



Qualitätssicherung und -entwicklung in der Hochschule: Methoden und Ergebnisse

*herausgegeben von
Annette Kluge und Kerstin Schüler*

Annette Kluge, Kerstin Schüler (Hrsg.)

Qualitätssicherung und -entwicklung
an Hochschulen:
Methoden und Ergebnisse



PABST SCIENCE PUBLISHERS
Lengerich, Berlin, Bremen, Miami,
Riga, Viernheim, Wien, Zagreb

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Das Werk, einschließlich aller seiner Teile, ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

© 2007 Pabst Science Publishers, D-49525 Lengerich
Druck: KM Druck, D-64823 Groß-Umstadt

ISBN 978-3-89967-336-4

Inhaltsverzeichnis

Kontext und Rahmen:

Die Hintergrundgeschichte(n) und das Ziel dieses Buches 5

Qualitätssicherung und Zulassungsverfahren 9

Benedikt Hell

Qualitätssicherung an Hochschulen durch Zulassungsverfahren 11

Margarete Litzenberger, Joachim F. Punter, Timo Gnams, Marco Jirasko und
Christiane Spiel

Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch
fundierter Wissensprüfung..... 23

Cornelia Schättle und Jürgen Wegge

Determinants of success in an international Master Program at the
LMU München: A field study..... 35

Qualitätssicherung und der Einsatz von Online-Verfahren 47

Ruth Rustemeyer, Claudia Callies, Peter Ferdinand und Michael Kutscher

Durchführung von Online-Klausuren mit Standard-E-Learning-Werkzeugen.
Ausgewählte Evaluationsergebnisse 49

Karen Tinsner und Markus Dresel

Onlinebefragung in der Lehrveranstaltungsevaluation: Ein faires, verzerrungsfreies
und ökonomisches Verfahren?..... 59

Kompetenzen als Qualitätskriterium..... 71

Edith Braun

Ergebnisorientierte Lehrveranstaltungsevaluation: Das Berliner
Evaluationsinstrument für studentische Kompetenzen 73

Manuela Paechter, Brigitte Maier, Alexandra Dorfer, Gudrun Salmhofer und
Alexandra Sindler

Kompetenzen als Qualitätskriterien für universitäre Lehre: Das Grazer
Evaluationsmodell des Kompetenzerwerbs (GEKo) 83

Karl-Josef Kluge und Julia Maria Siebert

Studieren heißt investieren - und zwar auf beiden Seiten
Qualitätsentwicklung & -sicherung im Lehrer-, Diplom- & Doktoranden-Studium
in der Universität zu Köln..... 95

Studienqualität	113
Hans-Peter Langfeldt, Susanne Frühauf und Tanja Nieder Beurteilung von Studienqualität durch Studierende	115
Hans-Peter Langfeldt, Tanja Nieder und Susanne Frühauf Der Frankfurter Studienqualitäts-Fragebogen (FSQF)	125
Walter H. Schreiber und Bernhard Wolf Was ist gut, was ist schlecht? Die Universität im Urteil von Studierenden dreier Diplomstudiengänge	133
Alexandra Hein, Ulrike Englisch und Helmut Niegemann Studienrichtungsevaluation zur Unterstützung der Qualitätsentwicklung im Bologna-Prozess:	145
Lehrevaluation - Auswirkung auf Lehrende	157
Jonas W. B. Lang und Martin Kersting Langfristige Effekte von regelmäßigem Feedback aus studentischen Lehrveranstaltungsevaluationen	159
Annette Kluge „Ob meine Veranstaltung erfolgreich war, sehe ich an den Gesichtern der Studierenden“ - eine explorative Studie zu Auslösern für Veränderungen und subjektive Qualitätskriterien von guten Veranstaltungen	169
Annette Kluge und Kerstin Schüler Der Einfluss von Erwartungen verschiedener Anspruchsgruppen auf Evaluationsergebnisse und potentiell Engagement für Qualitätsentwicklung in der Lehre	181
Markus Dresel, Heiner Rindermann und Karen Tinsner Beratung von Lehrenden auf der Grundlage studentischer Veranstaltungsbeurteilungen.....	193
Michael Henninger und Michael Balk Bewertung der Bewertung: Was halten Lehrende von der Evaluation ihrer Veranstaltungen durch Studierende?	205
Aline Vater, Paul Grohs und Peter Sedlmeier Die Beeinflussbarkeit studentischer Urteile in universitären Lehrevaluationen durch potenzielle Biasvariablen.....	219

Standardisierte Verfahren zur Erhebung von Publikations- und Zitationsdaten	233
Gabriel Schui und Günter Krampen Qualitätssicherung und -entwicklung an Universitäten im Bereich der Forschungsaktivitäten – Systematische Erhebung von Publikationsaktivitäten und internationaler Rezeption	235
Organisationaler Wandel und Image.....	243
Daniela Meiser, Ralf Stegmaier, Alexandra Michel und Karlheinz Sonntag Evaluation von Veränderungen im Hochschulbereich: Prozessmerkmale und Persönlichkeit als Determinanten von Zufriedenheit und Wechselbereitschaft	245
Alexandra Michel, Karlheinz Sonntag, Ralf Stegmaier und Daniela Meiser Auswirkungen von Veränderungsprozessen im Hochschulbereich auf Stresserleben und Arbeitszufriedenheit wissenschaftlicher Mitarbeiter	257
Peter Lindstrom Professional Institutional Accreditation: Is It Providing Utility to European Business Schools?	269
Verena Heukamp, Jan Schilling und Lutz Hornke Imageanalyse einer Hochschule – Das Fremdbild kennen, um es zu gestalten.....	279
Sascha Spoun und Jan Henrik Sieg Qualitätsmanagement durch universitäre Ideen in modernen Prozessen: Methoden und Ergebnisse aus der Praxis der Assessment-Stufe.....	297
Autorinnen- und Autorenverzeichnis	311

Kontext und Rahmen:

Die Hintergrundgeschichte(n) und das Ziel dieses Buches

Es gibt zwei Hintergrundgeschichten zu diesem Buch. Eine der Hintergrundgeschichten basiert auf meinen persönlichen Erfahrungen in einem Forschungsprojekt zur Qualitätssicherung und -entwicklung (QS/QE) an Hochschulen in der Schweiz im Jahr 2004-2005. Ich untersuchte damals die Frage, welche Auswirkungen standardisierte Lehrevaluationen auf die Qualitätsentwicklung haben. Dazu sprach ich mit KollegInnen anderer Universitäten im deutschsprachigen Raum und lernte dort Personen kennen, die wenig Zugang zu der fachlichen Diskussionen der Psychologie im Rahmen der Diagnostik oder Pädagogischen Psychologie, der HochschullehrerInnenpsychologie und der Organisationspsychologie haben und u. a. deshalb wenig der dort gewonnenen Erkenntnisse in ihrer Arbeit umsetzen können.

Fragen an mich lauteten z. B. „Welche Evaluationsinstrumente haben sich denn bewährt?“, „Zeigt sich ein Unterschied in den Bewertungen der Dozierenden, ob ich online oder in Papierform evaluiere?“, „Werden schwierige Veranstaltungen schlechter bewertet?“, „Wie kann man Lehrende nach erfolgter Evaluation unterstützen?“, „Zu welchen Ergebnissen führt eine EFQM (European Foundation for Quality Management) oder EQUIS (European Quality Improvement System) Akkreditierung?“, „Wie bewegt man die Studierenden und Dozierenden zum Mitmachen?“. Beim gemeinsamen Suchen nach Antworten zeigte sich für mich ein großer Graben zwischen den schon erarbeiteten und vorliegenden Ergebnissen und psychologischen Erkenntnissen einerseits und der täglichen Praxis der QS/QE Aktivitäten andererseits.

