele empirische Studien belegen die gute prognostische Validität der Schulnoten für einen späteren Studienerfolg mit Korrelationen um .40 (z. B. für deutsche Schulnoten Rindermann & Oubaid, 1999; Schmidt-Atzert, 2005; Steyer, Yousfi & Würfel, 2005). Davon zeugen auch mehrere Metaanalysen (Baron-Boldt, Schuler & Funke, 1988; Burton & Ramist, 2001; Robbins, Lauver, Le, Davis, Langley

¹ 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 Landesstiftung Baden-Württemberg durchgeführt wird.

& Carlstrom, 2004). Die Abiturdurchschnittsnote gilt als der valideste Einzelprädiktor für den Studienerfolg (Brandstätter & Farthofer, 2002; Schmidt-Atzert, 2005; Gold & Souvignier, 2005). Allerdings berechtigen einige Studien zu begründetem Zweifel an der Validität der Schulabschlussnoten. So fanden beispielsweise Peers und Johnston (1994) in ihrer Metaanalyse über 20 Studien aus den Jahren 1961 bis 1987 nur eine mittlere Validität von .28 für Schulnoten in Großbritannien.

Auch in Deutschland ist die Abiturnote trotz der gefundenen empirischen Zusammenhänge als Auswahlkriterium Kritik ausgesetzt. Die Vergleichbarkeit der Abiturnoten verschiedener Bundesländer und Schulen oder sogar die Vergleichbarkeit der Urteile verschiedener Lehrer wird häufig angezweifelt (Deidesheimer Kreis, 1997, S. 79 f.). Auch Untersuchungen zu einzelnen Noten zeigen weitgehend übereinstimmend, dass die mündlichen und schriftlichen Prüfungen oft nicht hinreichend objektiv, zuverlässig und valide sind, und dass deutliche Differenzen in den Bewertungsmaßstäben zwischen den Klassen und Fächern bestehen (Ingenkamp, 1997). Tent (2006) bescheinigt den Zensuren, dass sie «weder so schlecht, wie sie hingestellt wurden, noch so gut, wie sie ihrem Anspruch nach sein müssten» (S. 878) sind. Als alleiniges Entscheidungskriterium z. B. zur Studienzulassung sollten Lehrerurteile laut Tent (2006) allerdings nicht verwendet werden. Kritiker bemängeln an den Abiturnoten als Auswahlkriterium zur Studienzulassung vor allem auch die fehlende Spezifität im Hinblick auf die Anforderungen einzelner Studiengänge und/oder einzelner Hochschulen (Gold & Souvignier, 2005) und verlangen Auswahlkriterien, die die Eignung des Studienplatzbewerbers für einen speziellen Studiengang an einer bestimmten Hochschule ausweisen. Neben der Kritik an der mangelnden Vergleichbarkeit und Messgenauigkeit von Schulnoten wird auch die Frage gestellt, inwiefern die gymnasiale Bildung als Vorbereitung für ein Hochschulstudium angelegt ist. Wenngleich der erfolgreiche Abschluss der sekundären Bildungsstufe notwendige Voraussetzung für die tertiäre ist, werden Noten für eine Auswahl der Studierenden zwar häufig, aber keinesfalls unkritisch verwendet.

In den Vereinigten Staaten, die seit vielen Jahren Studierfähigkeitstests zur Auswahl von Studierenden einsetzen und zur Validierung dieser Verfahren groß angelegte

Studien durchführen, ist die gute Validität von Schulnoten belegt (Willingham, Lewis, Morgan & Ramist, 1990). Eine Metaanalyse bescheinigt dem *Scholastic Aptitude Test* (SAT) eine mittlere unkorrigierte Validität von .36, den Schulnoten sogar von .42 (Burton & Ramist, 2001). Durch die gemeinsame Verwendung von Test und Schulnoten kann die Vorhersagevalidität auf .52 gesteigert werden.

Nun lassen sich die US-amerikanischen Befunde aufgrund der unterschiedlichen Bildungssysteme sicherlich nicht ungeprüft auf europäische oder deutsche Verhältnisse übertragen, und seit der letzten im deutschsprachigen Raum durchgeführten Metaanalyse zur Validität von Schulnoten für die Studienleistung, die einen Zusammenhang von .46 feststellte (Baron-Boldt et al., 1988), sind beinahe 20 Jahre verstrichen. Daher werden die guten Validitätsbefunde der aktuelleren amerikanischen Metaanalysen und die neuerlich publizierten Primärstudien zu diesem Thema als Anlass für eine Überprüfung der Validität der europäischen Schulnoten für den Studienerfolg betrachtet. Die Ergebnisse stellen eine Ergänzung zu den vielfach dokumentierten Validitäten von amerikanischen Schulnoten dar und ermöglichen einen direkten Vergleich. Besonders im Zusammenhang mit der Neugestaltung der deutschen Hochschulzulassungsregelung sind zuverlässige und aktuelle Informationen über die Validität von möglichen Zulassungskriterien mehr als wünschenswert. Die vorliegende Untersuchung integriert daher die seit 1980 publizierten europäischen Studien zum Zusammenhang des häufigsten Auswahlkriteriums Schulnoten mit dem Studienerfolg (deutsche Studien wurden ab 1985 aufgenommen, um eine Überschneidung mit der Analyse von Baron-Boldt et al., 1988, zu vermeiden).

2 Methode

2.1 Literaturrecherche

Um möglichst viele relevante Studien aus Europa zu identifizieren, wurde zunächst in den Datenbanken PsycInfo und Psynindex recherchiert. Alle herangezogenen Studien wurden auf Hinweise zu weiteren verwendbaren Studien überprüft. Darüber hinaus wurden einschlägig arbeitende Kollegen konsultiert.

Tabelle 1
Anzahl der Validitätskoeffizienten und Stichprobengröße nach Prädiktor

Prädiktor	k	N	Md_N	SD_N	Min_N	Max_N
Durchschnittsnote	53	48178	146	2216.60	27	12496
Studienfachaffine Einzelnote	7	947	143	47.03	69	207
Mathematiknote	13	1766	137	44.99	79	216
Note im Fach Muttersprache	10	1387	136.5	48.50	79	216

Anmerkungen: k = Anzahl der unabhängigen Stichproben = Anzahl der analysierten Validitätskoeffizienten, N = Gesamtzahl der Stichprobengröße, Md_N = Median der Stichprobengröße, SD_N = Standardabweichung der Stichprobengröße, Min_N = Minimum der Stichprobengröße, Max_N = Maximum der Stichprobengröße

26 Studien konnten für das Erfolgskriterium Studiennoten herangezogen werden. Für die Prädiktoren durchschnittliche Schulnote, Einzelnote in einem studienfachnahen Schulfach, Einzelnote im Fach Mathematik und Einzelnote im Schulfach der Landessprache konnten insgesamt 83 Validitätskoeffizienten extrahiert werden. Die genaue Anzahl der unabhängigen Stichproben für jeden Prädiktor sind Tabelle 1 zu entnehmen. Insgesamt wurden Studien aus fünf Ländern (Deutschland, Tschechien, Großbritannien, Norwegen und Österreich) berücksichtigt, die in den Jahren 1980 bis 2005 veröffentlicht wurden. Der Studienerfolg wurde in allen Studien über das Kriterium Studiennoten operationalisiert. Für die Erfolgskriterien Studienabbruch, Studiendauer und Studienzufriedenheit lagen in keinem Fall mehr als vier Studien vor. Daher wurde auf eine Aggregation verzichtet.

2.2 Metaanalytischer Ansatz

Zur statistischen Aggregation der Validitätskoeffizienten wurde die Methode der Validitätsgeneralisierung nach Raju, Burke, Normand und Langlois (1991) eingesetzt und das dazugehörige Computerprogramm (Raju & Fleer, 2003) verwendet. Der Ansatz von Raju et al. geht auf die Methode von Hunter, Schmidt und Jackson (1982) zurück und entwickelt diese zugunsten genauerer Schätzungen der wahren Validität und ihrer Varianz weiter, wie in Monte Carlo-Studien gezeigt werden konnte (Mendoza & Reinhardt, 1991). Dieses Verfahren bietet neben dem verbesserten Schätzalgorithmus die Möglichkeit, die spezifischen Artefaktinformationen wie Reliabilitäten oder Varianzeinschränkungen aus den einzelnen Studien zu berücksichtigen. Für den Fall der fehlenden Information aus den Primärstudien kann ein Schätzwert festgelegt werden.

Das Programm berechnet die Varianz der mittleren gewichteten Validität und die Konfidenzintervalle sowohl im *Fixed Effects*-Modell (FE-Modell) als auch im *Random Effects*-Modell (RE-Modell). Inhaltlich sind die Annahmen des *Random Effects*-Modells angemessen, da nicht von einem einheitlichen Populationsparameter ausgegangen werden kann. In den meisten Fällen gingen allerdings zu wenig Validitätskoeffizienten in die Analyse ein, um eine zuverlässige Schätzung unter Annahme des *Random Effects*-Modells zu erhalten. Schulze (2004) stellt fest, dass verlässliche Ergebnisse im *Random Effects*-Modell erst ab einer Koeffizientenanzahl von 32 zu erwarten sind. Daher werden hier stets die Ergebnisse aus beiden Modellen berichtet.