Eine zweite parallele Hintergrundgeschichte beginnt mit der Organisation einer Arbeitsgruppe für den Kongress der Deutschen Gesellschaft für Psychologie in Nürnberg im September 2006. Ich wollte eine interdisziplinäre Arbeitsgruppe zum Thema QS/QE an Hochschulen anbieten, in denen die KollegInnen, die sich mit QS/QE aus unterschiedlichen psychologischen Disziplinen beschäftigen, ihre Ergebnisse präsentieren können. Der ursprünglich intendierte Fokus lag dabei auf der Schnittmenge zwischen Pädagogischer Psychologie und Organisationspsychologie und ich hatte die ernsthafte Befürchtung, dass alle diejenigen, die sich auf diese Weise disziplinübergreifend mit QS/QE beschäftigen, in einer kongressorganisatorischen „Restkategorie“ landen und das Thema trotz hoher Aktualität ein zerpfücktes Nischendasein fristen müsste. Auf die Einladung zum Einreichen von Abstracts über die Verteiler der Fachgruppen Pädagogische Psychologie und Arbeits- und Organisationspsychologie für eine solche Arbeitsgruppe erhielten wir 20 Beiträge. Diese positive Resonanz traf uns unerwartet, wir wollten aber auf keinen einzigen dieser Beiträge verzichten. Die Kongressausrichter zeigten sich zudem sehr offen für drei mögliche Arbeitsgruppen mit jeweils spezifischem Fokus auf einen Teilaspekt von QS/QE.

Begeistert von der Vielfalt der Themen und den interessanten und spannenden Fragen, die in den Beiträgen behandelt wurden, kamen mir wieder die Erfahrungen mit „dem Graben“ zwischen Theorie und Praxis in den Sinn. Wir überzeugten

zunächst Herrn Pabst vom Pabst Verlag und dann die AutorInnen der Kongressbeiträge sich an einem Herausgeberinnen-Band zu beteiligen.

Das Ziel dieses Buches

Über den eingangs beschriebenen Graben wollen wir mit unserem Herausgeberinnen-Band nun eine Brücke bauen. Ziel dieser Publikation ist es, den Personen und Verantwortlichen an Hochschulen und Universitäten, die in Bezug auf Aspekte zur Qualitätsentwicklung und -sicherung Entscheidungen treffen, den Stand der Forschung und praktischer Erfahrungen aus Deutschland, Österreich und der Schweiz nahe zu bringen.

Dabei gehen wir davon aus, dass die meisten Entscheidungsträger und -verantwortlichen überwiegend keine PsychologInnen oder PädagogInnen sind, sondern aus unterschiedlichen Fachrichtungen kommen und in Stäben und unterstützenden Bereichen im Hochschul- und Universitätswesen mit konkreten Fragen der Implementierung und des „Doing“ konfrontiert sind.

KollegInnen verschiedener psychologischer und pädagogischer sowie betriebswirtschaftlicher Fachrichtungen stellen ihre aktuellen Forschungsergebnisse vor und leiten daraus Handlungsvorschläge und -anregungen ab.

Die AutorInnen thematisieren Aspekte wie Studierendenauswahlverfahren, Erfahrungen mit on- und offlinebasierten Evaluations- und Test-Verfahren, Kompetenzerwerb als Qualitätskriterium an Hochschulen und Universitäten, Messung von Studienqualität, Lehrevaluationen und ihre Auswirkungen auf Lehrende, Evaluation von Publikationsaktivitäten als auch Auswirkungen von QS/QE-Verfahren auf eine Organisation und wie Wandelprozesse von den Beteiligten erlebt werden.

Fast alle Beiträge beginnen mit einer Zusammenfassung internationaler Forschungsergebnisse und einem „State-of-the-art“, aus der die jeweils eigene Forschungsfrage entwickelt wird. Die AutorInnen berichten dann von ihrer Untersuchungs- oder Forschungsmethodik mit der Absicht, den EntscheidungsträgerInnen und -verantwortlichen im Hochschulwesen Ideen und Anregungen für die QS/QE mitzugeben. Die AutorInnen waren auch sehr gerne bereit, die in den Untersuchungen eingesetzten Fragebogen und Instrumente detailliert zu erläutern oder sogar abdrucken zu lassen. Dafür bin ich den AutorInnen sehr dankbar, denn damit hat dieses Herausgeberinnen-Werk auch einen konkreten praktischen Nutzen für alle diejenigen, die nach konkreten Instrumenten als Orientierungshilfen suchen.

Ein weiterer Teil der Beiträge berichtet auch von eigenen Erfahrungen mit dem Leben des Qualitätsgedankens in der Gestaltung von Universität und Veranstaltungen.

Was es nicht ist!

Diese Aufsatzsammlung will nicht zur Diskussion des Qualitätsbegriffs als solchem und seiner Passung auf Bildungsfragen beitragen. Es geht nicht darum zu fragen, ob Studierende KundInnen sind, oder vielleicht die Gesellschaft oder die Unternehmen und Organisationen, in denen die AbgängerInnen nach ihrem Studienabschluss später arbeiten werden. Wir wollen nicht die Sinnhaftigkeit der Bologna Reform diskutieren oder kritisieren.

Die AutorInnen vergleichen in ihren Beiträgen viel mehr die gelebte Praxis und die darin verwendeten Methoden und Verfahren mit den selbst gesteckten und angestrebten Zielen dieser Veränderungen und Maßnahmen und berichten von ihren Ergebnissen und Erfahrungen. Das Verständnis von QS/QE wird jeweils von den AutorInnen mehr oder weniger implizit und explizit angesprochen. Im Vordergrund stehen in den Beiträgen jedoch die erlebte Praxis und die sich daraus ergebenden Probleme und Herausforderungen.

Vielen Dank an alle AutorInnen, die dieses Buch möglich gemacht haben!

St. Gallen, im Januar 2007

Prof. Dr. Annette Kluge

Qualitätssicherung und Zulassungsverfahren

Qualitätssicherung an Hochschulen durch Zulassungsverfahren¹

Benedikt Hell

Zusammenfassung: Dieser Beitrag zeigt auf, dass Zulassungsverfahren einen wichtigen Beitrag zur Qualitätssicherung im tertiären Bildungsbereich leisten können. In Form eines Gedankenspiels wird das Portersche Value Chain Framework herangezogen, um die besondere Stellung von Auswahlverfahren als Qualitätssicherungselement aufzuzeigen. Einen weiteren Schwerpunkt bildet die Analyse der prognostischen Aussagekraft und der Betrachtung des Nutzens von Auswahlverfahren. Hierzu werden Kennzahlen präsentiert, die durch den Einsatz von Auswahlverfahren in unterschiedlichen Auswahl-situationen zu erwarten sind. Verfahrensweisen zur Bestimmung des Nutzens von Auswahlverfahren werden ebenso vorgestellt wie potenzielle Defizite von Nutzenkalkulationen.

Schlagwörter: Studienerfolg, Studierendenauswahl, Zulassungskriterien, Trefferquoten, Nutzenkalkulation

Abstract: This contribution points out that educational admission and placement can add substantial value to quality management within higher education. Using the value chain framework as a thought experiment the significant position of admission procedures as first link of a value chain is underlined. Subsequently the validity and the use of admission criteria for different selection situations are discussed. A model for defining the utility of admission criteria and the potentially shortcomings are described.

Key Words: Academic Performance, Admission to Higher Education, Admission Criteria, Hit Rates, Utility of Admission Criteria

¹ Dieser Beitrag entstand innerhalb 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.

Hochschulen als Bestandteil einer Wertkette

In jüngster Zeit ist ein stark ansteigendes Interesse an Auswahl- und Zulassungsverfahren für den Hochschulbereich zu verzeichnen. Ein besonderer Impuls ging von der Novelle des Hochschulrahmengesetzes im Jahre 1998 aus. Den Hochschulen wurden erheblich mehr Freiheiten bei der Zulassung eingeräumt und in zunehmendem Ausmaße nutzen sie die neu entstandenen Freiheiten dazu, Auswahlverfahren zu etablieren (Heine, Briedis, Didi, Haase & Trost, 2006).

Es spricht tatsächlich Vieles dafür, Zulassungskriterien als wichtiges Element der Qualitätssicherung im tertiären Bildungsbereich anzusehen. Ein Gedankenspiel möge die Bedeutsamkeit der Zulassungsverfahren verdeutlichen: Der in der betriebswirtschaftlichen Literatur sehr prominente Value-Chain-Ansatz von Porter (1985) bietet einen Bezugsrahmen, innerhalb dessen die Studierendenauswahl als Element der „Wertschöpfungskette“ betrachtet werden kann. Porter formulierte das in Abbildung 1 skizzierte Wertketten-Rahmenmodell zur Analyse von Unternehmensstrukturen und -prozessen.



Abbildung 1. Wertketten-Rahmenmodell nach Porter (1985)

Das Modell systematisiert Wertschöpfungsprozesse beginnend bei der Eingangslogistik (Beschaffung von Rohstoffen, Bestandskontrolle, Transportplanung) über die Transformation (Fertigung, Verpacken, Wartung von Ausrüstungsgegenständen, Prüfung) bis hin zur Ausgangslogistik (Einlagerung, Auftragerfüllung, Transport, Verteilungsmanagement) und dem Vertrieb (Annoncieren, Verkaufen, Preiskalkulation). Es wurde bereits auf zahllose Branchen übertragen und auf unterschiedlichen Abstraktionsebenen eingesetzt und kann auch im Kontext der Studierendenauswahl Anregungen liefern.

Auch wenn eine Übertragung des Modells auf die Studierendenauswahl eine reduktionistische Sichtweise darstellt und förmlich zum Widerspruch auffordert, bietet es sich an, um die Bedeutsamkeit von Auswahlverfahren zu veranschaulichen: Aus welchen Elementen besteht die Wertkette eines Unternehmens? Den Ausgangspunkt der Wertschöpfung stellt in der Regel ein Rohstoff dar, der von verschiedenen Lieferanten angeboten wird. Im Rahmen der Eingangslogistik werden Lieferanten und eingehende Waren einer Qualitätskontrolle unterzogen. Bei der sich anschließenden Transformation der Rohstoffe zu einem Produkt ist das Unternehmen daran interessiert, die Durchlaufzeit und die Ausschussquote zu minimieren. Bevor das Produkt auf den Markt gelangen kann, muss die Qualität durch eine Endabnahme geprüft werden. Schließlich muss das Endprodukt durch geeignete verkaufsfördernde Werbemaßnahmen auf dem Markt platziert werden.

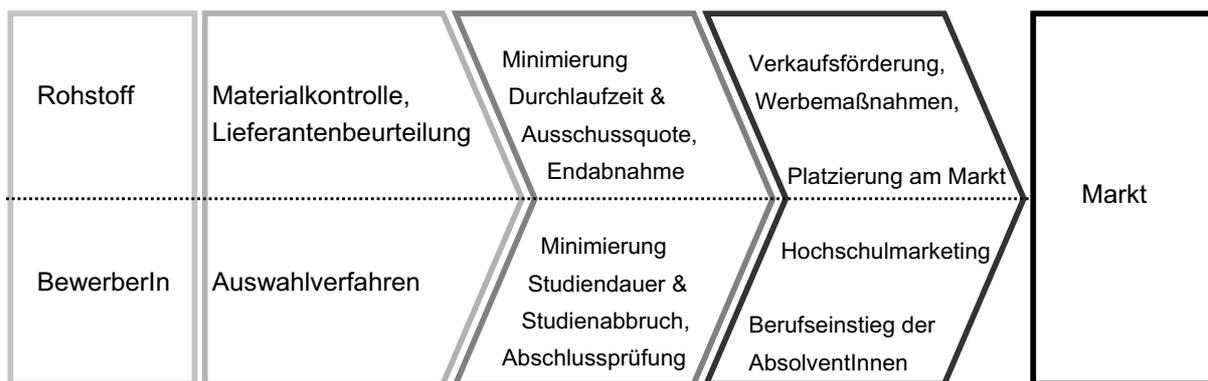


Abbildung 2. Auswahlverfahren als Schlüsselement einer Wertkette

Übertragen auf den Kontext der Studierendenauswahl käme den Zulassungsverfahren die Aufgabe zu, eine Qualitätskontrolle des „Rohstoffs“ BewerberIn – besonders unfreundlich formuliert: der humanen Ressource BewerberIn – zu sichern. Während des Studiums ist es für die Hochschule entscheidend, die Durchlaufzeit (Studiendauer) und die Ausschussquote (Studienabbruch) zu minimieren, sowie ein durch die Endabnahme (Abschlussprüfung) zertifiziertes „Produkt“ bereitzustellen. Das Hochschulmarketing kann anschließend den Berufseinstieg der AbsolventInnen durch geeignete Maßnahmen (z. B. Karrieremessen) fördern. Die inzwischen von den meisten Hochschulen als Wirkungsfeld erkannte Alumni-Arbeit kann als After Sales Management verstanden werden.

Selbstverständlich stellt die entwickelte Analogie eine *unzulässige Reduktion* dar, indem die Studierendenschaft als passive anonyme Masse betrachtet wird, die durch eine universitäre Ausbildung zur „Veredelung“ gelangt. Selbstverständlich ist es zweckmäßiger und angemessener, die Studierenden als *KundInnen* bzw. als PartnerInnen anzusehen. Die Hochschulen sind für die Studierenden da und nicht umgekehrt. Zudem haben die Hochschulen einen Bildungsauftrag und sollten nicht vorrangig auf eine „Minimierung der Durchlaufzeit“ und eine „Reduktion der Ausschussquote“ abzielen. Die hergestellte Analogie hilft aber, die Rolle von Zulassungsverfahren als Element der Qualitätssicherung zu verdeutlichen. Denn wird die Ausbildung von Studierenden als mehrgliedriger Prozess betrachtet, der von den BewerberInnen über die StudienanfängerInnen und AbsolventInnen bis zu reüssierenden BerufsanfängerInnen reicht, so wird unmittelbar ersichtlich, dass die Zulassungsverfahren den Beginn der Kette bilden und somit alle nachfolgenden Prozesse entscheidend beeinflussen: Je genauer die angenommenen BewerberInnen die Anforderungen erfüllen, die das Studium an sie stellen wird, desto kürzer werden die Studienzeiten und geringer die Abbruchquoten ausfallen. Letztlich sollten geeignete KandidatInnen bessere Abschlussprüfungen erzielen und sich schneller und erfolgreicher beruflich etablieren. Hochwertige Auswahlverfahren dienen letztlich auch dem Hochschulmarketing, wenn wie etwa in den Vereinigten Staaten offen damit geworben wird, wie anspruchsvoll die Eingangsvoraussetzungen der Hochschulen sind.

Aber nicht nur aus Sicht der Universitäten sondern auch aus Sicht der StudienplatzbewerberInnen sind valide Zulassungsverfahren wünschenswert. Ungeeigneten BewerberInnen werden die Frustrationen erspart, die mit der Wahl eines falschen Studiums und daraus resultierenden langen Studienzeiten oder gar Studienabbrüchen

verbunden sind. Ideelle und reale Kosten falsch investierter Lebenszeit werden durch aussagefähige Zulassungsverfahren reduziert. Die geeigneten BewerberInnen wiederum erhalten im Idealfall eine größere Chance auf einen Studienplatz in einem Studiengang, den sie mit einer größeren Wahrscheinlichkeit erfolgreich abschließen werden. Einen noch größeren Nutzen für die BewerberInnen haben Auswahlverfahren, die nicht nur Aufschluss über die Eignung für *einen* Studiengang geben, sondern eine *Zuordnungsberatung* erlauben. Sinnvoll wären Verfahren, die etwa das Fähigkeits- und Interessenprofil von BewerberInnen mit den Profilen von möglichst vielen Studiengängen abgleichen.

Nutzen von Verfahren der Studierendenauswahl

Ergebnisse der Hohenheimer Metaanalysen

Zur Klärung der bedeutsamen Frage, welche diagnostischen Verfahren geeignet sind, eine Prognose des Studienerfolgs zu ermöglichen, wurden in den vergangenen Jahrzehnten zahlreiche empirische Studien publiziert. Eine geeignete Methode zur Integration einer großen Zahl von Einzelstudien, die ihre Erkenntnisse mittels vergleichbarer Forschungsdesigns gewinnen, stellt die so genannte Metaanalyse bzw. die Methode der Validitätsgeneralisierung dar (Hunter, Schmidt & Jackson, 1982). Ein Teil dieses Beitrags wird sich auf die Ergebnisse aktueller Metaanalysen der Hohenheimer Arbeitsgruppe beziehen (Hell, Trapmann & Schuler, unter Begutachtung; Hell, Trapmann, Weigand & Schuler, in Druck; Trapmann, Hell, Weigand & Schuler, in Druck). In diesen Beiträgen werden die wichtigsten empirischen Studien, die die Prognosekraft von Schulnoten, Studierfähigkeitstests und Auswahlgesprächen untersuchen und in der Zeit zwischen 1980 und 2005 erschienen sind, statistisch und inhaltlich aggregiert.