Eine Fisher Z-Transformation der Validitätskoeffizienten wurde gemäß den Empfehlungen von Hunter und Schmidt (2004, p. 83) nicht durchgeführt.

2.3 Vorgehensweise bei der Berechnung der Metaanalysen

Ab einer Koeffizientenanzahl von fünf wurden für die einzelnen Prädiktoren metaanalytische Berechnungen

durchgeführt. Als Effektmaße dienten bivariate Produkt-Moment-Korrelationen. Validitätsangaben in abweichenden Effektmaßen aus den Primärstudien wurden – sofern möglich – nach den gängigen Formeln in Korrelationskoeffizienten transformiert (Rustenbach, 2003; Bortz & Döring, 1995). In die Analyse integriert wurden Primärstudien, die Korrelationen oder andere Effektmaße, die in Korrelationen umgerechnet werden können, zwischen Schulnoten (entweder Durchschnittsnote oder die Note in einem studienfachnahen Unterrichtsfach, im Fach Mathematik oder in dem Schulfach der jeweiligen Muttersprache) und Studiennoten berichten. Das Erfolgskriterium war in fast allen Fällen eine Durchschnittsnote. Nur in zwei Fällen diente eine einzelne Note als Maß für den Studienerfolg. Für jede Stichprobe ging je Prädiktor nur ein Wert in die Analyse ein. Lagen mehrere Korrelationen vor, so wurden diese unter Berücksichtigung der Interkorrelationen gepoolt oder – wo dies nicht möglich war – unter Verwendung einer Fisher Z-Transformation gemittelt. Fisher Z wurde an dieser Stelle verwendet (z. B. wenn Korrelationen zu verschiedenen Prüfungsleistungen angegeben waren), da diese Vorgehensweise bei kleinen Stichproben Unterschätzungen vermeiden kann (Silver & Dunlap, 1987). Bei den im Folgenden berichteten metaanalytisch aggregierten Validitäten handelt es sich *nicht* um Fisher Z-transformierte Werte, da Hunter und Schmidt (2004) davon abraten und es bei großen Stichproben zu Überschätzungen der mittleren Korrelationen kommen kann.

Je Validitätskoeffizient wurden 79 Einzelinformationen zu den folgenden Aspekten kodiert: Generelle Angaben zur Studie (z. B. Herkunftsland der Studie, Jahr der Publikation), Charakterisierung der Stichprobe (z. B. Größe, Geschlechterverteilung, Ethnie, Durchschnittsalter, Heterogenität der Stichprobe, Selektionsquote und Selektionsmethode), Charakterisierung des Prädiktors (z. B. Art und/oder Schulfach der Note, deskriptive Statistiken, Angaben zur Reliabilität), Charakterisierung des Kriteriums (z. B. Art des Kriteriums, Anzahl der Abstufungen, deskriptive Statistiken, Angaben zur Reliabilität), Informationen zur Validität (z. B. verwendete Statistik, Koeffizienten, methodische Besonderheiten) und zu Moderatorvariablen (z. B. Hochschulniveau, zeitlicher Abstand zwischen Prädiktor und Kriteriumsbestimmung, Selektionskontext oder freiwillige Teilnahme).

Es wurden Korrekturen von Artefaktquellen durchgeführt, die zu einer artifiziellen Veränderung der beobachteten Effektstärkevarianz beitragen. Bei allen Prädiktoren wurde eine Kontrolle des Stichprobenfehlers und eine Gewichtung der Einzelkorrelationen bei der Aggregation mit den Stichprobengrößen vorgenommen. Auch die Reliabilität des Erfolgskriteriums sowie Varianzeinschränkungen (Selektion der Studierenden durch das durchlaufene Auswahlverfahren) wurden berücksichtigt. Auf eine Korrektur der Reliabilität des Prädiktors Schulnoten wurde verzichtet, da diese Ungenauigkeit bei realen Auswahl-situationen ebenfalls hingenommen werden muss. Die be-

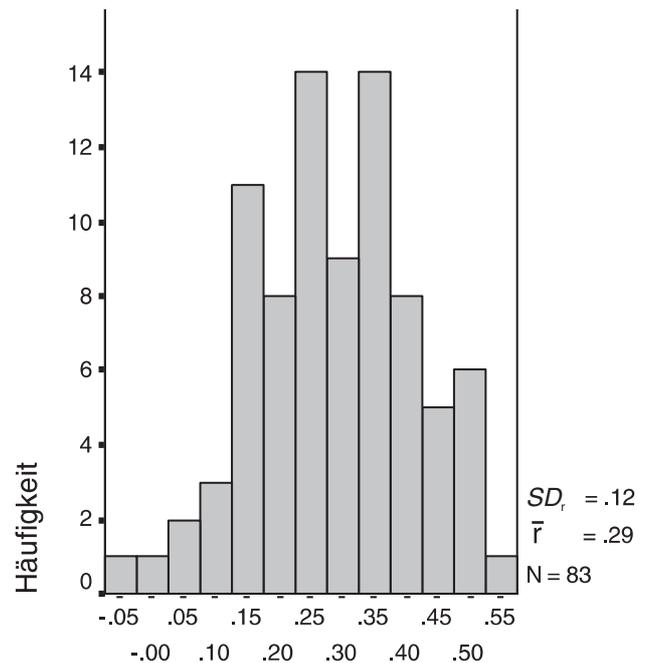
rechnerische Validität soll Auskunft über die Vorhersagegenauigkeit geben, die mit den in einer realen Auswahlssituation vorliegenden Schulnoten zu erwarten ist.

2.4 Artefaktinformationen

Die Studien enthielten kaum 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-Formula angewendet wird. Angenommen, eine Durchschnittsnote mit einer Reliabilität von $r_{tt} = .83$ besteht aus zwei Einzelnoten, so beträgt die Reliabilität der Einzelnoten gemäß Prophecy-Formula $r_{tt} = .71$. Setzt sich die Durchschnittsnote aus drei Einzelnoten zusammen, so ergibt sich eine Schätzung der Reliabilität der Einzelnoten von .62, bei vier Einzelnoten eine Schätzung von .55 und bei fünf Einzelnoten von .49. Empirische Studien, die die Generalisierbarkeitstheorie zur Schätzung der Reliabilität von Hochschulnoten nutzen, kommen auf Werte 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). Die Reliabilität schriftlicher Prüfungen liegt vermutlich etwas über diesem Niveau. Bei der vorliegenden Metaanalyse wurde der aus der Spearman-Brown-Prophecy-Formula extrapolierte Wert von $r_{tt} = .71$ (abgeleitet von der Reliabilität einer Gesamtnote, die sich aus zwei Einzelnoten zusammensetzt) als Reliabilität von Einzelnoten angenommen. Dieser vergleichsweise hohe Wert führt zu einer eher konservativen Korrektur der Validitätskoeffizienten, erscheint für die überwiegend schriftlichen Prüfungen aber angemessen.

Die Validitätskoeffizienten wurden darüber hinaus auch hinsichtlich der Varianzeinschränkung des Prädiktors durch Selektion der Studierenden korrigiert (simultane Berücksichtigung der Artefakte durch einen *compound artifact multiplier*; Hunter & Schmidt, 2004). Wenn in den Primärstudien Informationen zu Selektionsquoten vorlagen oder neben unkorrigierten Werten auch selektionskorrigierte Validitäten angegeben waren, wurden diese Informationen zur Berechnung des Korrekturwerts u verwendet (Rustenbach, 2003). Fehlten in den Primärstudien solche Angaben, wurde der mit der Stichprobengröße gewichtete Mittelwert aus insgesamt 72 u -Koeffizienten eingesetzt. Dieser beträgt .836 und ist vergleichbar mit dem Wert, den Kuncel et al. (2004) in ihrer Metaanalyse für Kohorten zu Beginn des Studiums ermittelten.



unkorrigierte Produkt-Moment-Korrelation

Abbildung 1. Häufigkeiten der in den Primärstudien berichteten unkorrigierten Korrelationen.

3 Ergebnisse

Tabelle 2 zeigt eine Auflistung aller einbezogenen Studien mit Stichprobeneigenschaften, Prädiktor- und Kriterienart sowie der gefundenen Effekte, deren Häufigkeiten in Abbildung 1 grafisch veranschaulicht werden.