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. Dieses Verständnis von Studienerfolg legt eine Operationalisierung über die Abbrecherquote nahe. 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 DozentInnen oder KommilitonInnen, sowie Angaben über die Zufriedenheit mit dem Studium via Selbsteinschätzung.

Da die weitaus meisten empirischen Studien zur Validierung von Auswahlverfahren das Erfolgskriterium Studiennoten heranziehen, wurde dieses Kriterium im Rahmen der Hohenheimer Metaanalysen zur Integration der Einzelbefunde verwendet. Als aussagekräftigster Prädiktor für die Vorhersage von Studiennoten erwiesen sich Durchschnittsnoten des Schulabschlusses mit einer Validität von $r=.52$ (Trapmann et al., in Druck). Die aktuelle Metaanalyse kommt somit zu einer etwas höheren Schätzung als die vorangegangene Metaanalyse der Hohenheimer Arbeitsgruppe zur Prognosekraft von Schulnoten (Schuler, Funke & Baron-Boldt, 1990). Angesichts der ausgezeichneten Validität der Schulnoten und des geringen Aufwands, diese im Kontext der Studienzulassung zu ermitteln, empfiehlt es sich, Schulnoten als

Zulassungskriterium heranzuziehen. Basis sollten primär die Abschlussdurchschnittsnoten sein, da diese in der Regel aussagekräftiger als Einzelnoten sind. Allerdings gibt es Studienfächer, bei denen zusätzlich zu Durchschnittsnoten berücksichtigte Einzelnoten einen ergänzenden Prognosebeitrag leisten. So hat sich beispielsweise bei einigen naturwissenschaftlichen Studienfächern gezeigt, dass die Aussagekraft der Abiturdurchschnittsnote durch eine separate Gewichtung von naturwissenschaftlichen Schulfächern noch gesteigert werden kann.

Wie ist die ausgezeichnete Validität der Schulnoten zu erklären? Zwei Aspekte sind zu differenzieren: Zum einen kann Validität auf die in den Schulnoten *abgebildeten Personenmerkmale* zurückgeführt werden. Noten sind nicht nur Indikatoren für das akkumulierte schulische Wissen einer Person, sondern es wirken eine ganze Reihe von Persönlichkeitseigenschaften und Fähigkeiten bei schulischen Leistungen zusammen, die auch für den Erfolg an der Hochschule bedeutsam sind. Zu nennen wären etwa kognitive Fähigkeiten, Ausdrucksfähigkeit, Lernbereitschaft, Fleiß, Ausdauer und Leistungsmotivation. Die außerordentliche Validität von Schulnoten ist darüber hinaus *messtheoretisch begründet*, denn es handelt sich nicht um eine einmalige – mitunter von der Tagesform abhängige – Messung, sondern um eine Aggregation zahlreicher einzelner Messungen, die von mehreren BeurteilerInnen über mehrere Jahre in mehreren Prüfungsmodalitäten (mündlich, schriftlich) vorgenommen wurden. Ein jeder, der sich durch einzelne Lehrerurteile zu schlecht eingestuft sieht, wird zugeben, dass er in anderen Fällen ungerechtfertigter Weise zu gut bewertet wurde. In statistischen Termini gesprochen führt die Aggregation zahlreicher Einzelmessungen mit normalverteilten Fehlerkomponenten im Mittel zu einem aussagekräftigen Leistungsindikator.

Sowohl allgemeine als auch fachspezifische Studierfähigkeitstests verfügen ebenfalls über eine sehr gute prognostische Güte ($r=.48$; Hell et al., unter Begutachtung). Sollte der Einsatz dieser Verfahrensklasse in Betracht gezogen werden, böte ein landesweit einheitlich für ein Studienfach angebotener Test gegenüber hochschulspezifischen Verfahren einige Vorteile. So wäre es den BewerberInnen möglich, den Test einmal zu absolvieren und sich anschließend bei verschiedenen Hochschulen mit einem Zertifikat zu bewerben; der durch mehrfache Testteilnahmen an mehreren Hochschulen entstehende, für BewerberInnen und Hochschulen erhebliche Aufwand würde insgesamt reduziert. Gleichzeitig könnte die Testentwicklung und -evaluation zentral erfolgen. Dadurch böte sich die Gelegenheit, kostengünstige und qualitativ hochwertige Verfahren bereitzustellen und ihre Qualitätssicherung auf Dauer zu gewährleisten.

Die mittlere Validität von Auswahlgesprächen beträgt $r=.16$, wobei die Vorhersagekraft durch Strukturierung bedeutsam auf $r=.21$ gegenüber $r=.11$ für unstrukturierte Interviews verbessert werden kann (Hell et al., in Druck). Interviews weisen für Studienerfolg im Sinne von Studiennoten also keine hohe Prognosegüte auf. Über die Prognosekraft von Interviews für die anderen eingangs erwähnten Erfolgskriterien liegen zu wenige aussagekräftige empirische Studien vor, als dass eine metaanalytisch abgesicherte Abschätzung sinnvoll wäre. Denkbar ist jedoch, dass das wichtige Studienerfolgskriterium Studienabschluss (vs. Studienabbruch) durch sorgsam geführte Auswahlgespräche in entscheidender Weise positiv beeinflusst werden kann. Interviews können neben der Funktion als Selektionsinstrument auch zur Studien- und Platzierungsberatung eingesetzt werden. Es ist möglich, im Gespräch auf den Bewerber/die Bewerberin einzugehen und die individuellen Besonderheiten

und Talente besonders zu berücksichtigen. Weiterhin spricht für den Einsatz von Interviews, dass aktuelle Metaanalysen strukturierten Interviews substanzielle Validität für Berufserfolg und inkrementellen Zusatznutzen gegenüber anderen Verfahren bescheinigen (zusammenfassend: Schuler, 2002). Eine auf den Berufserfolg ausgerichtete Auswahl von Studierenden ist in Fächern denkbar, mit denen ein klar definiertes Berufsbild verbunden ist (z. B. Lehramtsstudiengänge).

Inkrementelle Validität und Nutzensteigerung durch Kombination mehrerer Auswahlkriterien

Die gemeinsame Validität zweier Prädiktoren hängt zum einen von der eigenständigen Validität eines jeden Prädiktors und zum anderen von der Korrelation der beiden Prädiktoren untereinander ab. Sind die beiden Prädiktoren unkorreliert, ist die größte inkrementelle Validität gegenüber der eigenständigen Validität eines einzelnen Prädiktors zu erwarten. Je stärker die beiden Prädiktoren korreliert sind, sie also Ähnliches messen, desto geringer fällt die inkrementelle Validität aus.

Aufgrund des vergleichsweise hohen Aufwands, der bei der Entwicklung und Durchführung von Studierfähigkeitstests betrieben werden muss, ist die inkrementelle Validität von Studierfähigkeitstests über Schulnoten von besonderem Interesse. Nimmt man die berichteten Validitätskoeffizienten für Schulnoten und Studierfähigkeitstests als Grundlage, so sind die in Tabelle 1 in der rechten Spalte aufgeführten Validitätskennwerte in Abhängigkeit der in der linken Spalte aufgeführten Interkorrelationen der beiden Prädiktoren bei einem gemeinsamen Einsatz von Schulnoten und Studierfähigkeitstest zu erwarten. Dabei stellen die geschätzten Validitätskennwerte im Falle eines kombinierten Einsatzes von Schulnoten und Studierfähigkeitstests eine optimistische Schätzung dar, da die beiden Prädiktoren mittels multipler Regressionen optimal gewichtet wurden.

Tabelle 1. *Zu erwartende inkrementelle Validität durch den gemeinsamen Einsatz von Studierfähigkeitstests und Schulnoten*

Interkorrelation Test und Abiturnote	Gemeinsame Validität bei optimaler Gewichtung*
.00	.71
.20	.65
.40	.60
.60	.56
Mit folgenden angenommenen eigenständigen Validitätskoeffizienten: Schulnoten = .52, Studierfähigkeitstest = .48	

Anmerkung: * optimale Gewichtung mittels multipler Regression

Über das bislang im deutschsprachigen Raum am gründlichsten erforschte Instrument, den Test für medizinische Studiengänge (TMS), berichtet Trost im 1998 erschienenen Abschlussbericht, dass die Korrelation zwischen TMS und der „Durchschnittsnote der Schule im betrachteten Zeitraum minimal .37 und maximal .48 beträgt“ (S. 56). Legt man diese Unter- und Obergrenze der Interkorrelation und die berichteten eigenständigen Validitäten der beiden Prädiktoren zugrunde, so ergibt sich ein geschätzter Wertekorridor für die Validität eines kombinierten Verfahrenseinsatzes von Schulnoten und Studierfähigkeitstests, der von .58 (Interkorrelation Prädiktoren: .48) bis .61 (Interkorrelation Prädiktoren: .37) reicht. Für die nachfolgenden Überlegungen wird von einem mittleren Validitätskoeffizienten im Falle eines kombinierten Verfahrenseinsatzes von .60 ausgegangen.

Im Rahmen der psychologischen Diagnostik entwickelte Nutzenmodelle erlauben die genaue Bestimmung von Trefferquoten in Abhängigkeit der Validität von Auswahlverfahren und sind somit in der Lage, Informationen über die Sinnhaftigkeit von Verfahrenskombinationen zu liefern. Die Trefferquote eines Auswahlverfahrens ist neben der Validität von der Basisrate (Anteil der Geeigneten unter den Bewerbern) und von der Selektionsquote (Prozentsatz der angenommenen Bewerber) abhängig (vgl. Taylor & Russel, 1939). In der folgenden Tabelle werden Trefferquoten unter der Annahme, dass 40 Prozent der BewerberInnen grundsätzlich geeignet sind für verschiedene Konstellationen dieser Parameter aufgeführt.

Tabelle 2. *Trefferquote eines Auswahlverfahrens in Abhängigkeit von Selektionsquote und Validität (angenommene Basisrate: 40%)*

		Selektionsquote									
		.05	.15	.25	.35	.45	.55	.65	.75	.85	.95
Validität	.00	.40	.40	.40	.40	.40	.40	.40	.40	.40	.40
	.10	.48	.46	.45	.44	.43	.43	.42	.42	.41	.40
	.20	.56	.52	.50	.48	.47	.46	.44	.43	.42	.41

Wie leicht zu erkennen ist, steigt die Trefferquote eines Verfahrens mit einer positiven Validität gegenüber einer Zufallsauswahl (Validität = 0) um so stärker an, je strenger die Selektionsquote ausfällt. Für die Frage, ob sich der Einsatz von Tests zusätzlich zum Einsatz von Schulnoten empfiehlt, bedeutet dies, dass ein Testeinsatz insbesondere in Auswahl-situationen sinnvoll ist, in denen nur ein kleiner Teil der BewerberInnen zugelassen werden kann. *Erst bei strengen Selektionsquoten ergeben sich substantielle Verbesserungen hinsichtlich der Trefferquote bei steigender Validität.*

Eine genauere Vorstellung über den Nutzen von Studierfähigkeitstests gewinnt man durch einen Vergleich der zu erwartenden Trefferquoten für die eigenständige Validität von Schulnoten (.52) und der gemeinsamen Validität von Schulnoten und Studierfähigkeitstests (.60) in Abhängigkeit verschiedener Selektionsquoten und Basisraten. Im Anhang dieses Beitrags sind die zu erwartenden Trefferquoten aufgeführt. Mittels der tabellierten Trefferquoten kann eine Universität für sich entscheiden, ob die Berücksichtigung von Testergebnissen zusätzlich zu Schulnoten sinnvoll

ist. *Generell kann ein Test bei einem Anteil der Geeigneten an den BewerberInnen von 80 Prozent oder darüber kaum empfohlen werden, da der Anstieg der Treffergenauigkeit nur sehr gering ausfällt. Dabei ist zu beachten, dass der Anstieg in der Treffergenauigkeit bei sehr geringen Basisraten auch wieder abnimmt (nicht tabelliert)! Für den Kontext der Studierendenauswahl sind aber wohl keine geringeren Basisraten als die in Tabelle 3 aufgeführten zu erwarten.*

Die bereits formulierte Gesetzmäßigkeit, dass ein *Testeinsatz eher bei strengen Selektionsquoten sinnvoll ist, gilt für alle Basisraten*. Auch aus praktischen Gründen wird ein zusätzlich zu Schulnoten eingesetzter Test bei strengen Selektionsquoten interessant, denn gerade bei außerordentlich beliebten Studiengängen erreicht der *numerus clausus* unter Umständen ein so hohes Niveau, dass viele BewerberInnen auch durch Wartezeiten oder sonstige Boni die Zulassungsbedingungen nicht erfüllen können. Der Einsatz eines weiteren Verfahrens eröffnet dieser BewerberInnengruppe eine zusätzliche Chance auf einen Studienplatz.

Eine endgültige Nutzenabwägung im Sinne einer Kostenkalkulation ist im vorliegenden Kontext schwierig, da auf mehreren Seiten sowohl Vorteile als auch Nachteile mit einer Testverwendung verbunden sind. Aus Sicht der *Universitäten* verursachen ungeeignete Studierende dadurch Kosten, dass sie ihr Studium abbrechen oder überdurchschnittlich lange studieren und somit Studienplätze blockieren. Diese Kosten lassen sich beziffern und es ist möglich, mit den im Anhang aufgeführten Trefferquotendifferenzen die Einsparungen zu berechnen, die die Berücksichtigung von Tests bei der Auswahl einbringen würde. Diesem Betrag stehen die Kosten für die Testentwicklung, Durchführung und Auswertung gegenüber, die bei einem bundesweit einheitlichen Test gegebenenfalls von den BewerberInnen selbst getragen werden könnten.

Solche Nutzenüberlegungen greifen allerdings zu kurz, da sie lediglich die Perspektive der Hochschulen einnehmen. Ihr sind andere Gesichtspunkte an die Seite zu stellen: sind die Verfahren für die *BewerberInnen* „nützlich“? Sind z. B. die Trefferquoten bei allen Subgruppen ähnlich? Unter bestimmten Umständen sind Verzerrungen (Fehler 2. Art) zu erwarten und hierdurch wären aus Sicht der BewerberInnen erhebliche „Kosten“ im Sinne von verstellten Entwicklungsmöglichkeiten zu erwarten (vgl. Raatz, 1978; Schönemann & Thompson, 1996).

Aus Sicht der *Gesellschaft* wiederum stellt sich die Regelung der Zulassung zu einem Hochschulstudium eher als Zuordnungsentscheidung denn als Selektionsentscheidung dar. Ziel sollte es also sein, für jede Person den passendsten Studienplatz zu finden und nicht eine Bestenauswahl zu betreiben.

Einsatz von Zulassungskriterien und Auswahlverfahren in der Praxis

Drei Aspekten sollte vor Einführung neuer Auswahlverfahren besondere Aufmerksamkeit geschenkt werden:

1. Durchführung einer Anforderungsanalyse: Die Anforderungsanalyse bildet die Ausgangsbasis für die Entwicklung von Auswahlverfahren. Mit ihr werden die Anforderungen bestimmt, die die Studierenden erfüllen müssen. In einem zweiten Schritt erfolgt die „Übersetzung“ der Anforderungsdimensionen in Personenmerkmale und die Zuordnung geeigneter Instrumente.

2. Durchführung einer Bedingungsanalyse: Die Parameter Validität des Auswahlverfahrens (ggf. zu schätzen aus der Metaanalyse), Basisrate und Selektionsquote (aus den Statistiken der Zulassungsabteilung ermittelbar) des betreffenden Studiengangs können bestimmt werden und zur Bestimmung des (inkrementellen) Nutzens der Auswahlkomponenten herangezogen werden.
3. Ressourcenplanung: Aufwändige Verfahren verlangen nach Ressourcen – langfristig! Es ist nicht hinreichend, ein Interview oder einen Leistungstest einmalig zu konstruieren, denn die Fragen und die Testaufgaben werden den BewerberInnen bei mehrmaligem Einsatz bekannt und damit diagnostisch wertlos. Nur durch eine kontinuierliche Weiterentwicklung und Erneuerung der Auswahlverfahren bleibt die Aussagekraft erhalten.