Die mittleren gewichteten korrigierten Validitäten von Schulnoten zur Vorhersage des Studienerfolgs liegen alle oberhalb des Werts .20 (Tabelle 3). Keines der berechneten Konfidenzintervalle schließt die Null ein, und auch die unteren Grenzen der 90 Prozent-Kredibilitätsintervalle liegen deutlich über Null. Für alle Schulnoten kann also eine von Null verschiedene positive prognostische Validität angenommen werden. Für die Durchschnittsnoten wurde mit 53 integrierten Korrelationen aus Primärstudien und einer Gesamtstichprobengröße von 48 178 Personen ein Validitätskoeffizient von $\rho = .46$ (bei Korrektur der Kriteriumsreliabilität) bzw. $\rho = .52$ (bei Korrektur der Kriteriumsreliabilität und der Varianzeinschränkung durch Selektion) berechnet. Der ungewichtete unkorrigierte Mittelwert liegt bei .31. Für die fachnahe Einzelnote ergaben die sieben berücksichtigten Koeffizienten eine mittlere Validität von .31 bzw. .36 ($N = 947$) bei einem ungewichteten unkorrigierten Mittel von ebenfalls .31. Die Mathematiknote weist eine mittlere Validität von .27 bzw. .32 auf ($k = 13$, $N = 1766$). Der ungewichtete Mittelwert liegt hier bei .27. Die Einzelnote im Unterrichtsfach der jeweiligen Muttersprache korreliert korrigiert nur noch zu .22 bzw. .26 mit dem Studienerfolg ($k = 10$, $N = 1387$); das arithmetische Mittel der Korrelationen beträgt .19. Bei diesem Prädiktor

Tabelle 2
 Aufstellung der in die Metaanalyse aufgenommenen Studien

Primärstudie	Land	Prädiktor	<i>N</i>	<i>A</i>	<i>F</i>	Studienfach	Studienabschnitt	<i>r</i>
Barkowski (1986)	Deutschland	Durchschnittsnote	27	–	–	Zahnmedizin	vorklinisch	.25
			60	–	–	Zahnmedizin	klinisch	.26
			102	–	–	Zahnmedizin	klinisch	.13
Brandstätter & Fart- hofer (2002)	Österreich	Durchschnittsnote	705	19.0	43.2	Kombination mehrerer Fächer	Grund- und Hauptstudium	.48
Crum & Parikh (1983)	Großbritannien	Durchschnittsnote	146	–	24.1	Sozialwissenschaften	Bachelor	.16
Duff, Boyle, Dun- leavy & Ferguson (2004)	Großbritannien	Durchschnittsnote	146	24.3	74.7	Sozialwissenschaften	Bachelor	.27
Ferguson, James, O'Hehir & Sanders (2003)	Großbritannien	Durchschnittsnote	118	19.7	58.0	Humanmedizin	vorklinisch und klinisch	.13
Giesen, Gold, Hum- mer & Jansen (1986)	Deutschland	Fachaffine Einzelnote Mathematiknote	85	–	–	Medizin	vorklinisch	.26
		Fachaffine Einzelnote Mathematiknote Note Muttersprache	130	–	–	Mathematik und Naturwis- sensschaften	Grundstudium	.34
								.39
		Fachaffine Einzelnote Mathematiknote Note Muttersprache	143	–	–	Mathematik und Naturwis- sensschaften	Grundstudium	.34
								.32
		Fachaffine Einzelnote Mathematiknote Note Muttersprache	146	–	–	Ingenieurwissenschaften	Grundstudium	.36
								.40
		Mathematiknote Note Muttersprache	113	–	–	Wirtschaftswissenschaften	Grundstudium	.37
						.20		
Mathematiknote Note Muttersprache	89	–	–	Rechtswissenschaft	Hauptstudium	.28		
						.21		
Mathematiknote Note Muttersprache	216	–	–	Sprach- und Kulturwissen- schaften	Grund- und Hauptstudium	.16		
						.34		
Mathematiknote	157	–	–	Lehramt	Hauptstudium	.25		
Hell (2003)	Deutschland	Durchschnittsnote	311	21.0	46.8	Wirtschaftswissenschaften	Grundstudium	.54
Höppel & Moser (1993)	Deutschland	Durchschnittsnote	295	–	100	Agrarwissenschaften	Grund- und Hauptstudium	.47
			567	–	0	Agrarwissenschaften	Grund- und Hauptstudium	.43
Hörschgen, Cierpka, Friese & Steinbach (1993)	Deutschland	Durchschnittsnote	371	–	22.5	Wirtschaftswissenschaften	Grund- und Hauptstudium	.40
Höschl & Kozeny (1997)	Tschechien	Mathematiknote Note Muttersprache	92	18.4	43.5	Humanmedizin	vorklinisch und klinisch	.34 .24
Kurz, Fischer & Wagner (1995)		Durchschnittsnote Fachaffine Einzelnote Mathematiknote Note Muttersprache	78	–	0	Maschinenbau	Fachhochschule	.42
			69	–	0	Maschinenbau	Fachhochschule	.50
			79	–	0	Maschinenbau	Fachhochschule	.34
			79	–	0	Maschinenbau	Fachhochschule	.04
McManus & Ri- chards (1986)	Großbritannien	Durchschnittsnote	485	–	–	Humanmedizin	vorklinisch	.23
Meier (2003)	Deutschland	Durchschnittsnote Mathematiknote Note Muttersprache	212	–	44.8	Rechtswissenschaft	Hauptstudium	.22
								.22
								.12

Primärstudie	Land	Prädiktor	<i>N</i>	<i>A</i>	<i>F</i>	Studienfach	Studienabschnitt	<i>r</i>
Melamed (1992)	Großbritannien	Durchschnittsnote	265	–	–	Psychologie	Bachelor	.20
Montague & Odds (1990)	Großbritannien	Fachaffine Einzelnote	207	–	–	Humanmedizin	vorklinisch und klinisch	.16
		Mathematiknote	137	–	–	Humanmedizin	vorklinisch und klinisch	–.03
Ohlsen (1985)	Deutschland	Durchschnittsnote	125	25.8	24.0	Wirtschaftswissenschaften	Hauptstudium	.24
Schiefele, Streblow, Ermgassen & Moschner (2003)	Deutschland	Durchschnittsnote	280	–	63.2	Kombination mehrerer Fächer	Grundstudium	.27
Schmidt-Atzert (2005)	Deutschland	Durchschnittsnote	78	–	–	Psychologie	Grundstudium	.37
Steyer, Yousfi & Würfel (2005)	Deutschland	Durchschnittsnote	167	–	80.0	Psychologie	Hauptstudium	.28
		Fachaffine Einzelnote						.19
		Mathematiknote						.11
		Note Muttersprache						.17
Stumpf & Fay (1991)	Deutschland	Durchschnittsnote	57	–	–	Veterinärmedizin	vorklinisch und klinisch	.41
			46	–	–	Veterinärmedizin	klinisch	.15
			84	–	–	Veterinärmedizin	klinisch	.45
			74	–	–	Veterinärmedizin	klinisch	.29
			94	–	–	Zahnmedizin	klinisch	.16
			49	–	–	Zahnmedizin	klinisch	.33
			78	–	–	Zahnmedizin	vorklinisch	.26
			59	–	–	Zahnmedizin	klinisch	.06
			58	–	–	Zahnmedizin	klinisch	.26
			Stumpf & Nauels (1990)	Deutschland	Durchschnittsnote	1688	–	–
1649	–	–				Medizin	vorklinisch und klinisch	.30
1244	–	–				Medizin	vorklinisch und klinisch	.41
1119	–	–				Medizin	vorklinisch und klinisch	.33
Trost (1998)	Deutschland	Durchschnittsnote	12496	–	–	Humanmedizin	vorklinisch	.45
			7504	–	–	Humanmedizin	vorklinisch	.50
			7876	–	–	Humanmedizin	vorklinisch	.45
			1444	–	–	Veterinärmedizin	vorklinisch	.50
			564	–	–	Veterinärmedizin	vorklinisch	.49
			572	–	–	Veterinärmedizin	vorklinisch	.41
			2807	–	–	Zahnmedizin	vorklinisch	.35
			1277	–	–	Zahnmedizin	vorklinisch	.35
1449	–	–	Zahnmedizin	vorklinisch	.32			
Trost & Freitag (1991)	Deutschland	Durchschnittsnote	164	–	–	Wirtschaftswissenschaften	Grundstudium	.31
			51	–	–	Informatik	Grundstudium	.41
			512	–	–	Ingenieurwissenschaften	Grundstudium	.48
Trost & Piel (1991)	Deutschland	Durchschnittsnote	47	–	–	Wirtschaftswissenschaften	Grundstudium	.28
			37	–	–	Wirtschaftswissenschaften	Grundstudium	.18

Primärstudie	Land	Prädiktor	<i>N</i>	<i>A</i>	<i>F</i>	Studienfach	Studienabschnitt	<i>r</i>
			43	–	–	Wirtschaftswissenschaften	Grundstudium	.23
			45	–	–	Wirtschaftswissenschaften	Grundstudium	.33
Vollmer (1984) ¹	Norwegen	Durchschnittsnote	70	–	100	Kombination mehrerer Fächer	Bachelor	.09
			48	–	0	Kombination mehrerer Fächer	Bachelor	.01
Wallace, Sanderson, Mercer & Gilmore (1988)	Großbritannien	Durchschnittsnote	52	–	–	Humanmedizin	vorklinisch und klinisch	.16
			143	–	–	Humanmedizin	vorklinisch und klinisch	.17
			140	–	–	Humanmedizin	vorklinisch und klinisch	.21

Anmerkungen: Bei dem Prädiktor handelt es sich immer um Noten des zum Hochschulzugang berechtigenden Zeugnisses, bei dem Kriterium immer um durchschnittliche Studiennoten; ¹ bei dieser Studie war das Kriterium eine Einzelnote. *N* = Stichprobengröße, *A* = Durchschnittsalter der Stichprobe, *F* = Anteil an weiblichen Probanden in Prozent, *r* = unkorrigierte Produkt-Moment-Korrelation.