Abgesehen von der Validität für Studienerfolg sollten bei der Beurteilung von Auswahlverfahren weitere Gesichtspunkte berücksichtigt werden. Hierbei ist an die Akzeptanz der Verfahren bei den BewerberInnen (Hell & Schuler, 2005), die Fairness der Verfahren, die Imagewirkung für die Hochschule, hochschulspezifische Zielsetzungen (z. B. Diversifizierung) oder Honorierung von gesellschaftlichem Einsatz zu denken. Angesichts der komplexen Bewertungsgrundlage wäre es unangemessen, generell ein Zulassungskriterium oder ein Auswahlverfahren für alle Zulassungssituationen zu empfehlen. Vielmehr müssen Auswahlverfahren für die jeweilige Situation maßgeschneidert werden.

Literatur

- Heine, C., Briedis, K., Didi, H.-J., Haase, K., & Trost, G. (2006). *Auswahl- und Eignungsfeststellungsverfahren beim Hochschulzugang in Deutschland und ausgewählten Ländern*. HIS-Kurzinformation A3/2006.
- Hell, B. & Schuler, H. (2005). Verfahren der Studierendenauswahl aus Sicht der Bewerber. *Empirische Pädagogik*, 19, 361-376.
- Hell, B., Trapmann, S. & Schuler, H. (unter Begutachtung). Eine Metaanalyse der Validität von Studierfähigkeitstests im deutschsprachigen Raum.
- Hell, B., Trapmann, S., Weigand, S. & Schuler, H. (in Druck). Die Validität von Auswahlgesprächen im Rahmen der Hochschulzulassung – eine Metaanalyse. *Psychologische Rundschau*.
- Hunter, J., Schmidt, F. & Jackson, G. (1982). *Meta-analysis: cumulating research findings across studies*. London: Sage.
- Porter, M. (1985). *Competitive Advantage: Creating and Sustaining Superior Performance*. New York: The Free Press.
- Raatz, U. (1978). *Was können Tests bei Entscheidungsproblemen leisten? Lernzielorientierter Unterricht*, 2, 30-37.
- Schönemann, P. H. & Thompson, W. W. (1996). Hit-rate bias in mental testing. *Current Psychology of Cognition*, 15, 3-28.
- 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.

- Taylor, H.C. & Russel, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: Discussion and tables. *Journal of Applied Psychology*, 23, 565-578.
- 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. (Hrsg.) (1998). *Evaluation des Tests für medizinische Studiengänge (TMS): Synopse der Ergebnisse*. Bonn: Institut für Test- und Begabungsforschung.

Anhang

Trefferquote bei kombiniertem Einsatz von Studierfähigkeitstest und Abiturdurchschnittsnote (Validität: .60) gegenüber alleiniger Berücksichtigung der Abiturdurchschnittsnote (Validität: .52) in Abhängigkeit verschiedener Selektionsquoten und Basisraten

Selektionsquote

.05	.15	.25	.35	.45	.55	.65	.75	.85	.95
-----	-----	-----	-----	-----	-----	-----	-----	-----	-----

Basisrate: .20

.52	.60	.48	.42	.37	.33	.30	.28	.25	.23	.21
.60	.68	.54	.46	.40	.36	.32	.29	.26	.23	.21
Δ	.08	.06	.04	.03	.03	.02	.01	.01	.00	.00

Basisrate: .40

.52	.83	.73	.68	.63	.59	.55	.52	.48	.45	.42
.60	.88	.79	.72	.67	.62	.58	.53	.49	.46	.42
Δ	.05	.06	.04	.04	.03	.03	.01	.01	.01	.00

Validität

Basisrate: .60

.52	.94	.89	.85	.81	.78	.75	.72	.69	.66	.62
.60	.96	.92	.88	.85	.81	.78	.74	.71	.67	.63
Δ	.02	.03	.03	.04	.03	.03	.02	.02	.01	.01

Basisrate: .80

.52	.99	.97	.95	.94	.93	.91	.89	.87	.85	.82
.60	.99	.98	.97	.96	.94	.93	.91	.89	.86	.83
Δ	.00	.01	.02	.02	.01	.02	.02	.02	.01	.01

Anmerkung: Δ = Differenz Trefferquote

Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch fundierter Wissensprüfung

Margarete Litzenberger, Joachim F. Punter, Timo Gnambs, Marco Jirasko und Christiane Spiel

Zusammenfassung: An der Fakultät für Psychologie der Universität Wien wurde ein Auswahlverfahren für Psychologie-Studierende eingerichtet, welches die Absolvierung der Ringvorlesung „Psychologie als Wissenschaft“ umfasst. Als Prüfungsform wurde eine lehrzielorientierte und EDV-unterstützte Prüfung mit Multiple-Choice-Antwortformat eingeführt. Die Itemkonstruktion orientierte sich an mehreren Kriterien, die der Qualitätssicherung des Auswahlverfahrens dienen sollten. Zunächst wurden – in Anlehnung an die Taxonomie kognitiver Lernziele (Bloom, 1956) – Richtlinien zur Gestaltung von Multiple-Choice-Aufgaben erarbeitet. Darauf aufbauend erfolgte die Itementwicklung im Rahmen eines mehrstufigen Prozesses der Qualitätskontrolle. Dieses Procedere wird in diesem Beitrag ausführlich und beispielhaft dargestellt. Die Ergebnisse der Itemanalyse nach erfolgter Studierendenauswahl lassen auf eine zufriedenstellende Güte schließen.

Schlagwörter: Multiple-Choice, Studierendenauswahl, Auswahlverfahren

Abstract: The Faculty of Psychology (University of Vienna) established a student selection procedure based on the lecture “Psychology as a Science”. The examination was designed as a multiple-choice exam, based on educational objectives, and can be analyzed computer-assisted. The item-construction follows defined criteria to guarantee the quality of the selection procedure. In the first step, scientific guidelines for the construction of fair multiple-choice-questions were elaborated, based on the taxonomy of cognitive educational objectives (Bloom, 1956). In the second step, the item-development was carried out in a stepwise process of quality control. This procedure is presented in detail. The results of the item analysis provide satisfactory psychometric properties.

Key Words: Multiple-choice, Student Selection, Proficiency Assessment Procedure

Ausgangslage

Im Zusammenhang mit dem Urteil des Gerichtshofes der Europäischen Gemeinschaften vom 07.07.2005 zu den „Voraussetzungen des Zugangs zum Hochschulstudium – Diskriminierung“ wurde an österreichischen Universitäten mit einem möglichen Ansturm ausländischer Studierender aus dem EU-Raum gerechnet. Diese Regelung schien insbesondere für Studierende aus Deutschland eine gangbare Alternative, den im Heimatland notwendigen *Numerus clausus* (vor allem in den Studienfächern Psychologie, Pharmazie und Medizin) zu umgehen. Eine Verordnung des Rektorats der Universität Wien² ermöglichte schließlich die Durchführung eines Auswahlverfahrens zur Auswahl der am geeignetsten erscheinenden BewerberInnen. Da der Numerus clausus aus Fairnessgründen nicht in Frage kam, jedoch hoher Zeitdruck bestand (Studienbeginn: 1. Oktober 2005), wurde als zukunftsorientierte Maßnahme eine in vorhergehenden Jahren bewährte Prüfung zu einer Ringvorlesung als Auswahlkriterium festgelegt.

Die Ringvorlesung „Psychologie als Wissenschaft“ ist eine einführende Lehrveranstaltung, die Grundwissen in den insgesamt zwölf Prüfungsfächern des Diplomstudiums Psychologie an der Universität Wien vermittelt und in der ersten Semesterwoche jedes Wintersemesters in Form einer Blocklehrveranstaltung angeboten wird. Auf diese Weise können angehende Psychologie-Studierende bereits frühzeitig – im Sinne des Konzepts der *Berufswahlreife* (*vocational maturity*; vgl. Crites, 1974 bzw. Seifert, 1984) – umfassende Informationen über das Studienfach erwerben und so ihre Studienwahlentscheidung bestärken bzw. revidieren. Die positive Absolvierung der „Psychologie als Wissenschaft“ (zweiteilige Lehrveranstaltung zu je 1 SWS) ist laut Studienplan Voraussetzung für die Teilnahme an allen Lehrveranstaltungen mit immanentem Prüfungscharakter (Übungen, Proseminare). Seit dem Studienjahr 2005/06 ist nun an diese Prüfung zusätzlich das Auswahlverfahren gekoppelt.

Seit Einführung dieser Eingangslehrveranstaltung wurden jährlich etwa 1500 Prüfungen (mit freiem Antwortformat) pro Studienjahr abgenommen; die investierte Arbeitszeit für die Abwicklung betrug etwa 1000 Stunden und konnte nur mit Unterstützung aller FakultätsmitarbeiterInnen jeweils zeitgerecht bewerkstelligt werden. Zur Handhabung des (infolge des EuGH-Urteils) erwarteten hohen Ansturms von Studierenden aus dem EU-Raum und der enormen Prüfungsbelastung, wurde von der Fakultät für Psychologie die Prüfung zur Vorlesung „Psychologie als Wissenschaft“ vom freien Antwortformat auf das Multiple-Choice-Antwortformat umgestellt. Ziel dieser Umstellung war damit auch die Implementierung eines fairen Auswahlverfahrens durch die Konstruktion von eindeutigen Multiple-Choice-Aufgaben. Mit der Umsetzung wurde ein Projektteam beauftragt.

Lernpsychologische und messtheoretische Grundlagen

Entsprechend der oben genannten Ziele und Vorgaben wurden zunächst, durch Aufarbeitung und Dokumentation der einschlägigen Literatur, Grundlagen für die Erstellung von computerisiert auswertbaren Multiple-Choice-Prüfungsfragen erarbeitet, die

² Verordnung des Rektorats der Universität Wien vom 08.09.2005 (Mitteilungsblatt, Studienjahr 2004/2005, ausgegeben am 08.09.2005, 39. Stück)

das reine „Wiedererkennen von Faktenwissen“ übersteigen. Besonderes Anliegen bei der Implementierung des Auswahlverfahrens war die lernpsychologische Fundierung der angestrebten Wissensprüfung. Technische und administrative Notwendigkeiten für die Durchführung derartiger Prüfungen wurden ermittelt und schließlich ein ausreichend umfangreicher Itempool für die Prüfung „Psychologie als Wissenschaft“ konstruiert. Zur Sicherung der hohen Qualitätsansprüche eines Auswahlverfahrens wurde ein mehrstufiges Kontrollregulativ zur Prüfung der formalen und inhaltlichen Qualität der Items implementiert. Diese Vorgehensweise zur Qualitätssicherung wird im Folgenden vorgestellt.

Taxonomien zur Generierung von Wissensfragen

Taxonomien, um die kognitive Fähigkeit zur Lösung eines Testitems zu spezifizieren, wurden verschiedentlich entwickelt. Die bekannteste stammt von Bloom (1956), eine weitere, welche der Itemkonstruktion zugrunde liegt, wurde von Roid und Haladyna (1982) veröffentlicht. Während die Bloomsche Taxonomie eher ein theoretisches Gerüst darstellt und zur Klassifikation von Items geeignet ist (Legg, 1991), bietet sich letztere Taxonomie speziell als Grundlage für den Prozess der Itemkonstruktion von Multiple-Choice-Tests an.

Bloom unterscheidet drei Bereiche (kognitiv, affektiv und psychomotorisch), denen er eine Reihe hierarchisch gegliederter Lernziele zuordnet. Leistungsbeurteilungen im Rahmen von Wissenstests, wie die Eingangsprüfung zur Studierendenauswahl in der Psychologie, sind primär dem kognitiven Bereich zuzuordnen. Es handelt sich hierbei um Prozesse des Wissens, Verstehens, Begründens u. ä.. Affektive und psychomotorische Lernziele finden deshalb im Folgenden keine Berücksichtigung (ohne deren Bedeutung für die psychologische Ausbildung schmälern zu wollen). Im Detail unterscheidet Bloom (1956) im kognitiven Bereich sechs Lernziele:

1. *Wissen (Knowledge)*: Informationen (von spezifischen Fakten bis zu ganzen Theorien) sollen aus dem Gedächtnis wiedergegeben werden, z. B. Reproduktion von Fachausdrücken, spezifische Fakten, Grundprinzipien.
2. *Verständnis (Comprehension)*: Die Bedeutung des Erlernten soll erfasst werden ohne es notwendigerweise auf andere (neue) Bereiche zu übertragen oder Implikationen daraus abzuleiten, z. B. Graphiken und Diagramme interpretieren, verbales Material in mathematische Formeln übertragen, Konzepte und Prinzipien verstehen.
3. *Anwendung (Application)*: Erlerntes (Regeln, Methoden, Konzepte, Theorien etc.) soll in neuen und konkreten Situationen angewendet werden, z. B. Konzepte in neuen Sachlagen anwenden, Gesetze und Theorien auf praktische Anwendungsfelder umlegen.
4. *Analyse (Analysis)*: Material soll in seine Bestandteile zerlegt werden, z. B. unausgesprochene Annahmen oder logische Fehlschlüsse erkennen, den Aufbau des Materials zerteilen können.
5. *Synthese (Synthesis)*: Aus Teilen, Elementen etc. soll etwas neues Ganzes zusammengestellt werden, z. B. Untersuchungsdesigns abwandeln, Forschungsabläufe in die richtige Reihenfolge bringen.

6. *Beurteilung (Evaluation)*: Die Güte eines vorgegebenen Materials soll für einen bestimmten Zweck eingeschätzt werden, z. B. Korrektheit von Interpretationen aufgrund der Datenlage beurteilen, Material anhand interner bzw. externer Standards beurteilen.

Die Bloomschen Lernziele sind nicht unabhängig voneinander zu betrachten. Sie stehen vielmehr in einer geordneten Rangreihe zueinander. Wurde ein untergeordnetes Lernziel (wie z. B. Wissen) nicht erfolgreich erreicht, wird auch ein übergeordnetes Lernziel eher nicht zu bewältigen sein. Jemand der z. B. nicht weiß, was eine unabhängige Variable ist (*Wissen*), wird auch nicht in der Lage sein, diese in einem konkreten Beispiel zu identifizieren (*Anwendung*).

Ergänzend bzw. als Überbau der Testkonstruktion schlagen Roid und Haladyna (1982) eine kriteriumsbezogene Testentwicklung vor, wobei „kriteriumsbezogen“ so verstanden wird, dass die Items eines Tests eine repräsentative Auswahl der Lehrziele eines Fachs darstellen. Dementsprechend bedeutet ein Testscore von 85%, dass der Studierende 85% des Lehrstoffs beherrscht. Zur Entwicklung eines kriteriumsbezogenen Tests schlagen die Autoren einen fünfstufigen Prozess vor:

1. *Abklärung der Lehrziele*: Der Lehrende/Prüfende sollte eine klare Vorstellung über die Lehrziele haben, diese können eher abstrakt (z. B. „die Funktionsweise des menschlichen Gehirns verstehen“) oder auch sehr konkret (z. B. „den Grundaufbau einer Nervenzelle beschreiben können“) sein.
2. *Operationalisierung der Lehrziele*: Hier geht es um die Spezifikation der zuvor definierten Lehrziele und der Festlegung konkreter Regeln zur Itemkonstruktion mit dem Zweck, Lehrziele messbar zu machen.
3. *Itementwicklung*: Je nach Lehrziel und Schwierigkeitsniveau der Items können verschiedene Techniken zur Itemkonstruktion angewandt bzw. Typen von Multiple-Choice-Aufgaben (z. B. Mehrfachwahl-Aufgaben, Lückentext, Zuordnungsaufgaben) realisiert werden.
4. *Itemrevision*: In diesem Schritt sollen die Items einer logischen und einer empirischen Itemrevision unterzogen werden. Die generierten Items werden einerseits von ExpertInnen des Fachs hinsichtlich fachlicher Fehler bzw. der Konsistenz zwischen Item und Lehrziel geprüft und andererseits einer empirischen Kontrolle unterzogen um Mehrdeutigkeiten auszuschließen.
5. *Testentwicklung*: Nach der ersten Erprobung und Revision der Items liegt schließlich ein Itempool vor, aus dem zufällig eine Itemauswahl für eine kriteriumsbezogene Testerstellung im oben beschriebenen Sinn getroffen werden kann.

Die Itemkonstruktion im Rahmen der Implementierung der Multiple-Choice Prüfung zur „Psychologie als Wissenschaft“ orientierte sich vornehmlich an den beiden vorgestellten Taxonomien, wenngleich aus organisatorischen Gründen einige Abstriche gemacht werden mussten. Dennoch wurde die Testkonstruktion damit auf eine allgemein anerkannte und lernpsychologisch fundierte Basis gestellt.