Tabelle 3
Ergebnisse der Metaanalyse differenziert nach Art der Schulnoten

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	<i>SE</i> _{FE}	95 % <i>CI</i> _{FE}	<i>SE</i> _{RE}	95 % <i>CI</i> _{RE}	90 % <i>CV</i>
Durchschnittsnote	48178	53	.305								
Korrektur Reliabilität Kriterium				.459	.007	13.3	.004	[.451, .468]	.012	[.435, .483]	.353
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.517	.009	12.4	.005	[.507, .527]	.014	[.489, .545]	.393
Fachaffine Einzelnote	947	7	.307								
Korrektur Reliabilität Kriterium				.309	.005	57.7	.032	[.247, .372]	.042	[.227, .391]	.217
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.360	.006	58.0	.035	[.290, .429]	.047	[.269, .451]	.258
Einzelnote Mathematik	1766	13	.266								
Korrektur Reliabilität Kriterium				.274	.010	41.7	.024	[.227, .321]	.037	[.202, .346]	.144
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.320	.014	40.2	.027	[.267, .372]	.042	[.236, .403]	.168
Einzelnote Muttersprache	1387	10	.193								
Korrektur Reliabilität Kriterium				.220	.000	100.0	.028	[.165, .276]	.028	[.165, .275]	.220
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.260	.000	100.0	.033	[.196, .324]	.032	[.196, .323]	.260

Anmerkungen:

N = Größe der Gesamtstichprobe

k = Anzahl der unabhängigen Stichproben

\bar{r} = mittlere ungewichtete unkorrigierte Validität

ρ = mittlere gewichtete korrigierte Validität

σ_p^2 = geschätzte Varianz von ρ

90 % *CV* = credibility value: untere Grenze des 90 %-Kreditibilitätsintervalls

% = prozentualer Anteil der Gesamtvarianz, der durch die Artefakte erklärt wird

*SE*_{FE} = Standardfehler von ρ , Modell: Feste Effekte

95 % *CI*_{FE} = Konfidenzintervall mit $p = 95\%$ um ρ , Modell: Feste Effekte

*SE*_{RE} = Standardfehler von ρ , Modell: Zufällige Effekte

95 % *CI*_{RE} = Konfidenzintervall mit $p = 95\%$ um ρ , Modell: Zufällige Effekte

Tabelle 4
Homogenitätsprüfung zur Moderatoranalyse

Prädiktor	N	k	Fixed Effects-Model			Random Effects-Model		
			Q	df	p	Q	df	p
Durchschnittsnoten								
Korrektur Reliabilität Kriterium	48178	53	290.77	52	.000	72.01	52	0.035
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			335.60	52	.000	63.56	52	0.131
Studienfachaffine Einzelnote								
Korrektur Reliabilität Kriterium	947	7	12.82	6	.046	7.61	6	0.268
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			13.43	6	.037	7.68	6	0.262
Mathematiknote								
Korrektur Reliabilität Kriterium	1766	13	29.39	12	.003	12.42	12	0.413
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			26.60	12	.003	12.09	12	0.439
Muttersprache								
Korrektur Reliabilität Kriterium	1387	10	10.64	9	.301	10.64	9	0.301
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung			10.74	9	.294	10.74	9	0.294

Anmerkungen: N = Gesamtzahl der Stichprobengröße, k = Anzahl der unabhängigen Stichproben = Anzahl der analysierten Validitätskoeffizienten, Q = Homogenitätsstatistik, df = Freiheitsgrade, p = Signifikanzniveau des χ^2 -Tests

wird die Varianz zu 100 Prozent durch die Artefakte erklärt. Bei den anderen Prädiktoren sind Moderatoreffekte wahrscheinlich, da der prozentuale Anteil der Gesamtvarianz, der durch die Artefakte erklärt wird, weniger als 75 Prozent beträgt (Hunter & Schmidt, 2004).

3.1 Moderatoren der Validität

Neben der Überprüfung der 75 Prozent-Regel nach Hunter und Schmidt (2004) wurde für alle Prädiktoren ein Homogenitätstest (Q-Statistik) berechnet. Diese beiden Kennwerte geben Hinweise auf das Vorhandensein von Moderatoreffekten. Folgende mögliche Moderatoren wurden bei ausreichender Informationsmenge von mindestens drei Validitätskoeffizienten pro Moderatorstufe in Moderatoranalysen überprüft: Publikationsdatum, Abstand des Erhebungszeitpunkts des Prädiktors zu dem des Kriteriums, mittleres Alter der Versuchspersonen, Heterogenität der Stichprobe in Bezug auf unterschiedliche Hochschulen, verschiedene Jahrgänge und verschiedene Fächer, Anzahl der Abstufungen im Kriterium, Studienabschnitt (Grundstudium/Bachelor vs. Hauptstudium/Master bzw. vorklinischer vs. klinischer Studienabschnitt bei medizinischen Studiengängen), Vorliegen einer echten Auswahl-situation als Kontext der Datenerhebung in der Primärstudie, Art der Publikation (Artikel in einem *peer-reviewed* Journal vs. Publikationen ohne *Peer Review*-Verfahren), Herkunftsland der Untersuchung sowie das Studienfach. Bei den medizinischen Disziplinen wurde bei der Analyse des Moderators Studienabschnitt nach vorklinischem und klinischem

Abschnitt differenziert, da diese Aufteilung international gebräuchlich ist. Hier erwarten wir für den ersten, vorklinischen Studienabschnitt nicht nur aufgrund zeitlicher, sondern auch aufgrund inhaltlicher Nähe von Prädiktor und Kriterium höhere Validitäten der Schulnoten als für den zweiten, klinischen Abschnitt. Q-Statistiken (im FE-Modell und teilweise im RE-Modell) ergaben in Übereinstimmung mit der 75 Prozent-Regel Hinweise auf Moderator-einflüsse bei drei Prädiktoren (Tabelle 4).

Gemäß den Ergebnissen des Q-Tests und der Überprüfung der 75 Prozent-Regel wurden für alle Prädiktoren mit Ausnahme der Note im Fach Muttersprache Moderatoranalysen berechnet, sofern ausreichend Informationen aus den Primärstudien vorlagen.

Zur statistischen Prüfung der Moderatoreinflüsse wurde bei kontinuierlichen oder dummy-kodierten Variablen auf Regressionen und bei kategorialen Variablen auf Varianzanalysen zurückgegriffen. Ab einer Mindestanzahl von 32 Validitätskoeffizienten (k) werden die Ergebnisse unter Annahme des *Random Effects*-Modells berichtet, ansonsten wurde mit den Varianzen nach dem *Fixed Effects*-Modell gerechnet. Aufgrund der großen Anzahl von überprüften Moderatorvariablen wurde das α -Fehler-Niveau nach einer Bonferroni-Adjustierung auf $p = .004$ (overall: $p = .05$) festgelegt. Die Statistiken zu den Regressionsanalysen beziehen sich im Folgenden aus Gründen der Übersichtlichkeit auf die nicht selektionskorrigierten Validitäten.

Für den *Prädiktor Durchschnittsnote* konnten signifikante Effekte für folgende Moderatoren nachgewiesen werden.

Moderator Herkunftsland der Studie: Dieser wichtige

Tabelle 5
Ergebnisse der Moderatoranalyse, differenziert nach Herkunftsland der Studie für den Prädiktor Durchschnittsnote

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Deutschland	45860	42	.335								
Korrektur Reliabilität Kriterium				.467	.005	14.1	.004	[.458, .476]	.012	[.444, .490]	.376
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.525	.008	11.8	.005	[.515, .534]	.014	[.497, .553]	.413
Großbritannien	1495	8	.191								
Korrektur Reliabilität Kriterium				.223	.000	100.0	.027	[.169, .277]	.015	[.194, .252]	.223
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.276	.000	100.0	.033	[.212, .340]	.019	[.239, .313]	.276
Norwegen	118	2	.054								
Korrektur Reliabilität Kriterium				.072	.000	100.0	.110	[-.143, .287]	.033	[.008, .136]	.072
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.086	.000	100.0	.131	[-.170, .342]	.039	[.010, .162]	.086

Anmerkungen: siehe Tabelle 3

Tabelle 6
Ergebnisse der Metaanalyse differenziert nach Art der Schulnoten für deutsche Studien

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Durchschnittsnote	45860	42	.335								
Korrektur Reliabilität Kriterium				.467	.005	14.1	.004	[.458, .476]	.012	[.444, .490]	.376
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.525	.008	11.8	.005	[.515, .534]	.014	[.497, .553]	.413
Fachaffine Einzelnote	740	6	.332								
Korrektur Reliabilität Kriterium				.347	.002	80.8	.035	[.277, .416]	.039	[.270, .424]	.293
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.402	.002	81.4	.039	[.326, .479]	.043	[.317, .487]	.344
Einzelnote Mathematik	1537	11	.286								
Korrektur Reliabilität Kriterium				.295	.004	65.2	.025	[.246, .345]	.031	[.234, .357]	.217
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.345	.005	64.2	.028	[.289, .401]	.036	[.275, .414]	.254
Einzelnote Muttersprache	1295	9	.188								
Korrektur Reliabilität Kriterium				.217	.001	92.6	.029	[.160, .275]	.030	[.158, .277]	.186
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.256	.001	93.9	.034	[.189, .323]	.035	[.187, .325]	.223

Anmerkungen: siehe Tabelle 3

Moderator wurde untersucht, obwohl nicht für alle Moderatorstufen (Länder) mehr als drei Koeffizienten vorhanden waren. Die Subgruppenanalyse (Tabelle 5) zeigt die Ergebnisse für die Länder, aus denen mindestens zwei Koeffizienten vorlagen. Die Differenzen sind signifikant ($Q = 25.036$, $df = 2$, $p < .001$, $k = 52$). Da die Länder so unterschiedlich häufig vertreten sind und bedeutsame Unterschiede angenommen werden müssen, werden im Folgen-

den nur noch Moderatoranalysen für *deutsche* Studien berichtet. Für die Befunde aus Großbritannien und Norwegen besteht statistisch keine Veranlassung, nach weiteren Moderatorvariablen zu suchen, da die Varianzaufklärung durch Artefakte 100 Prozent erreicht. Tabelle 6 zeigt die Ergebnisse der Metaanalyse für alle vier Prädiktoren nur für die deutschen Studien.

Moderatorvariable Größe des zeitlichen Abstands zwi-

Tabelle 7

Ergebnisse der Moderatoranalyse, differenziert nach Studienabschnitt für den Prädiktor Durchschnittsnote für deutsche Studien

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Grundstudium/Bachelor	1568	10	.340								
Korrektur Reliabilität Kriterium				.444	.009	39.2	.024	[.397, .492]	.038	[.369, .520]	.323
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.506	.010	42.0	.027	[.454, .558]	.041	[.426, .587]	.380
Hauptstudium/Master	504	3	.247								
Korrektur Reliabilität Kriterium				.269	.000	100.0	.046	[.180, .359]	.016	[.237, .301]	.269
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.317	.000	100.0	.052	[.215, .419]	.019	[.280, .354]	.317
Medizinische Disziplinen: vorklinischer Abschnitt	36094	11	.394								
Korrektur Reliabilität Kriterium				.489	.003	8.1	.005	[.479, .498]	.017	[.456, .522]	.421
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.548	.005	5.0	.005	[.538, .558]	.022	[.504, .592]	.455
Medizinische Disziplinen: klinischer Abschnitt	626	9	.232								
Korrektur Reliabilität Kriterium				.254	.001	93.8	.041	[.173, .334]	.042	[.171, .337]	.213
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.291	.000	100.0	.048	[.197, .385]	.043	[.207, .376]	.291

Anmerkungen: siehe Tabelle 3

Tabelle 8

Ergebnisse der Moderatoranalyse, differenziert nach Art der Studienfächer für den Prädiktor Durchschnittsnote für deutsche Studien

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Sprach- und Kulturwissenschaften	245	2	.325								
Korrektur Reliabilität Kriterium				.339	.000	100.0	.062	[.217, .461]	.033	[.275, .403]	.339
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.395	.000	100.0	.069	[.261, .530]	.036	[.325, .465]	.395
Mathematik, Natur- und Ingenieurwissenschaften	641	3	.437								
Korrektur Reliabilität Kriterium				.513	.000	100.0	.034	[.445, .580]	.016	[.481, .545]	.513
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.581	.000	100.0	.036	[.511, .651]	.017	[.548, .614]	.581
Medizinische Disziplinen	42477	25	.328								
Korrektur Reliabilität Kriterium				.470	.005	10.2	.005	[.461, .479]	.015	[.441, .499]	.382
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.528	.007	8.1	.005	[.517, .538]	.018	[.492, .563]	.417
Rechts-, Wirtschafts- und Sozialwissenschaften	1355	9	.304								
Korrektur Reliabilität Kriterium				.395	.011	37.9	.027	[.342, .447]	.043	[.310, .480]	.263
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.453	.011	41.5	.030	[.394, .511]	.046	[.362, .544]	.316

Anmerkungen: siehe Tabelle 3

Tabelle 9

Ergebnisse der Moderatoranalyse, differenziert nach Art der Studienfächer für den Prädiktor Durchschnittsnote für deutsche Studien, Detailanalyse

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Wirtschaftswissenschaften	1143	8	.314								
Korrektur Reliabilität Kriterium				.423	.008	44.6	.029	[.367, .480]	.043	[.338, .508]	.307
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.484	.009	49.2	.032	[.421, .547]	.046	[.394, .574]	.365
Humanmedizin	33576	7	.397								
Korrektur Reliabilität Kriterium				.486	.003	5.1	.005	[.476, .496]	.022	[.443, .529]	.414
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.551	.004	4.3	.005	[.541, .561]	.025	[.501, .600]	.467
Zahnmedizin	6060	11	.248								
Korrektur Reliabilität Kriterium				.362	.001	68.0	.013	[.336, .388]	.016	[.331, .393]	.324
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.375	.000	100.0	.015	[.346, .404]	.014	[.347, .404]	.375
Veterinärmedizin	2841	7	.386								
Korrektur Reliabilität Kriterium				.511	.003	44.3	.017	[.477, .544]	.026	[.461, .561]	.446
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.576	.003	45.8	.019	[.539, .613]	.028	[.521, .630]	.506

Anmerkungen siehe Tabelle 3

Tabelle 11

Ergebnisse der Moderatoranalyse, differenziert nach Art der Studienfächer für den Prädiktor Mathematiknote

	<i>N</i>	<i>k</i>	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Sprach- und Kulturwissenschaften*	540	3	.173								
Korrektur Reliabilität Kriterium				.187	.000	100.0	.046	[.098, .277]	.035	[.119, .256]	.187
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.222	.000	100.0	.053	[.118, .326]	.041	[.143, .301]	.222
Mathematik, Natur- und Ingenieurwissenschaften*	498	4	.363								
Korrektur Reliabilität Kriterium				.401	.000	100.0	.042	[.318, .483]	.019	[.363, .438]	.401
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.463	.000	100.0	.046	[.373, .553]	.021	[.423, .503]	.463
Medizinische Disziplinen	314	3	.207								
Korrektur Reliabilität Kriterium				.187	.028	26.4	.057	[.074, .300]	.112	[-.032, .407]	-.026
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.217	.038	25.2	.066	[.089, .346]	.131	[-.039, .473]	-.033
Rechts-, Wirtschafts- und Sozialwissenschaften*	414	3	.290								
Korrektur Reliabilität Kriterium				.301	.000	100.0	.049	[.205, .397]	.040	[.222, .379]	.301
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.352	.000	100.0	.055	[.244, .459]	.045	[.264, .440]	.352

Anmerkungen: siehe Tabelle 3. * hier lagen nur deutsche Studien vor

Tabelle 10

Ergebnisse der Moderatoranalyse, differenziert nach Art der Studienfächer und Studienabschnitt für den Prädiktor Durchschnittsnote für deutsche Studien

	N	k	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Wirtschaftswissenschaften, Grundstudium/Bachelor	483	5	.313								
Korrektur Reliabilität Kriterium				.483	.013	43.1	.044	[.396, .570]	.068	[.351, .616]	.337
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.545	.014	47.6	.050	[.447, .642]	.072	[.403, .686]	.395
Humanmedizin, vorklinisch	27876	3	.467								
Korrektur Reliabilität Kriterium				.509	.001	12.6	.005	[.499, .518]	.014	[.481, .536]	.480
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.577	.001	12.8	.005	[.567, .587]	.014	[.549, .605]	.548
Zahnmedizin, vorklinisch	5638	5	.306								
Korrektur Reliabilität Kriterium				.374	.000	100.0	.013	[.348, .400]	.009	[.357, .391]	.374
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.385	.000	100.0	.014	[.358, .413]	.007	[.372, .399]	.385
Zahnmedizin, klinisch	422	6	.200								
Korrektur Reliabilität Kriterium				.205	.000	100.0	.051	[.105, .305]	.038	[.131, .279]	.205
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.241	.000	100.0	.059	[.125, .357]	.042	[.160, .322]	.241
Veterinärmedizin, vorklinisch	2580	3	.467								
Korrektur Reliabilität Kriterium				.525	.001	50.5	.016	[.492, .557]	.023	[.479, .570]	.489
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.593	.001	50.0	.017	[.560, .626]	.024	[.546, .639]	.556
Veterinärmedizin, klinisch	204	3	.297								
Korrektur Reliabilität Kriterium				.356	.002	85.9	.069	[.221, .491]	.074	[.210, .502]	.294
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.395	.000	100.0	.082	[.235, .555]	.068	[.262, .528]	.395

Anmerkungen: siehe Tabelle 3

Tabelle 12

Ergebnisse der Moderatoranalyse differenziert nach Studienabschnitt für den Prädiktor Mathematiknote

	N	k	\bar{r}	ρ	σ_p^2	%	SE_{FE}	95 % CI_{FE}	SE_{RE}	95 % CI_{RE}	90 % CV
Grundstudium/Bachelor*	532	4	.370								
Korrektur Reliabilität Kriterium				.406	.000	100.0	.041	[.326, .485]	.017	[.371, .440]	.406
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.469	.000	100.0	.044	[.383, .554]	.019	[.431, .506]	.469
Hauptstudium/Master*	625	4	.215								
Korrektur Reliabilität Kriterium				.227	.000	100.0	.042	[.145, .308]	.034	[.161, .293]	.227
Korrektur Reliabilität Kriterium & Variabilitätseinschränkung				.228	.000	100.0	.048	[.174, .362]	.039	[.191, .345]	.268

Anmerkungen: siehe Tabelle 3. * hier lagen nur deutsche Studien vor

schen Prädiktor und Kriterium: Sinkende Validitäten lassen sich bei größerem Abstand zwischen den Erhebungszeitpunkten von Prädiktor und Kriterium feststellen ($R^2 = .51$, $\beta = -.71$, $p < .004$, $k = 17$).