Richtlinien zur Konstruktion von Multiple-Choice Aufgaben

Multiple-Choice stellt nicht ein spezifisches Antwortformat dar, sondern kann als Überbegriff verschiedener geschlossener Antwortformate verstanden werden. Grundsätzlich werden zwei Gruppen unterschieden: dichotome (*true/false* bzw. *alternate choice*) und polytome Antwortformate, bei denen eine (die beste) Antwort zu wählen ist. Ein wesentlicher Unterschied dieser beiden Ansätze besteht im Informationsgehalt der Antwortoptionen. Bei Richtig/Falsch-Items muss jede Antwortoption eindeutig und absolut richtig oder falsch sein, während bei Best-Answer-Items eine graduelle Abstufung der Korrektheit vorliegen kann.

Das dichotome Antwortformat wird sehr kontrovers diskutiert, wobei die Vorteile vor allem darin liegen, dass sie einfach zu konstruieren sind, mehr Inhalte in der gleichen Vorgabezeit als bei klassischem Multiple-Choice abgefragt werden können und auch höhere Lernziele erfassbar sind (vgl. Haladyna, 1992). Diese Vorteile sind jedoch durch eine Reihe von Studien (vgl. Downing, Baranowsk, Grosso & Norcini, 1995) in Frage gestellt. So können zwar mehr Items in diesem Format vorgegeben werden – Testscores von klassischen Multiple-Choice Items sind jedoch reliabler. Zudem zeigten sich Items als unterschiedlich schwierig, je nachdem ob der Stamm positiv oder negativ formuliert wurde und sie richtig bzw. falsch sind (Candida, Petersen & Petersen, 1976). Falsche Items weisen zudem eine größere Trennschärfe als Richtige auf (Ebel, 1972, zitiert nach Haladyna, 1992). Vergleichbare Kritik trifft multiple Richtig/Falsch-Items (d. h. ein Item besteht aus einem Aufgabenstamm und mehreren Antwortoptionen, die alle mit richtig/falsch zu beantworten sind). Während die Akzeptanz dieses Antwortformats bei einzelner Verrechnung der Antwortoptionen bei den Studierenden deutlich höher ist (Macher, 2005), dürften Ratewahrscheinlichkeiten und Antworttendenzen problematisch sein – empirische Studien hierzu fehlen allerdings bislang.

Best-Answer-Items sind in der Regel länger als klassische Multiple-Choice-Items, es muss daher mit längeren Bearbeitungszeiten gerechnet werden, wodurch in der gleichen Zeitspanne weniger Items vorgegeben und weniger Inhalte abgefragt werden können. Darüber hinaus kann hier Halbwissen (z. B. Ausschluss einzelner Antwortoptionen) helfen, die richtige Lösung zu finden, indem Distraktoren eliminiert werden können (d. h. das Format birgt Hinweise auf die Lösung für testefahrene Studierende). Da die primären Anliegen der Itemkonstruktion eine automatisierte Auswertbarkeit und eine einfache Scorebildung (= Anzahl richtig bearbeiteter Items) waren, werden spezifische Typen von Best-Answer-Antwortformaten (z. B. Zuordnungsaufgaben, kontextbezogene Multiple-Choice-Aufgaben) hier nicht näher in Betracht gezogen.

Multiple-Choice mit Mehrfachwahl bestehen aus einem Aufgabenstamm und mehreren Antwortoptionen. Im Unterschied zu klassischen Multiple-Choice-Items kann es dabei nicht nur eine, sondern mehrere richtige Antworten geben. Die Auswertung kann dabei einerseits in Anlehnung an multiple Richtig/Falsch-Items erfolgen (für jede richtigerweise ausgewählte bzw. nicht ausgewählte Antwortmöglichkeit wird ein Punkt vergeben) und andererseits nach dem Alles-oder-Nichts-Prinzip (nur wenn alle richtigen und keine falschen Antwortmöglichkeiten markiert werden, wird das Item als „gelöst“ gewertet und mit einem Punkt verrechnet). Letzterer Verrechnungsmodus erhöht die Schwierigkeit der Items deutlich, stößt aber bei Studierenden naturgemäß auf geringere Akzeptanz (vgl. Macher, 2005).

Um die Ratewahrscheinlichkeit zu minimieren und gleichzeitig die Möglichkeit auszuschließen, dass Halbwissen zur Lösung einzelner Items führen kann, wurde bei der Festlegung des Antwortformats der Prüfung zur Ringvorlesung für ein Multiple-Choice mit Mehrfachwahl und Auswertung nach dem Alles-oder-Nichts-Prinzip entschieden. Die Anzahl richtiger Antwortmöglichkeiten pro Item wurde variiert, wobei es Teil der Aufgabe der Studierenden war, die Anzahl Richtiger je Item herauszufinden. Auf diese Weise wurde sichergestellt, dass nur die Beherrschung des Lehrstoffs zur erfolgreichen Absolvierung der Prüfung führt und Zufallsfaktoren weitestgehend ausgeschlossen werden können. Zur Vereinheitlichung und damit effizienteren Auswertbarkeit wurden fünf Antwortoptionen pro Item festgelegt, wobei entsprechend der Empfehlung von Haladyna, Downing und Rodriguez (2002) niemals keine oder alle Antwortmöglichkeiten richtig sein konnten.

Zur optimalen Anzahl der Distraktoren liegen eine Reihe (teils widersprüchlicher) Ergebnisse vor. Am konklusivsten erscheinen die Metaanalysen von Haladyna und Downing (1989) sowie Rodriguez (2005). Danach sinkt zwar die Itemschwierigkeit bei drei Antwortoptionen im Vergleich zu vier oder fünf Optionen leicht, aber Trennschärfe und Reliabilität steigen, während die Validität davon unberührt bleibt. Wie Haladyna und Downing (1993) im Rahmen einer umfangreichen Reanalyse verschiedener Multiple-Choice-Tests beobachten konnten, sind (unabhängig von der tatsächlich vorgegebenen Anzahl an Distraktoren) bei über zwei Drittel der Items lediglich ein oder zwei Distraktoren auch effektiv wirksam. Nur ein bis acht Prozent aller Items weisen drei wirksame Distraktoren auf. Kubinger, Holocher-Ertl und Frebort (2006) konnten allerdings zeigen, dass Items mit weniger als $k = 6$ Antwortmöglichkeiten (mit lediglich einer richtigen Antwortoption) psychologisch-diagnostisch kaum vertretbar sind. Zusätzlich ergaben ihre Analysen, dass das Antwortformat 2 (aus 5) richtige Antwortoptionen zumindest nicht leichter ist als das freie Antwortformat. Da es aufgrund des Zeitdrucks keine Möglichkeit gab die Güte der Distraktoren vorab zu ermitteln und man nicht Gefahr laufen wollte, zu leichte Distraktoren anzubieten, wurde demgemäß für ein fünfstufiges Antwortformat entschieden.

Nach Festlegung von Kriterien zur Gestaltung der Multiple-Choice Aufgaben wurden ausgehend von einer umfangreichen Literaturrecherche (vgl. Haladyna & Downing, 1989; Moreno, Martínez & Muñiz, 2006) Richtlinien zur konkreten Itemgestaltung festgelegt. Diese umfassen inhaltliche Empfehlungen (z. B. *Jedes Item muss sich an konkreten inhaltlichen Lehrzielen des Fachgebiets orientieren*), formale Gestaltungshinweise (z. B. *Die Items sollen so kurz wie möglich, aber so lang wie für das Verständnis notwendig sein*) und Formulierungshilfen (z. B. *Die Items sollten positiv formuliert sein, Negationen sind nur in Ausnahmefällen zulässig*). Die Richtlinien sollten den ItemkonstrukteurInnen als wissenschaftliche Grundlage und zur Orientierung an einheitlichen Standards dienen.

Maßnahmen zur Qualitätssicherung

Zur Gewährleistung der hohen Qualitätsansprüche eines Auswahlverfahrens wurde ein mehrstufiges Kontrollregulativ zur Prüfung der formalen und inhaltlichen Qualität der Items implementiert. Die Vorgehensweise zur Qualitätssicherung wird im Folgenden vorgestellt.