Moderator Studienabschnitt: Die Validität von Durchschnittsnoten ist für die Leistungen im Hauptstudium/Master bzw. klinischen Studienabschnitt der medizinischen Disziplinen kleiner als für die Leistungen im Grundstudium/Bachelor bzw. vorklinischen Studienabschnitt ($R^2 = .38$, $\beta = -.62$, $p < .004$, $k = 13$ bzw. $R^2 = .21$, $\beta = -.46$, $p < .004$, $k = 20$). Tabelle 7 zeigt die Subgruppenanalyse für die Moderatorvariable Studienabschnitt.

Moderator Studienfach: Es wurden nach einer Aufstellung des statistischen Bundesamts vier Kategorien gebildet: Sprach- und Kulturwissenschaften inklusive Lehramtsstudiengänge, Mathematik, Natur- und Ingenieurwissenschaften, medizinische Disziplinen sowie Rechts-, Wirtschafts- und Sozialwissenschaften. Die höchsten Validitäten ließen sich feststellen für die Natur- und Ingenieurwissenschaften mit einer mittleren gewichteten Validität von $\rho = .51$ resp. $\rho = .58$ nach Selektionskorrektur ($k = 3$, $N = 641$), gefolgt von den medizinischen Disziplinen mit einer mittleren gewichteten Validität von $.47$ resp. $.53$ ($k = 25$, $N = 42477$). Studienleistungen in den Fachbereichen Rechts-, Wirtschafts- und Sozialwissenschaften kovariieren in geringerem Ausmaß mit den schulischen Durchschnittsnoten; hier konnte eine mittlere gewichtete Validität von $\rho = .40$ resp. $\rho = .45$ ($k = 9$, $N = 1355$) festgestellt werden. Niedrigste Validitäten finden sich für das Studienfach Psychologie als einzigem Vertreter der Kategorie Sprach- und Kulturwissenschaften. Tabelle 8 zeigt die Ergebnisse der Subgruppenanalyse Studienfach.

Während für die Studienfächer Sprach- und Kulturwissenschaften sowie Mathematik, Natur- und Ingenieurwissenschaften je 100 Prozent der Varianz durch die Artefakte aufgeklärt wird und demnach keine Moderatoren zu erwarten sind, veranlasst die niedrige Aufklärung in den beiden anderen Studienfachgruppen von 8 bis 42 Prozent (Tabelle 8) zu einer differenzierteren Betrachtung. Eine Aufteilung speziell der Gruppe der Rechts-, Wirtschafts- und Sozialwissenschaften empfiehlt sich zudem, da in der Metaanalyse von Baron-Boldt et al. (1988) in Subgruppenanalysen für die Wirtschaftswissenschaften die höchste und für die Rechtswissenschaften die niedrigste Validität gefunden wurde.

Für die Rechtswissenschaften wurde nur *eine* mit einem Koeffizienten von $.22$ Untersuchung gefunden, der recht niedrig ausfällt. Demgegenüber beträgt die mittlere gewichtete reliabilitätskorrigierte Validität der Schulnoten für wirtschaftswissenschaftliche Studiengänge $.42$, reliabilitäts- und selektionskorrigiert $.48$. Die Varianzaufklärung durch die Artefakte steigt allerdings nur geringfügig an und bleibt unter 50 Prozent. Bei der getrennten Analyse der medizinischen Disziplinen Human-, Zahn- und Veterinärmedizin ergaben sich die höchsten Validitäten für den Studiengang Veterinärmedizin mit $\rho = .51$ (reliabilitätskorrigiert) und $\rho = .58$ (reliabilitäts- und selektionskorrigiert). Die Va-

lidität der Schulnoten für die humanmedizinischen Studiengänge fällt mit $\rho = .48$ bzw. $\rho = .55$ nur geringfügig niedriger aus. Eine deutliche Steigerung über 75 Prozent der Varianzaufklärung durch die Artefakte ließ sich bei dieser Analyse aber nur für die Zahnmedizin feststellen. Tabelle 9 zeigt Detailinformationen.

Die vorangegangenen Analysen belegen, dass durchschnittliche Schulnoten in Deutschland zumeist über beachtliche prädiktive Validität für Studiennoten verfügen. Moderiert wird dieser Zusammenhang vor allem durch den Studienabschnitt und das Studienfach. Die Subgruppenanalysen konnten nicht in allen Fällen zu einer Varianzaufklärung durch die Artefakte von über 75 Prozent führen. Daher wurden – wo es auf Grund der vorliegenden Koeffizientenzahlen möglich war – beide Moderatoren simultan analysiert. In Tabelle 10 sind die Ergebnisse der Subgruppenanalyse dargestellt. Mit Ausnahme der Wirtschaftswissenschaften war diese differenzierte Betrachtung nur noch für die drei medizinischen Disziplinen möglich. Zumindest für die Zahn- und Veterinärmedizin wurde der Prozentsatz aufgeklärter Varianz durch die Auftrennung nach klinischem und vorklinischem Studienabschnitt bedeutsam gesteigert. In beiden Fällen ist die Validität für Studiennoten des vorklinischen Studienabschnitts erwartungsgemäß höher als für den klinischen Abschnitt. Die Konfidenzintervalle berühren sich zwar in einigen Fällen, überschneiden sich aber nicht.

Ein ähnliches Bild wie für die Durchschnittsnoten zeichnet sich für die Validität der *Mathematiknote* ab (Tabelle 11): Die höchsten Validitäten werden für die Studiengänge Mathematik, Natur- und Ingenieurwissenschaften berichtet, die niedrigsten für Sprach- und Kulturwissenschaften sowie für die medizinischen Disziplinen. Bei letztgenannten schließen das Konfidenzintervall im RE-Modell und das Kreditätsintervall den Wert Null ein. Die niedrigen Anteile der durch Artefakte aufgeklärten Varianz weisen auf die Möglichkeit weiterer bedeutsamer Einflussfaktoren hin und zeigen eine mangelnde Generalisierbarkeit dieses Ergebnisses an. Hierbei kann es sich um einen Effekt des Herkunftslands der Studie handeln, da nur für den Bereich der medizinischen Disziplinen Validitäten der Mathematiknote verschiedener Länder aufgenommen werden konnten. Für deutsche Mathematiknoten liegt nur ein einziger Validitätskoeffizient vor ($r = .31$; Giesen, Gold, Hummer & Jansen, 1986). Aufgrund der Datenlage war auch eine Aufspaltung in die einzelnen medizinischen Disziplinen als potenzieller weiterer Moderator nicht möglich. Darüber hinaus liegen für alle Subgruppen nur äußerst wenige Validitätskoeffizienten vor.

Neben dem Moderator Studienfach erwiesen sich die Moderatoren Abstand des Erhebungszeitpunkts von Prädiktor und Kriterium ($R^2 = .76$, $\beta = -.87$, $p < .004$, $k = 10$) sowie Grund- vs. Hauptstudium ($R^2 = .74$, $\beta = -.86$, $p < .004$, $k = 8$) als statistisch signifikant für die Beziehung zwischen Schulnoten im Fach Mathematik und Studiennoten. Wie auch schon bei den Durchschnittsnoten ergeben sich höhere Validitäten für den Studienerfolg im Grundstu-

dium als für die Noten im Hauptstudium. Die Ergebnisse der Subgruppenanalyse nach dem Studienabschnitt zeigt Tabelle 12. Für diese Analyse konnten nur Studien aus Deutschland herangezogen werden.

Für den Prädiktor studienfachaffine Einzelnote wurden keine signifikanten Moderatorwirkungen gefunden. Allerdings konnten nur die Moderatoren Abstand zwischen Prädiktor- und Kriteriumserhebung und Publikationsart überprüft werden, da für die anderen Moderatorvariablen nicht ausreichend viele Validitätskoeffizienten pro Moderatorstufe vorlagen. Für deutsche Studien werden Moderatorwirkungen ausgeschlossen, da die Varianzaufklärung durch die Artefakte mehr als 75 Prozent beträgt (vgl. Tabelle 7) und der Q -Test nicht signifikant wird ($Q = 7.9402, p > .05, k = 5, \text{Fixed Effects-Modell}$).

4 Diskussion

Die prädiktive Validität von durchschnittlichen Schulabschlussnoten liegt mit $\rho = .46$ (korrigierter Wert für die Reliabilität des Kriteriums) bzw. $\rho = .52$ (reliabilitäts- und selektionskorrigierter Wert) an der Spitze der Einzelprädiktoren zur Vorhersage des Studienerfolgs. Für deutsche Schulabschlussnoten wurden sogar noch etwas höhere Validitäten von $\rho = .47$ bzw. $\rho = .53$ errechnet. Ähnlich hohe Werte erreichen nur noch Studierfähigkeitstests (Hell, Trapmann & Schuler, 2007). Für die anderen europäischen Länder ergaben sich deutlich niedrigere Werte, die allerdings auf wesentlich weniger Primärstudien fußen. Einzelne Fachnoten erreichen immerhin noch mittlere korrigierte Validitäten zwischen .22 und .40. Zur Vorhersage des Studienerfolgs in natur- und ingenieurwissenschaftlichen Studiengängen bzw. der Grundstudiums- oder Bachelornote werden sogar korrigierte Validitäten bis .46 bzw. .47 erreicht. Die höhere Validität von Durchschnittsnoten im Vergleich zu Einzelnoten überrascht nicht. Dieser Befund zeigte sich bereits in der Metaanalyse von Baron-Boldt et al. (1988). Die Autoren führten damals an, dass «die Durchschnittsnote [...] aufgrund der Aggregation die messtheoretischen Anforderungen an einen Prädiktor besser als die einzelnen Fachnoten» (S. 86) erfüllt. Tatsächlich handelt es sich bei den deutschen Abiturnoten um Urteile, die über zwei Jahre hinweg von mehreren Beurteilern in Prüfungen verschiedener Modalitäten (schriftliche Klausuren und Tests, mündliche Mitarbeit, mündliche Abiturprüfung) gebildet werden. Darüber hinaus spiegeln die Schulabschlussnoten Fähigkeiten und Fertigkeiten wider, die sowohl für gute Schulnoten als auch für gute Studiennoten relevant sind, wie z. B. kognitive Leistungsfähigkeit, Lernbereitschaft, Leistungsmotivation, Fleiß und sprachliche Ausdrucksfähigkeit.

Bei genauerer Betrachtung der Ergebnisse sowohl zur Durchschnittsnote als auch zur studienfachnahen Einzelnote und der Mathematiknote fällt die geringe Varianzaufklärung durch die Artefakte von 12 bis 58 Prozent ins Auge.

Die positiven und von Null verschiedenen unteren Werte der Kredititäts- und Konfidenzintervalle zeigen zwar an, dass eine positive Validität generalisierbar ist (Schmidt, Hunter, Pearlman & Hirsh, 1985), angesichts der geringen Varianzaufklärung und der signifikanten Q -Tests sollte jedoch nicht von «der» Validität gesprochen werden. Vielmehr werden mehrere Moderatorwirkungen signifikant. So unterscheiden sich die Validitäten in Abhängigkeit vom Herkunftsland der Studie, von der Größe des zeitlichen Abstands zwischen Prädiktor- und Kriteriumserhebung, vom Studienabschnitt und vom untersuchten Studienfach. Jeweils einzeln betrachtet kann auch für die Moderatorstufen nicht immer eine Varianzaufklärung von 100 Prozent oder über 75 Prozent erreicht werden, so dass hier von weiteren Moderatorvariablen oder einer gemeinsamen Wirkung mehrerer Variablen ausgegangen werden kann. Diese Hypothese konnte in der vorliegenden Untersuchung aufgrund zu geringer Zellbesetzungen nur in einigen Bereichen überprüft werden. Die deutlichen Unterschiede zwischen den betrachteten europäischen Ländern legten für weitere Analysen allerdings die exklusive Betrachtung deutscher Ergebnisse nahe. Die weiteren Moderatoranalysen fanden daher bereits auf einer zweiten Ebene statt. Besonders auffällig ist die geringe Varianzaufklärung durch die Artefakte bei den medizinischen Disziplinen und hier besonders im vorklinischen Studienabschnitt der Humanmedizin. Eine Klärung dieser Befundlage ist allerdings bei vorliegender Studienanzahl nicht möglich. Auch die Ergebnisse der Teilanalysen mit nur sehr geringer Studienzahl (z. B. die Subgruppenanalysen zur Mathematiknote) sind trotz vollständiger Varianzaufklärung durch die Artefakte nur eingeschränkt generalisierbar, da anzunehmen ist, dass die eingegangenen Effekte keine repräsentative Abbildung aller möglichen Effekte dieser Prädiktor-Kriteriums-Verbindung darstellen.

Insgesamt betrachtet bieten die mit Ausnahme der Mathematiknote als Prädiktor für die Studienleistungen in den medizinischen Disziplinen durchgängig positiven unteren Werte der Konfidenz- und Kredititätsintervalle eine Absicherung der Validität von deutschen Schulnoten gegen Null.

Die beste Vorhersagemöglichkeit ergibt sich sowohl mit der Durchschnittsnote als auch mit der Mathematiknote für die Studienfächer Mathematik, Natur- und Ingenieurwissenschaften. Diese Studiengänge scheinen in ihren Anforderungen vergleichbarer mit der weiterführenden Schule zu sein als Studiengänge der Sprach- und Kulturwissenschaften, für die sich weitaus niedrigere Validitätskoeffizienten ergaben. Neben der inhaltlichen Ähnlichkeit ist es auch denkbar, dass die Kriterien in manchen Studiengängen messgenauer sind als in anderen.

Über die Zeit scheinen deutsche Schulnoten insgesamt leicht an Validität gewonnen zu haben, wie der Vergleich mit der Metaanalysen von Baron-Boldt et al. (1988) zeigt. Damals wurde eine mittlere reliabilitäts- und selektionskorrigierte Validität von .46 berichtet gegenüber einem Rho von .53 in der aktuellen Analyse. Diese Tendenz fin-

det sich im Unterschied zur genannten älteren Untersuchung (Baron-Boldt et al., 1988; vgl. auch Schuler, 2006) auch *innerhalb* der vorliegenden Metaanalyse. Jüngere deutsche Studien berichten im Mittel höhere Validitäten. Die Moderatoranalyse (*Random Effects*-Modell) zum Publikationsdatum verfehlte die adjustierte Signifikanzgrenze allerdings knapp ($R^2 = .16$, $\beta = .40$, $p = .006$, $k = 42$). Es ist denkbar, dass es sich hier um einen Effekt handelt, der durch eine verbesserte Messung der Studiennoten, etwa durch veränderte Prüfungsmodalitäten, zustande kommt. Aber auch auf Seiten des Prädiktors hat es Veränderungen, z. B. im schulischen Unterricht und der Notengebung, gegeben, so dass auch dort mögliche Ursachen des Validitätsanstiegs liegen könnten.

Die Moderatoren Studienabschnitt und zeitlicher Abstand zwischen Prädiktor- und Kriteriumserhebung hängen so hoch zusammen ($r = .914$), dass nicht davon ausgegangen werden kann, dass diese Variablen wirklich unterschiedliches beinhalten. Ob in erster Linie die unterschiedlichen Studienabschnitte als Moderator wirken oder ob es sich um einen primär zeitlichen Einfluss handelt, ist unklar und analytisch anhand der gefundenen Ergebnisse nicht zu trennen. Es ist möglich, dass die Abschlussnote des Hauptstudiums weniger stark mit Schulnoten zusammenhängt, weil ein nicht unerheblicher Anteil durch die meist unrelieblere Note einer Abschlussarbeit (z. B. Diplomarbeit) zustande kommt. Auch die teilweise vorliegende Relevanz praktischer Leistungen im Hauptstudium gegenüber der theoretischen Leistungen im Vordiplom können eine inhaltliche Ursache des geringeren Zusammenhangs sein. Auf der anderen Seite können Reifeprozesse durch den größeren Zeitabstand bedeutsamer werden. Neben der Möglichkeit eines realiter geringer werdenden Zusammenhangs zwischen Schul- und Studiennoten im Zeitverlauf könnte dieser Befund auch ein statistisches Artefakt sein, das durch Studienabbruch und die dadurch bedingte Varianzeinschränkung zustande kommt. Leider fehlten in den hier berücksichtigten Primärstudien überwiegend die Angaben zu Studienabbruchquoten.

5 Ausblick

In der vorliegenden Metaanalyse haben sich die Schulnoten einmal mehr als valider Prädiktor für Studiennoten erwiesen. Besonders deutsche Durchschnittsnoten erreichen hohe Validitäten, die auch von studienfachnahen Schulnoten nicht übertroffen werden. Zur Berechnung zukünftiger Metaanalysen ist bei der Publikation empirischer Primärstudien unbedingt die Angabe von Selektions- und Drop-out-Quoten zu fordern. Desweiteren wäre es wünschenswert, dass zukünftig die Interkorrelationen zwischen den einzelnen verwendeten Auswahlkriterien und verschiedenen Studienerfolgsindikatoren berichtet werden.

Für eine verlässliche Abschätzung der Validität der Schulnoten für andere Erfolgskriterien wie Studienabbruch, Studienzufriedenheit, Studiendauer oder auch Be-

rufserfolgskriterien liegen derzeit noch zu wenige hochwertige Primärstudien vor. Hier besteht weiterhin Forschungsbedarf.

Angesichts der guten Validitäten und der günstigen und einfachen Einsetzbarkeit wäre es sicher ein Fehler, Schulnoten bei der Auswahl von Studienplatzbewerbern nicht zu berücksichtigen.

Autorenhinweis

Die Autoren danken Nambury Raju†, Stephan J. Rustenbach, Frank L. Schmidt und Ralf Schulze für ihre Unterstützung bei der Klärung methodischer Detailfragen.

Literatur

Studien, die in die Metaanalyse einbezogen wurden, sind mit einem Stern gekennzeichnet.

- *Barkowski, D. (1986). *Vorhersage von Studien- und Berufserfolg bei Zahnmedizinern. Längsschnittvalidierung einer apparativen Testbatterie zur Eignungsdiagnose für das Studium der Zahnmedizin*. München: Profil.
- Baron-Boldt, J., Schuler, H. & Funke, U. (1988). Prädiktive Validität von Schulabschlussnoten: Eine Metaanalyse. *Zeitschrift für Pädagogische Psychologie*, 2, 79–90.
- Bortz, J. & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Berlin: Springer-Verlag.
- *Brandstätter, H. & Farthofer, A. (2002). Studienerfolgsprognose – konfigurativ oder linear additiv? *Zeitschrift für Differentielle und Diagnostische Psychologie*, 23, 381–391.
- Burton, N. W. & Ramist, L. (2001). *Predicting success in college: SAT studies of classes graduating since 1980*. The College Board Research Report, 2001–2002.
- *Crum, R. & Parikh, A. (1983). Headmasters' reports, admissions and academic performance in social sciences. *Educational Studies*, 9, 169–184.
- Deidesheimer Kreis. (1997). *Hochschulzulassung und Studienerfolg*. Göttingen: Vandenhoeck & Ruprecht.
- *Duff, A., Boyle, E., Dunleavy, K. & Ferguson, J. (2004). The relationship between personality, approach to learning and academic performance. *Personality and Individual Differences*, 36, 1907–1920.
- *Ferguson, E., James, D., O'Hehir, F. & Sanders, A. (2003). Pilot study of the roles of personality, references, and personal statements in relation to performance over the five years of a medical degree. *British Medical Journal*, 326, 429–431.
- *Giesen, H., Gold, A., Hummer, A. & Jansen, R. (1986). *Prognose des Studienerfolgs*. Frankfurt am Main: Institut für Pädagogische Psychologie.
- Gold, A. & Souvignier, E. (2005). Prognose der Studierfähigkeit – Ergebnisse aus Längsschnittanalysen. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie*, 37, 214–222.
- Heine, C., Briedis, K., Didi, H.-J., Haase, K. & Trost, G. (2006). *Bestandsaufnahme von Auswahl- und Eignungsfeststellungs-*

- verfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern. Hannover: HIS-Kurzinformation A 3/2006.
- *Hell, B. (2003). *Kognitive Leistungsfähigkeit in der Berufseignungsdiagnostik – Forschungsüberblick und Entwicklung eines neuen Instruments zur Erfassung Schlussfolgernden und Kreativen Denkens*. Berlin: dissertation.de.
- Hell, B., Trapmann, S. & Schuler, H. (2007). *Eine Metaanalyse der Prognosekraft von Studierfähigkeitstests*. Zur Veröffentlichung eingereicht.
- *Höppel, D. & Moser, K. (1993). Die Prognostizierbarkeit von Studiennoten und Studiendauer durch Schulabschlußnoten. *Zeitschrift für Pädagogische Psychologie*, 7, 25–32.
- *Hörschgen, H., Cierpka, R., Friese, M. & Steinbach, R. (1993). *Erfolg im Studium und Beruf*. Universität Hohenheim – Institut für Betriebswirtschaftslehre.
- *Höschl, C. & Kozeny, J. (1997). Predicting academic performance of medical students: The first three years. *American Journal of Psychiatry*, 154, 87–92.
- 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.
- Ingenkamp, K. (1997). *Lehrbuch der Pädagogischen Diagnostik*. Weinheim: Beltz.
- 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.
- *Kurz, G., Fischer, W. & Wagner, H. (1995). Prognose des Studienerfolgs in Studiengängen des Maschinenbaus Untersuchungen an der FHT Esslingen und an der FH Karlsruhe. *Empirische Pädagogik*, 9, 331–360.
- Law, K. S. (1995). The use of Fisher's Z in Schmidt-Hunter-type meta-analyses. *Journal of Educational and Behavioral Statistics*, 20, 287–306.
- *McManus, I. C. & Richards, P. (1986). Prospective survey of performance of medical students during preclinical years. *British Medical Journal*, 293, 124–127.
- *Meier, B.-D. (2003). Ist der Erfolg im Jurastudium vorhersagbar? Empirische Befunde zum Zusammenhang zwischen Schulnoten und Abschneiden im Ersten Juristischen Staatsexamen. *Beiträge zur Hochschulforschung*, 25, 18–35.
- *Melamed, T. (1992). Use of biodata for predicting academic success over thirty years. *Psychological Reports*, 71, 31–38.
- Mendoza, J. L. & Reinhardt, R. N. (1991). Validity generalization procedures using sample-based estimates: A comparison of six procedures. *Psychological Bulletin*, 110, 596–610.
- *Montague, W. & Odds, F. C. (1990). Academic selection criteria and subsequent performance. *Medical Education*, 24, 151–157.
- Novy, D. M., Kopel, K. F. & Swank, P. R. (1996). Psychometrics of oral examinations for psychology licensure: The Texas examination as an example. *Professional Psychology: Research & Practice*, 27, 415–417.
- *Ohlsen, U. (1985). *Eine empirische Untersuchung der Einflußgrößen des Examenserfolgs für Absolventen wirtschaftswissenschaftlicher Studiengänge der Universität Münster*. Frankfurt: Lang.
- 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.
- 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.
- Robbins, S. B., Lauer, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin*, 130, 261–288.
- Rustenbach, S. J. (2003). *Metaanalyse – Eine anwendungsorientierte Einführung*. Bern: Huber.
- *Schiefele, U., Streblow, L., Ermgassen, U. & Moschner, B. (2003). Lernmotivation und Lernstrategien als Bedingungen der Studienleistung: Ergebnisse einer Längsschnittstudie. *Zeitschrift für Pädagogische Psychologie*, 17, 185–198.
- 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-Atzert, L. (2005). Prädiktion von Studienerfolg bei Psychologiestudenten. *Psychologische Rundschau*, 56, 131–133.
- Schuler, H. (2006). Noten als Prädiktoren von Studien- und Berufserfolg. In D. H. Rost (Hrsg.), *Handwörterbuch Pädagogische Psychologie* (3. Aufl., S. 535–541). Weinheim: Beltz.
- Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.
- Silver, N. C. & Dunlap, W. P. (1987). Averaging correlation coefficients: Should Fisher's z transformation be used? *Journal of Applied Psychology*, 72, 146–148.
- *Steyer, R., Yousfi, S. & Würfel, K. (2005). Prädiktion von Studienerfolg: Der Zusammenhang zwischen Schul- und Studiennoten im Diplomstudiengang Psychologie. *Psychologische Rundschau*, 56, 129–131.
- *Stumpf, H. & Fay, E. (1991). Zur prognostischen Validität des «Tests für medizinische Studiengänge» (TMS) in den Studiengängen Tier- und Zahnmedizin. *Diagnostica*, 37, 213–225.
- *Stumpf, H. & Nauels, H.-U. (1990). Zur prognostischen Validität des «Tests für medizinische Studiengänge» (TMS) im Studiengang Humanmedizin. *Diagnostica*, 36, 16–32.
- Tent, L. (2006). Zensuren. In D. H. Rost, *Handwörterbuch Pädagogische Psychologie* (S. 873–880). Weinheim: Beltz.
- *Trost, G. (Hrsg.). (1998). *Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse*. Bonn: Institut für Test- und Begabungsforschung.
- *Trost, G. & Freitag, G. (1991). Prognostische Validität studienfeldbezogener Tests zur Beratung von Studierwilligen. In H. Schuler & U. Funke (Hrsg.), *Eignungsdiagnostik in Forschung und Praxis* (S. 191–193). Stuttgart: Verlag für Angewandte Psychologie.
- *Trost, G. & Piel, P. (1991). Der Aufnahmewettbewerb der WHU

- eine Zwischenbilanz. *Hochschulnachrichten aus der Wissenschaftlichen Hochschule für Unternehmensführung*, 6, 60–65.
- *Vollmer, F. (1984). Expectancy and academic achievement. *Motivation & Emotion*, 8, 67–76.
- *Wallace, W. F. M., Sanderson, G., Mercer, C. G. & Gilmore, R. S. (1988). Short- and long-term academic predictors of medical student performance. *Ulster Medical Journal*, 57, 149.
- 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.
- Willingham, W. W., Lewis, C., Morgan, R. & Ramist, L. (1990).

Predicting college grades: An analysis of institutional trends over two decades. Princeton: Educational Testing Service.

Dipl.-Psych. Sabrina Trapmann

Universität Hohenheim
 Lehrstuhl für Psychologie (540F)
 D-70593 Stuttgart
 Tel. +49 711 459-23325
 E-Mail trapmann@uni-hohenheim.de